A Bayesian framework for risk perception

van Erp, Noel

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

A Bayesian framework for risk perception

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.Ch.A.M. Luyben;
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 4 december 2017 om 12:30 uur

door

Henri Robert Noël VAN ERP
Doctorandus Psychologie, Rijks Universiteit Groningen
geboren te Den Haag, Nederland

ii

Dit proefschrift is goedgekeurd door de promotoren:
Prof. ir. drs. J.K. Vrijling
Prof. dr. ir. P.H.A.J.M. van Gelder

Samenstelling promotiecommisssie:
Rector Magnificus   Voorzitter
Prof. ir. drs. J.K. Vrijling   Technische Universiteit Delft, Horvat, Promotor
Prof. dr. ir. P.H.A.J.M. van Gelder   Technische Universiteit Delft, Promotor

Onafhankelijke leden:
Prof. dr. ir. M. Kok   Technische Universiteit Delft, HKV$_{\text{LIJN IN WATER}}$
Prof. dr. B.T. Adey   Eidgenössiche Technische Hochschule Zürich
Prof. dr. J. Groeneweg   Technische Universiteit Delft
Dr. N. Khakzad   Technische Universiteit Delft
Dr. ir. R.B. Jongejan   Jongejan Risk Management Consulting

Contents

Chapter 1

Algemene Inleiding

In deze dissertatie zal er een Bayesiaans raamwerk voor risicoperceptie worden gegeven. Dit raamwerk omvat plausibiliteitsoordelen, besliskunde en het stellen van vragen. Plausibiliteitsoordelen worden gemodelleerd middels de Bayesiaanse kansrekening, besliskunde middels een Bayesiaanse besliskunde, en relevantie-oordelen middels een Bayesiaanse informatietheorie. Deze theorieen worden, respectievelijk, beproken in Deel I, II, en III van deze these.

De Bayesiaanse kansrekening is relatief wel bekend, aangezien deze theorie nu ook in de niet-Engelstalige landen in opmars is. De Bayesiaanse kansrekening is niet alleen een krachtig instrument voor data analyse, het is tevens een model voor de wijze waarop we (impliciet) inductie plegen, dat wil zeggen, de manier waarop we plausibiliteitsoordelen maken op de basis van onvolledige informatie. In Deel I van deze thesis beschouwen we dat de Bayesiaanse kansrekening niets meer is dan een gequantificeerde *common sense*.

De Bayesiaanse besliskunde, zoals gegeven in deze thesis, is een directe afgeleide van de Bayesiaanse kansrekening. In deze besliskunde vergelijken we utiliteitskansverdelingen, welke we construeren door utiliteiten, oftewel, subjectieve waardeoordelen, toe te kennen aan de objectieve uitkomsten van onze uitkomstenkansverdelingen, welke afgeleid zijn middels de Bayesiaanse kansrekening.

Wanneer de uitkomsten in onze uitkomstenkansverdelingen monetair zijn, dan mogen we gebruik maken van de psycho-physische Weber-Fechner wet, oftewel, Bernoulli's utiliteitsfunctie, om utiliteiten toe te kennen aan deze uitkomsten. Deze *mapping* van uitkomsten naar utiliteiten transformeert onze uitkomstenkansverdelingen naar hun corresponderende utiliteitskansverdelingen.

De utiliteitskansverdeling welke meer naar rechts licht op de utiliteits-as zal, afhankelijk van de context, danwel meer winstgevender danwel minder verliesgevender zijn dan de utiliteitskansverdeling welke meer naar links licht. Hieruit volgt dat we geneigd zullen zijn die beslissing te nemen welke onze utiliteitskansverdelingen 'maximaliseert'. In Deel II van deze these zullen we de Bayesiaanse besliskunde toepassen op zowel een investeringsprobleem als ook twee verzekeringsproblemen.

Niet alle vragen zijn eender, somige vragen, indien beantwoord, zullen informatiever zijn dan andere. Met andere woorden, vragen kunnen verschillen wat betreft hun relevantie ten aanzien van een gegeven *issue of interest* welke we uitgezocht wensen te hebben. Dit feit wordt verwoord in het adagium: '*to know the question, is to have gone half the journey*'.

Bayesiaanse informatie theorie, middels een wiskundige operationalisatie van wat een vraag is, stelt ons in staat om te bepalen welke vraag, indien beantwoord, het meest informatief zal zijn ten aanzien van een gegeven *issue of interest*. In de Bayesiaanse informatietheorie worden er relevanties toegekend aan alle mogelijke vragen welke gesteld kunnen worden. Op deze relevanties worden dan middels de informatietheoretische produkt- en somregels uitgevoerd, om zo de relevantie van een vraag in relatie tot een gegeven *issue of interest* te kunnen bepalen.

De Bayesiaanse informatietheorie is een uitbreiding van het 'kanvas van de rationaliteit' en, daarmee, van de mogelijke psychologische phenomena welke zich lenen voor een mathematische analyse. Zo kunnen we niet alleen relevanties toekennen aan vragen, maar ook aan boodschappen welke aan ons gecommuniceerd worden door een bron van informatie.

De relevantie van een boodschap representeert de bruikbaarheid van die boodschap, indien ontvangen, voor het bepalen van een gegeven issue of interest. Door een relevantie aan een boodschap toe te kennen, kennen we indirect een relevatie toe aan de bron van informatie zelf; mogelijke voorbeelden van bronnen van informatie zijn de media, wetenschappers, en overheidsinstanties. In Deel III van deze these zullen we een informatietheoretische analyse geven van een simpel risicocommunicatie scenario.

De Bayesiaanse kansrekening heeft haar axiomatische wortels in de lattice theorie, daar de produkt- en somregels van de Bayesiaanse kansrekening afgeleid kunnen worden middels consistentie *constraints* op de lattice van uitspraken. Op een zelfde wijze, middels consistentie *constraints* op de lattice van vragen, kunnen we ook de produkt- en somregel van de Bayesiaanse informatietheorie afleiden.

Dus, indien we rationaliteit, oftewel, consistentie *constraints* op lattices, nemen als ons leidend principe in de afleiding van onze theorieen van inferentie, dan krijgen we aan de ene kant de Bayesiaanse kansrekening, met de Bayesiaanse besliskunde als een specifieke toepassing van deze kansrekening, en aan de andere kant krijgen we een Bayesiaanse informatietheorie. Door zo te doen verkrijgen we een omvattend, coherent, en krachtig raamwerk waarmee we het menselijk redeneren, in de breedste zin van het woord, kunnen modelleren.

Chapter 2

General Introduction

We present here a Bayesian framework of risk perception. This framework encompasses plausibility judgments, decision making, and question asking. Plausibility judgments are modeled by way of Bayesian probability theory, decision making is modeled by way of a Bayesian decision theory, and relevancy judgments are modeled by way of a Bayesian information theory. These theories are discussed in Parts I, II, and III, respectively, of this thesis.

Bayesian probability theory is fairly well known and well established. Bayesian probability theory is not only a powerful tool of data analysis, but it also may function as a model for the way we (implicitly) do induction, that is, the way we make plausibility judgments on the basis of incomplete information. In Part I of this thesis we will make the case that Bayesian probability theory is nothing but common sense quantified.

The Bayesian decision theory, as proposed in this thesis, derives directly from Bayesian probability theory. In this decision theory we compare utility probability distributions, which are constructed by way of assigning utilities, that is, subjective worths, to the objective outcomes of our outcome probability distributions, which are derived by way of Bayesian probability theory.

When the outcomes under consideration are monetary, then we may use the Weber-Fechner law of psychophysics, or, equivalently, Bernoulli's utility function, to assign utilities to these outcomes. This mapping of outcomes to utilities, transforms our outcome probability distributions to their corresponding utility probability distributions.

That utility probability distribution which is located more to the right on the utility axis will tend to be, depending on the context of our problem of choice, either more profitable or less disadvantageous than the utility probability distribution that is more to the left. So, we will tend to prefer that decision which 'maximizes' our utility probability distributions. This then, in a nutshell, is the whole of our Bayesian decision theory. In Part II of this thesis, we will apply the Bayesian decision theory to both investment and insurance problems.

Not all questions are equal, some questions, when answered, may give us more information than others. Stated differently, questions may differ in their

relevancy, in relation to some issue of interest we wish to see resolved. This is borne out by the well known adage that, 'to know the question, is to have gone half the journey'.

Bayesian information theory, by way of a mathematical operationalization of the concept of a question, allows us to determine which question, when answered, will be the most informative in relation to some issue of interest. The Bayesian information theory does this by assigning relevancies to the questions under consideration. These relevancies are then operated upon, by way of the information theoretical product and sum rules, in order to determine the relevancy of some question in relation to the issue of interest.

The Bayesian information theory constitutes an expansion of the 'canvas of rationality', and, consequently, of the range of psychological phenomena which are amenable to mathematical analysis. For example, we may assign relevancies not only to questions, but also to the messages that are communicated to us by some source of information.

The relevancy of a message represents the usefulness of that message, when received, in determining some issue of interest. By assigning a relevancy to the message, we indirectly assign a relevancy to the sources of information itself; possible examples of sources of information being the media, scientists, and governmental institutions. In Part III of this thesis, we will give an information theoretical analysis of a simple risk communication problem.

Bayesian probability has its axiomatic roots in lattice theory, as the product and sum rule of Bayesian probability theory may be derived by way of consistency requirements on the lattice of statements. One may derive, likewise, by way of consistency requirements on the lattice of questions, the product and sum rules of Bayesian information theory.

So, if we choose rationality, that is, consistency requirements on lattices, as our guiding principle in the derivation of our theories of inference, then we get on the one hand a Bayesian probability theory, with as its specific application a Bayesian decision theory, and on the other hand we get a Bayesian information theory. In doing so, we obtain a comprehensive, coherent, and powerful framework with which to model human reasoning, in the widest sense.

Part I

Bayesian Probability Theory

Chapter 3

Introduction

The Bayesian decision theoretic framework proposed in this thesis has as one of its basic assumptions that Bayesian probability theory, by construction, is common sense quantified. So, it is felt that an explicit treatment of Bayesian probability theory needs to be included, for those readers not yet familiar with Bayesian probability theory in general and Jaynes' *Probability Theory; The logic of science*, [47], in particular.

We will demonstrate in this part of the thesis that Bayesian probability theory is an extended logic, in that the strong and weak syllogisms of Aristotelian logic, as well as the even weaker plausible syllogisms, may be derived by way of the product and sum rules [47]. Some worked out examples will be given to further strengthen the claim that Bayesian probability theory is indeed common sense quantified. We will also give an outline of the axiomatic underpinnings of Bayesian probability theory, as well as a very short and rough historical overview of the Bayesian probability theory, from its inception by Laplace in the 18th century until now.

Chapter 4

Is Induction Bayesian?

In this chapter we first will give a discussion on the product and sum rules of Bayesian probability theory. We then construct a symbolic Bayesian Network by way of the product and sum rules of Bayesian probability theory, in order to demonstrate the qualitative correspondence of these rules with common sense.

4.1 Bayesian Probability Theory

The whole of Bayesian probability theory flows forth from two simple rules, the product and sum rules [47],

$$P(A)\,P(\,B|\,A) = P(AB) = P(B)\,P(\,A|\,B)\,, \qquad (4.1)$$

where $P(\,B|\,A)$ is the probability of B being true given that A is true, $P(\,A|\,B)$ the probability of A being true given that B is true, and $P(AB)$ the probability of both A and B being true, and

$$P(\overline{A}) = 1 - P(A)\,, \qquad (4.2)$$

where $\overline{A}$ is the negation of A and $P(\overline{A})$ is the probability of not-A being true.

Now, at first glance, it may seem to be somewhat surprising that the whole of Bayesian probability theory flows forth from the product and sum rules, (4.1) and (4.2). But it should be remembered that Boolean algebra, on an operational level, is nothing more than a repeated application of AND- and NOT-operations on logical propositions.

In the product and sum rules, (4.1) and (4.2), we have the plausibility operators of the logical conjunction AB and negation $\overline{A}$, respectively. So, the plausibility of any proposition that is generated in the Boolean algebra may be arrived at by repeated applications of the product rule and sum rules [47].

For example, the AND- and NOT-operations combine, by way of the identity of de Morgan, in an OR-operation:

$$A + B = \overline{\overline{A}\,\overline{B}}. \qquad (4.3)$$

where the symbol '+' stands for the OR-operator, or, equivalently, logical dis-
junction. By way of the identity of de Morgan and the product and sum rules,
(4.3), (4.1) and (4.2), the plausibility of the logical disjunction $A + B$, also
known as the generalized sum rule, may be derived as follows:

$$P(A + B) = P\left(\overline{\overline{A}\,\overline{B}}\right)$$

$$= 1 - P\left(\overline{A}\,\overline{B}\right)$$

$$= 1 - P\left(\overline{A}\right) P\left(\overline{B}\middle|\,\overline{A}\right)$$

$$= 1 - P\left(\overline{A}\right) \left[1 - P\left(B\middle|\,\overline{A}\right)\right]$$

$$= P(A) + P\left(\overline{A}B\right) \tag{4.4}$$

$$= P(A) + P(B) P\left(\overline{A}\middle|\,B\right)$$

$$= P(A) + P(B) \left[1 - P\left(A\middle|\,B\right)\right]$$

$$= P(A) + P(B) - P(AB)\,.$$

The generalized sum rule (4.4) is one of the most useful in applications. It
gives rise to probability distributions, be they discrete or continuous, as well as
the Bayesian practice of summating over those parameters one is not directly
interested in (i.e., the nuissance parameters.)

   If we have two propositions that are exhaustive and mutually exclusive, for
example A and its complement $\overline{A}$, then the probability that either A or its
complement $\overline{A}$ will occur is one:

$$P\left(A + \overline{A}\right) = 1, \tag{4.5}$$

whereas the probability of A and $\overline{A}$ occurring at the same time is zero:

$$P\left(A\overline{A}\right) = 0. \tag{4.6}$$

By way of (4.4) and (4.6), we then have

$$P\left(A + \overline{A}\right) = P(A) + P\left(\overline{A}\right). \tag{4.7}$$

If we combine (4.5) and (4.7), we find that the probabilities of a Bernoulli
distribution should sum to one:

$$P(A) + P\left(\overline{A}\right) = 1. \tag{4.8}$$

For n propositions that are exhaustive and mutually exclusive, (4.5) generalizes to

$$\sum_{i=1}^{n} P(A_i) = 1. \tag{4.9}$$

And if we both let the A_i correspond with numbers on the real axis in the range (a, b) and let the number of propositions n tend to infinity, then we may go from discrete probability distributions to a continuous probability distributions:

$$\int_{a}^{b} p(A) \, dA = 1, \tag{4.10}$$

as summation tends to integration.

From both the fact that a disjunction of an exhaustive and mutually exclusive set of propositions always holds true and the Boolean property of the distributivity of propositions, we have that

$$B = (A + \overline{A}) B = AB + \overline{A}B. \tag{4.11}$$

Substituting (4.11) into (4.4), we find that

$$P(B) = P(AB) + P(\overline{A}B) - P(A\overline{A}B), \tag{4.12}$$

where we have made use of the Boolean property of idempotence, $BB = B$. Now, as A and $\overline{A}$ are mutually exclusive, it follows that they cannot occur in conjunction. So, the proposition $A\overline{A}B$ represents an impossibility:

$$P(A\overline{A}B) = 0. \tag{4.13}$$

Substituting (4.13) into (4.12), we have that

$$P(B) = P(AB) + P(\overline{A}B). \tag{4.14}$$

By way of a set of n exhaustive and mutually exclusive propositions $A_i = \{A_1, \ldots, A_n\}$, (4.14) may be generalized to

$$P(B_j) = \sum_{i=1}^{n} P(A_i B_j). \tag{4.15}$$

Furthermore if the set of m propositions $B_j = \{B_1, \ldots, B_m\}$ is also exhaustive and mutually exclusive, then from a repeated application of (4.9) we have that

$$\sum_{j=1}^{m} P(B_j) = 1, \tag{4.16}$$

or, equivalently,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} P(A_i B_j) = 1. \tag{4.17}$$

And if let the A_i and B_j correspond with numbers on the real axis in the ranges (a, b) and (c, d), respectively, and let the number of propositions n and m tend to infinity, then we may go from discrete probability distributions to a continuous probability distributions:

$$p(B) = \int_a^b p(A, B)\, dA. \tag{4.18}$$

and

$$\int_c^d p(B)\, dB = 1, \tag{4.19}$$

or, equivalently,

$$\int_c^d \int_a^b p(A, B)\ dA\, dB = 1, \tag{4.20}$$

as summation tends to integration. Note that in the derivation of both (4.15) and (4.18) we have the rationale behind the Bayesian practice of summation and integration over nuisance parameters that are of no direct interest.

In closing, Bayesian probability theory is very simple on a conceptual level; one just needs to apply the product and sum rules, (4.1) and (4.2). However, on an implementation level, when doing an actual data-analysis, it may be quite challenging. In close analogy, Boolean algebra is simple on the conceptional level; one just needs to apply the AND- and NOT-operators. However, on the implementation level it may be quite challenging, when, say, we use this Boolean algebra to design logic circuits for computers.

We refer the interested reader to Skilling's [92], for a first cursory overview on the considerations that come with a Bayesian data-analysis. Though the absolute authority is Jaynes' [47]. But the reading of this tome would require a considerable time investment on the part of the reader. But then again, as Calculus is the highway to the exact sciences, so we have that *Probability Theory: The Logic of Science* is the highway to Bayesian statistics.

4.2   A Symbolic Bayesian Network

Bayesian probability theory is not only said to be common sense quantified, but also common sense amplified. If Bayesian probability theory were not common sense amplified, then it could not ever hope to enjoy the successes it currently enjoys in the various fields of science; astronomy, astrophysics, chemistry, image recognition, etc., having a much higher 'probability resolution' than our human brains can ever hope to achieve [47]. This statement is in accordance with the finding that, if presented with some probability of a success p subjects fail to draw the appropriate binomial probability distribution of the number of successes r in n draws. Since experimental subjects manage to find the expected number of successes, but they fail to accurately determine the probability spread of the r successes [51].

Behavioral economists see this finding as evidence that humans are fundamentally non-Bayesian in the way they do their inference, [51]. We instead propose that human common sense is not hard-wired for problems involving sampling distributions. Otherwise there would be no need for such a thing as data-analysis, as we only would have to take a quick look at our sufficient statistics after which we then would draw the probability distributions of interest. However, humans do seem to be hard-wired for the day to day problems of inference.

For example, if we are told that our burglary alarm has gone off, after which we are also told that a small tremor has occurred in the vicinity of our house around the time that the alarm went off. Then common sense would suggest that the additional information concerning the occurrence of a small tremor will somehow modify our probability assessment of there actually being a burglar in our house.

We may use Bayesian probability theory to examine how the knowledge of a small earthquake having occurred translates to our state of knowledge regarding the plausibility of a burglary. The narrative we will formally analyze is taken from [76]:

> Fred lives in Los Angeles and commutes 60 miles to work. Whilst at work, he receives a phone-call from his neighbor saying that Fred's burglar alarm is ringing. While driving home to investigate, Fred hears on the radio that there was a small earthquake that day near his home.

The propositions that will go in our Bayesian inference network are the following:

$$B = \text{Burglary},$$
$$\overline{B} = \text{No burglary},$$

$$A = \text{Alarm},$$
$$\overline{A} = \text{No alarm},$$

$$E = \text{Small earthquake},$$
$$\overline{E} = \text{No earthquake},$$

where we will distinguish between two prior states of knowledge:

I_1 = State of knowledge where hypothesis of earthquake is also entertained

I_2 = State of knowledge where hypothesis of earthquake is not entertained

We assume that the neighbor would never phone if the alarm is not ringing and that radio reports are fully trustworthy too. Furthermore, we assume that the occurrence of a small earthquake and a burglary are independent. We also

assume that a burglary alarm is almost certainly triggered by either a burglary
or a small earthquake or both, that is,

$$P\left(A\middle|\,B\overline{E}I_1\right) = P\left(A\middle|\,\overline{B}EI_1\right) = P\left(A\middle|\,BEI_1\right) \to 1, \tag{4.21}$$

whereas alarms in the absence of both a burglary and a small earthquake are
assumed to be extremely rare, that is,

$$P\left(A\middle|\,\overline{B}\,\overline{E}I_1\right) \to 0. \tag{4.22}$$

But if in our prior state of knowledge we do not entertain the possibility of an
earthquake, then (4.21) and (4.22) will, respectively, collapse to

$$P(A|\,BI_2) \to 1, \tag{4.23}$$

and

$$P\left(A\middle|\,\overline{B}I_2\right) \to 0. \tag{4.24}$$

Let

$$P(E) = e, \qquad P(B) = b. \tag{4.25}$$

Then we have, by way of the sum rule (4.2),

$$P\left(\overline{E}\right) = 1 - e, \qquad P\left(\overline{B}\right) = 1 - b. \tag{4.26}$$

If we are in a state of knowledge where we allow for an earthquake, we have, by
way of the product rule (4.1), as well as (4.21), (4.22), (4.25), and (4.26), that

$$P\left(AB\overline{E}\middle|\,I_1\right) = P\left(A\middle|\,B\overline{E}I_1\right) P(B)\,P\left(\overline{E}\right) \to b\,(1 - e)\,,$$

$$P\left(A\overline{B}E\middle|\,I_1\right) = P\left(A\middle|\,\overline{B}EI_1\right) P\left(\overline{B}\right) P(E) \to (1 - b)\,e,$$

$$\tag{4.27}$$

$$P(ABE|\,I_1) = P(A|\,BEI_1)\,P(B)\,P(E) \to be,$$

$$P\left(A\overline{B}\,\overline{E}\middle|\,I_1\right) = P\left(A\middle|\,\overline{B}\,\overline{E}I_1\right) P\left(\overline{B}\right) P\left(\overline{E}\right) \to 0.$$

By way of 'marginalization', that is, an application of the generalized sum rule,
(4.4), we obtain the probabilities

$$P\left(A\overline{B}\middle|\,I_1\right) = P\left(A\overline{B}\,\overline{E}\middle|\,I_1\right) + P\left(A\overline{B}E\middle|\,I_1\right) \to (1 - b)\,e,$$

$$P(AB|\,I_1) = P(ABE|\,I_1) + P\left(AB\overline{E}\middle|\,I_1\right) \to b, \tag{4.28}$$

$$P(A|\,I_1) = P(AB|\,I_1) + P\left(A\overline{B}\middle|\,I_1\right) \to b + e - be$$

and

$$P\left(A\overline{E}\middle|\,I_1\right) = P\left(AB\overline{E}\middle|\,I_1\right) + P\left(A\overline{B}\,\overline{E}\middle|\,I_1\right) \to b\,(1-e)\,,$$

$$\text{(4.29)}$$

$$P(\,AE|\,I_1) = P(\,ABE|\,I_1) + P\left(A\overline{B}E\middle|\,I_1\right) \to e.$$

But if we are in a state of knowledge where we do not allow for an earthquake, we have, by way of the product rule (4.1), as well as (4.23), (4.24), (4.25), and (4.26), that

$$P(\,AB|\,I_2) = P(\,A|\,BI_2)\,P(B) \to b,$$

$$\text{(4.30)}$$

$$P\left(A\overline{B}\middle|\,I_2\right) = P\left(A\middle|\,\overline{B}I_2\right) P\left(\overline{B}\right) \to 0,$$

By way of 'marginalization', that is, an application of the generalized sum rule, (4.4), we obtain the probability

$$P(\,A|\,I_2) = P(\,AB|\,I_2) + P\left(A\overline{B}\middle|\,I_2\right) \to b. \qquad \text{(4.31)}$$

The moment Fred hears that his burglary alarm is going off, then there are two possibilities. One possibility is that Fred may be new to Los Angeles and, consequently, overlook the possibility of a small earthquake triggering his burglary alarm, that is, his state of knowledge is I_2, which will make his prior probability of his alarm going off go to (4.31). Fred then assesses, by way of the product rule (4.1), (4.30) and (4.31), the likelihood of a burglary to be

$$P(\,B|\,AI_2) = \frac{P(\,AB|\,I_2)}{P(\,A|\,I_2)} \to \frac{b}{b} = 1, \qquad \text{(4.32)}$$

which leaves him greatly distressed, as he drives to his home to investigate.

Another possibility is that Fred is a veteran Los Angeleno and, as a consequence, instantly will take into account the hypothesis of a small tremor occurring near his house, that is, his state of knowledge is I_1. Fred then assesses, by way of the product rule (4.1) and (4.28), the likelihood of a burglary to be

$$P(\,B|\,AI_1) = \frac{P(\,AB|\,I_1)}{P(\,A|\,I_1)} \to \frac{b}{b+e-be} \approx \frac{b}{b+e}, \qquad \text{(4.33)}$$

seeing that $b+e >> be$. And if earthquakes are somewhat more common than burglaries, then Fred, based on his (4.32), may still hope for the best as he drives home to investigate, seeing that chances of a burglary will then be lower than fifty percent.

Either way, the moment that Fred hears on the radio that a small earthquake has occurred near his house, around the time when the burglary alarm went off, then, by way of the product rule (4.1), (4.27) and (4.29), Fred updates the likelihood of a burglary to be

$$P(\,B|\,AEI_1) = \frac{P(\,ABE|\,I_1)}{P(\,AE|\,I_1)} \to \frac{be}{e} = b. \qquad \text{(4.34)}$$

Stated differently, in the presence of an alternative explanation for the triggering of the burglary alarm, that is, a small earthquake occurring, the burglary alarm has lost its predictive power over the prior probability of a burglary, seeing that, (4.25) and (4.34),

$$P(B|\,AEI_1) = P(B)\,. \tag{4.35}$$

Consequently, Fred's fear for a burglary, as he rides home, after having heard that a small earthquake did occur, will only be dependent upon his assessment of the general likelihood of a burglary occurring. If we assume that Fred lives in a nice neighborhood, then we can imagine that Fred will be somewhat relieved after hearing the earthquake report on his radio.

4.3   Discussion

One of the arguments made against Bayesian probability theory as a normative model for human rationality is that people are generally numerical illiterate. Hence, the Bayesian model is deemed to be too numerical a model for human inference, [97]. However, it should be noted that the Bayesian analysis given here was purely qualitative, in that no actual numerical values were given to our probabilities, apart from (4.21), (4.23), (4.22), and (4.24), which are limit cases of certainty and, hence, in a sense, may also be considered to be qualitative. Moreover, the result of this qualitative analysis seems to be intuitive enough. And it is to be noted that the qualitative correspondence of the product and sum rules with common sense has been noted and demonstrated time and again by many researchers, including Laplace [73], Keynes [56], Jeffreys [48], Polya [84, 85], Cox [18], Tribus [100], de Finetti [20], Rosenkrantz [86], and Jaynes [47].

Chapter 5

Plausible Reasoning

If Bayesian probability theory is indeed common sense quantified, as we claim, then it should, at a very minimum, by commensurate with the formal rules of deductive and inductive logic [40, 47]. So, we now proceed to demonstrate how the Aristotelian syllogisms, may be derived by way of the rules of Bayesian probability theory. We then proceed to derive a new class of plausible syllogisms [47].

5.1 The Aristotelian Syllogisms

The rules of Bayesian probability theory are the product and sum rules [47]:

$$P(A)\,P(\,B|\,A) = P(AB) = P(B)\,P(\,A|\,B) \tag{5.1}$$

and

$$P\big(\overline{A}\big|\,B\big) = 1 - P(\,A|\,B)\,, \tag{5.2}$$

where AB is the proposition 'both A and B are true' and $\overline{A}$ is the proposition 'not-A is true'.

5.1.1 Strong Aristotelian Syllogisms

The strong syllogisms in Aristotelian logic correspond with the process of deduction. The first strong syllogism is

$$\begin{aligned} \text{Premise}: & \quad \text{If } A \text{ then also } B \\ \text{Observation}: & \quad A \\ \text{Conclusion}: & \quad \text{therefore } B \end{aligned} \tag{5.3}$$

Under the premise in (5.3), proposition AB is logically equivalent to the proposition A, that is, they have the same 'truth value':

$$A = AB. \tag{5.4}$$

The most primitive assumption of probability theory is that consistency demands that propositions which are logically equivalent, that is, have the same truth values, should be assigned equal plausibilities [47]. So, by way of (5.4), the premise of (5.3) translates to

$$P(A) = P(AB).\tag{5.5}$$

Because of the product rule, (5.1), we have

$$P(AB) = P(A)\,P(\,B|\,A)\,.\tag{5.6}$$

Substituting (5.5) into (5.6), it follows that after having observed A the proposition B has a probability 1 of being true, that is,

$$P(\,B|\,A) = 1.\tag{5.7}$$

The second strong syllogism is

$$\begin{aligned}\text{Premise}: \quad &\text{If } A \text{ then also } B\\ \text{Observation}: \quad &\overline{B}\\ \text{Conclusion}: \quad &\text{therefore } \overline{A}\end{aligned}\tag{5.8}$$

The premise in the second strong syllogism is the same as the premise in the first strong syllogism. Therefore, we may use the results of the first strong syllogism in the derivation of the second strong syllogism. From the sum rule and the first strong syllogism, (5.2) and (5.7), it follows that

$$P\big(\overline{B}\big|\,A\big) = 1 - P(\,B|\,A) = 0.\tag{5.9}$$

From the product rule, (5.1), we have

$$P(A)\,P\big(\overline{B}\big|\,A\big) = P\big(A\overline{B}\big) = P\big(\overline{B}\big)\,P\big(A|\,\overline{B}\big)\,.\tag{5.10}$$

From (5.9) and (5.10), it follows that, for $P\big(\overline{B}\big) > 0$,

$$P\big(A|\,\overline{B}\big) = P(A)\,\frac{P\big(\overline{B}\big|\,A\big)}{P\big(\overline{B}\big)} = 0.\tag{5.11}$$

Substituting (5.11) into the sum rule (5.2), we find that after having observed $\overline{B}$ the proposition $\overline{A}$ has a probability 1 of being true, that is,

$$P\big(\overline{A}\big|\,\overline{B}\big) = 1 - P\big(A|\,\overline{B}\big) = 1.\tag{5.12}$$

5.1.2   Weak Aristotelian Syllogisms

The weak syllogisms in Aristotelian logic correspond with the process of induction. The first weak Aristotelian syllogism is

$$\begin{aligned}\text{Premise}: \quad &\text{If } A \text{ then also } B\\ \text{Observation}: \quad &B\\ \text{Conclusion}: \quad &\text{therefore } A \text{ more plausible}\end{aligned}\tag{5.13}$$

From the product rule, (5.1), we have

$$P(A|\,B) = P(A)\,\frac{P(B|\,A)}{P(B)}. \tag{5.14}$$

Substituting (5.7) into (5.14), we find

$$P(A|\,B) = P(A)\,\frac{1}{P(B)}. \tag{5.15}$$

Excluding both absolute certainty and impossibility of B, we have that

$$0 < P(B) < 1. \tag{5.16}$$

From (5.16), it then follows that (5.15), translates to the inequality

$$P(A|\,B) > P(A). \tag{5.17}$$

In words, after having observed B the proposition A has become more probable.
The second weak Aristotelian syllogism is

$$\begin{aligned}
\text{Premise}: &\quad \text{If } A \text{ then also } B \\
\text{Observation}: &\quad \overline{A} \\
\text{Conclusion}: &\quad \text{therefore } B \text{ less plausible}
\end{aligned} \tag{5.18}$$

From the product rule (5.1) it follows that

$$P(\overline{A})\,P(B|\,\overline{A}) = P(\overline{A}B) = P(B)\,P(\overline{A}|\,B). \tag{5.19}$$

Rewriting (5.19), we get

$$\frac{P(\overline{A}|\,B)}{P(\overline{A})} = \frac{P(B|\,\overline{A})}{P(B)}. \tag{5.20}$$

By applying the sum rule (5.2) to (5.17), we find

$$P(\overline{A}|\,B) = 1 - P(A|\,B) < 1 - P(A) = P(\overline{A}). \tag{5.21}$$

Combining (5.21) with (5.20), we obtain the inequality

$$\frac{P(B|\,\overline{A})}{P(B)} = \frac{P(\overline{A}|\,B)}{P(\overline{A})} < 1, \tag{5.22}$$

or, equivalently,

$$P(B|\,\overline{A}) < P(B). \tag{5.23}$$

In words, after having observed $\overline{A}$ the proposition B has become less probable.
This concludes our derivation of the second weak syllogism of inductive logic.

5.2   The Plausibility Syllogisms

The four Aristotelian syllogisms all share the same certainty premise: 'If A then also B'. Now, in real life we are often forced to do our reasoning based incomplete information, that is, on a plausibility premises of the type: 'If A then B more plausible'. So, we now relax the certainty premise into a plausibility premise and derive the corresponding plausibility syllogisms by way of the product and sum rules. Then we show that as the plausibility premise tends to the certainty premise the plausibility syllogisms will tend, in some cases trivially, to their Aristotelian counterparts.

5.2.1   Strong Plausibility Syllogisms

The first strong plausibility syllogism is

$$
\begin{aligned}
\text{Premise}: &\quad \text{If } A \text{ then } B \text{ more plausible} \\
\text{Observation}: &\quad A \\
\text{Conclusion}: &\quad \text{therefore } B \text{ more plausible}
\end{aligned}
\tag{5.24}
$$

The plausibility premise in (5.24) translates to:

$$
P(B\,|\,A) > P(B)\,.
\tag{5.25}
$$

From this premise the conclusion in (5.24) follows trivially. This concludes our derivation of the first strong syllogism of plausible reasoning.

Note that, as we let the plausibility premise tend to the certainty premise, that is,

$$
P(B\,|\,A) \to 1,
\tag{5.26}
$$

then the conclusion of the first strong plausibility syllogism (5.24) tends, trivially, to the conclusion of the first strong Aristotelian syllogism (5.4).

The second strong plausibility syllogism is

$$
\begin{aligned}
\text{Premise}: &\quad \text{If } A \text{ then } B \text{ more plausible} \\
\text{Observation}: &\quad \overline{B} \\
\text{Conclusion}: &\quad \text{therefore } A \text{ less plausible}
\end{aligned}
\tag{5.27}
$$

From the sum rule (5.2) and the plausibility conclusion (5.25), we have

$$
P\!\left(\overline{B}\,\middle|\,A\right) = 1 - P(B\,|\,A) < 1 - P(B) = P\!\left(\overline{B}\right).
\tag{5.28}
$$

Multiplying (5.28) with $P(A)$ and applying the product rule (5.1), we obtain

$$
P\!\left(A\overline{B}\right) < P(A)\,P\!\left(\overline{B}\right).
\tag{5.29}
$$

Dividing (5.29) with $P\!\left(\overline{B}\right)$ and applying the product rule (5.1), we obtain

$$
P\!\left(A\,\middle|\,\overline{B}\right) < P(A)\,.
\tag{5.30}
$$

This concludes our derivation of the second strong syllogism of plausible reasoning.

We now will show how as the premise of the second strong plausibility syllogism tends in a limit of certainty to the premise of the second Aristotelian strong syllogism, the conclusion of the former will also tend to the conclusion of the latter:

$$
\begin{aligned}
\text{Premise}: \quad & B \text{ tends to certainty if we observe } A \\
\text{Observation}: \quad & \overline{B} \\
\text{Conclusion}: \quad & \text{therefore } \overline{A} \text{ tends to certainty}
\end{aligned}
\tag{5.31}
$$

The premise of syllogism (5.31) translates to (5.26)

$$
P(B|A) \to 1.
$$

Making use of the sum rule, (5.2), we have

$$
P(\overline{B}|A) = 1 - P(B|A) \to 0.
\tag{5.32}
$$

From the product rule (5.1), we have

$$
P(A|\overline{B}) = P(A) \frac{P(\overline{B}|A)}{P(\overline{B})}.
\tag{5.33}
$$

Because of (5.32), we have that equality (5.33), for $P(A) > 0$ and $P(\overline{B}) > 0$, tends to

$$
P(A|\overline{B}) \to 0.
\tag{5.34}
$$

Substituting (5.34) into the sum rule (5.2), we find

$$
P(\overline{A}|\overline{B}) = 1 - P(A|\overline{B}) \to 1.
\tag{5.35}
$$

It follows that from a plausibility premise we may approach the second strong Aristotelian syllogism in a limit of certainty. Seeing that the same holds, trivially, for the first strong Aristotelian syllogism, we have that all of deduction is just a specific limit case of plausible reasoning [47].

5.2.2   Weak Plausibility Syllogisms

The first weak plausibility syllogism is

$$
\begin{aligned}
\text{Premise}: \quad & \text{If } A \text{ then } B \text{ more plausible} \\
\text{Observation}: \quad & B \\
\text{Conclusion}: \quad & \text{therefore } A \text{ more plausible}
\end{aligned}
\tag{5.36}
$$

From the product rule (5.1), we have

$$
P(A|B) = P(A) \frac{P(B|A)}{P(B)}.
\tag{5.37}
$$

where, from the plausibility premise (5.25),

$$\frac{P(B|A)}{P(B)} > 1. \tag{5.38}$$

It follows, from (5.37) and (5.38), that

$$P(A|B) > P(A), \tag{5.39}$$

which is the conclusion of (5.36). This concludes our derivation of the first weak syllogism of plausible reasoning. Note that the first weak Aristotelian syllogism (5.13) is a special case of the more general first weak plausible syllogism[1] (5.36).

The second weak plausibility syllogism is

$$\begin{aligned}
\text{Premise}: &\quad \text{If } A \text{ then } B \text{ more plausible} \\
\text{Observation}: &\quad \overline{A} \\
\text{Conclusion}: &\quad \text{therefore } B \text{ less plausible}
\end{aligned} \tag{5.40}$$

By multiplying the plausibility premise (5.25) with $P(A)$ and by applying the product rule (5.1), we find

$$P(A)\,P(B|A) = P(AB) > P(A)\,P(B). \tag{5.41}$$

Dividing (5.41) with $P(B)$ and by applying the product rule (5.1), we obtain

$$P(A|B) = \frac{P(AB)}{P(B)} > P(A). \tag{5.42}$$

By way of the sum rule (5.2) and (5.42), it follows

$$P(\overline{A}|B) = 1 - P(A|B) < 1 - P(A) = P(\overline{A}). \tag{5.43}$$

Multiplying (5.43) with $P(B)$ and applying the product rule (5.1), we find

$$P(\overline{A}B) < P(B)\,P(\overline{A}). \tag{5.44}$$

Dividing (5.44) with $P(\overline{A})$ and applying the product rule (5.1), we obtain the desired inequality:

$$P(B|\overline{A}) < P(B). \tag{5.45}$$

This concludes our derivation of the second weak syllogism of plausible reasoning. And again we have that (5.18) is a special case of the more general plausible syllogism (5.40).

---

[1]As long as B is not always true irrespective of the truth value of A, that is, as long as $P(B) < 1$.

5.3 Discussion

The fact that Bayesian probability theory, that is, the product and sum rules, hold the Aristotelian syllogisms of deduction and induction as a special case leads Jaynes to the statement that Bayesian probability theory is an extension of logic [47]. And it is shown by Polya that even mathematicians will use the weaker forms of reasoning, that is, the plausibility syllogisms, (5.24), (5.27), (5.36), and (5.40), most of the time when still in the exploratory phase of their research [84, 85]. Only on publishing their new theorems will mathematicians try their hardest to invent an argument which uses only the strong Aristotelian syllogisms of deduction, (5.3) and (5.8). But the reasoning process which led to their theorems in the first place almost always involve the syllogisms of plausible reasoning.

Chapter 6

Bayesian Data Analysis

In the previous chapters we have focused on Bayesian probability theory as a general model of inference. But Bayesian probability theory, as such, is also a data analysis tool. In order to demonstrate this point, we will now proceed to give both the orthodox and the Bayesian derivation of the Student-t distribution. Also, by putting the orthodox and Bayesian derivations side by side we allow the interested reader to get some sense for the differences between the orthodox and Bayesian approaches. Moreover, the Bayesian derivation of the Student-t distribution may serve as a low-level introduction into the practice of Bayesian data analysis.

In the following we will leave in all the intermediate derivation steps, so that one may get a general sense of these steps as one reads along.

6.1   The Orthodox Approach

In what follows we will give a short outline on how orthodox statistics comes to the Student-t distribution. We will leave out the actual proofs of the first three preliminary theorems and their corollaries .

Preliminary Theorem 1 *If $X_1, \ldots, X_n$, are observations of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$, then the distribution of the sample mean*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{6.1}$$

is $N(\mu, \sigma^2/n)$, or, equivalently, the statistic $\overline{X}$ admits the probability distribution:

$$p(\overline{X} \,|\, \mu, \sigma, n) = \frac{n}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n}{2\sigma^2} \left(\overline{X} - \mu\right)^2\right]. \tag{6.2}$$

A corollary of this theorem is that the statistic

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}. \tag{6.3}$$

is standard normal distributed $N(0,1)$, or, equivalently, Z admits the probability distribution:

$$p(Z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z^2}{2}\right). \tag{6.4}$$

Preliminary Theorem 2 *If $X_1, \ldots, X_n$, are observations of a random sample of size n from the normal distribution $N\left(\mu, \sigma^2/n\right)$ and*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2, \tag{6.5}$$

then the statistic

$$U = \frac{(n-1)\,S^2}{\sigma^2} \tag{6.6}$$

is chi-squared distributed $\chi^2(n-1)$, or, equivalently,

$$p(U \,|\, n) = \frac{1}{2^{(n-1)/2}\, \Gamma[(n-1)\,/2]}\, U^{(n-1)/2-1} \exp\left(-\frac{U}{2}\right), \tag{6.7}$$

where Γ is the gamma function.

Preliminary Theorem 3 *If $X_1, \ldots, X_n$, are observations of a random sample of size n from the normal distribution $N\left(\mu, \sigma^2/n\right)$, then the statistics $\overline{X}$ and S^2, (6.1) and (6.5), are independently distributed. A corollary of this theorem is that the statistics Z and U, (6.3) and (6.6), are independently distributed.*

Note that the first two preliminary theorems have proofs that take up a couple of pages in [37], while the proof of the third preliminary theorem is not given, as this latter proof is deemed to be too involved for an introduction text on statistics.

   With these three preliminary theorems in hand we can now proof the Student-t distribution theorem.

Theorem 1 *If $X_1, \ldots, X_n$, are observations of a random sample of size n from the normal distribution $N\left(\mu, \sigma^2/n\right)$, then the statistic*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}. \tag{6.8}$$

has a Student-t distribution with $n-1$ degrees of freedom, or, equivalently,

$$p(T \,|\, n) = \frac{\Gamma(n/2)}{\sqrt{n-1}\,\Gamma[(n-1)\,/2]\,\Gamma(1/2)} \left(1 + \frac{T^2}{n-1}\right)^{-n/2}. \tag{6.9}$$

Proof. The statistic (6.8) may be rewritten in terms of the statistics (6.3) and (6.6):

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

$$= \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \frac{\sigma}{S}$$

$$= \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{(n-1)\,\sigma^2}{(n-1)\,S^2}} \tag{6.10}$$

$$= Z \sqrt{\frac{n-1}{U}}$$

$$= \frac{Z}{\sqrt{U/(n-1)}}.$$

Because of the corollary of the third preliminary theorem, we have that that the distribution of both Z and U may be factored as, (6.4) and (6.7),

$$p(Z, U \,|\, n) = p(Z)\, p(U \,|\, \sigma, n)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z^2}{2}\right) \frac{U^{(n-1)/2-1}}{2^{(n-1)/2-1}\,\Gamma[(n-1)/2]} \exp\left(-\frac{U}{2}\right)$$

$$= \frac{U^{(n-1)/2-1}}{2^{n/2}\,\Gamma(1/2)\,\Gamma[(n-1)/2]} \exp\left(-\frac{Z^2}{2} - \frac{U}{2}\right), \tag{6.11}$$

where we have made use of the identity $\Gamma(1/2) = \sqrt{\pi}$. Because of (6.10), we may make the transformations

$$T = \frac{Z}{\sqrt{U/(n-1)}} \qquad \text{and} \qquad U' = U, \tag{6.12}$$

or, equivalently,

$$Z = T\sqrt{U'/(n-1)} \qquad \text{and} \qquad U = U'. \tag{6.13}$$

The corresponding Jacobian is

$$|J| = \begin{vmatrix} \frac{\partial}{\partial T}Z & \frac{\partial}{\partial U'}Z \\ \frac{\partial}{\partial T}U & \frac{\partial}{\partial U'}U \end{vmatrix} = \begin{vmatrix} \sqrt{\frac{U'}{(n-1)}} & \frac{T}{2\sqrt{U'/(n-1)}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{U'}{(n-1)}}. \tag{6.14}$$

Substituting (6.13) into (6.11) and multiplying with the Jacobian (6.14), we obtain the transformed bivariate distribution:

$$|J|\ p(T, U'|\, n) = \frac{\sqrt{U'/\,(n-1)}\ U'^{(n-1)/2-1}}{2^{n/2}\,\Gamma(1/2)\ \Gamma[(n-1)\,/2]}\ \exp\!\left(-\frac{T^2\,U'}{2\,(n-1)} - \frac{U'}{2}\right)$$

(6.15)

$$= \frac{U'^{n/2-1}}{2^{n/2}\,\sqrt{n-1}\,\Gamma(1/2)\ \Gamma[(n-1)\,/2]}\ \exp\!\left[-\frac{U'}{2}\left(1 + \frac{T^2}{n-1}\right)\right].$$

If we integrate out U' of (6.15), then we obtain the probability distribution of the T statistic (6.8) which is the Student-t distribution (6.9):

$$p(T|\, n) = \int_0^\infty |J|\ p(T, U'|\, \sigma, n)\, dU'$$

$$= \int_0^\infty \frac{U'^{n/2-1}}{2^{n/2}\,\sqrt{n-1}\,\Gamma(1/2)\ \Gamma[(n-1)\,/2]}\ \exp\!\left[-\frac{U'}{2}\left(1 + \frac{T2}{n-1}\right)\right] dU'$$

$$= \frac{\Gamma(n/2)}{\sqrt{n-1}\,\Gamma(1/2)\ \Gamma[(n-1)\,/2]}\left(1 + \frac{T^2}{n-1}\right)^{-n/2}.$$

(6.16)

■

Now, if we substitute (6.8) into (6.16) and multiply (6.16) with the differential which corresponds with the change of variable from the T statistic to the unknown parameter μ:

$$\frac{dT}{d\mu} = \left|\frac{d}{d\mu}\frac{\overline{X} - \mu}{S/\sqrt{n}}\right| = \frac{\sqrt{n}}{S},$$

(6.17)

or, equivalently,

$$dT = \frac{\sqrt{n}}{S}\, d\mu,$$

(6.18)

then we obtain the probability distribution for the unknown parameter μ, as implied by the Student-t distribution (6.9):

$$p\big(\mu|\,\overline{X}, S, n\big)\, d\mu = \frac{\Gamma(n/2)}{\Gamma[(n-1)\,/2]\,\Gamma(1/2)}\sqrt{\frac{n}{(n-1)\,S^2}}\left[1 + \frac{n\,(\mu - \overline{X})^2}{(n-1)\,S^2}\right]^{-n/2} d\mu.$$

(6.19)

And we say 'implied', because orthodox statistics, on ideological grounds, for lack of a better description, only allows for probability distributions of statistics like $\overline{X}$, U, and T, (6.1), (6.6), and (6.8), respectively. But it does not allow for probability distributions of unknown parameters like μ, even if the Jacobian transformation from (6.16) to (6.19) is perfectly legitimate from a purely mathematical point of view.

So the best an orthodox statistician can do is to compute the confidence bound percentiles, say, $t_{0.025}$ and $t_{0.975}$ from the standardized Student-t distribution (6.9) for the given degrees of freedom $n - 1$, and then compute the 95%-confidence interval for the unknown constant μ:

$$95\% \text{ CI} = \left[\overline{X} + t_{0.025} \frac{S}{\sqrt{n}}, \overline{X} + t_{0.975} \frac{S}{\sqrt{n}} \right]. \tag{6.20}$$

where $t_{0.50} = 0$, and percentiles below and above this median are negative and positive, respectively.

In regards to the interpretation of the confidence interval (6.20), it is to be noted that in orthodox statistics probability statements can be made only about random variables. So, from this perspective, it is meaningless to speak of the probability that μ lies in a certain interval, because μ is not a random variable, but only an unknown constant. And it is on these grounds that it is held by orthodox statistics to be very important that we use the words, "the probability that the interval covers the true value of μ", rather than "the probability that the true value lies in the interval" [44].

6.2 The Bayesian Approach

In what follows we will give the Bayesian derivation of the Student-t distribution. But first we give a short outline on the algorithmic steps of any Bayesian data-analysis and, by so doing, introduce the reader to the four central constructs of Bayesian probability theory.

In Bayesian probability theory one first has to *assign* a **likelihood function** and a **prior distribution** to the set of unknown parameters $\{\theta\}$. One then combines the likelihood and the prior, by way of the product rule, in order to get the probability distribution of both the data D and the parameters $\{\theta\}$. By integrating out the parameters $\{\theta\}$ out of this probability distribution one may *compute* the **evidence**, which is both a normalizing constant as well as a scalar which is of paramount importance in Bayesian model selection, and proceed to *construct* the **posterior distribution**.

1. The likelihood (function) of the unknown parameters $L(\{\theta\})$ links both the data D and the expert judgment with the set of unknown parameters $\{\theta\}$. It may be helpful to realize that the likelihood is just $p(D|\{\theta\})$, that is, the probability of the data D, conditional on the unknown parameters $\{\theta\}$.

2. The prior (distribution) $p(\{\theta\}|I)$ is the probability distribution of the unknown parameters $\{\theta\}$, conditional on our background information I. The prior links the expert judgment with the set of unknown parameters and expresses our state of knowledge regarding $\{\theta\}$ independent of the data D. If we are in a state of ignorance about the true values of $\{\theta\}$, then our prior $p(\{\theta\}|I)$ should express this ignorance.

3. The evidence $p(D|I)$ is a normalizing constant which also may be used for Bayesian model selection. In order to compute the evidence, first the likelihood and prior have to be combined by way of the product rule:

$$p(D, \{\theta\}|I) = p(\{\theta\}|I)\, p(D|\{\theta\})\,, \qquad (6.21)$$

then the by integrating out the $\{\theta\}$, by way of the sum rule, as summation goes to into integration in a limit of infinitely many propositions, we obtain the evidence:

$$p(D|I) = \int p(D, \{\theta\}|I)\, d\,\{\theta\}\,. \qquad (6.22)$$

Note that the evidence has a built-in Occam's razor, which rewards the goodness of fit and at the same time penalizes the size of the prior parameter space; the larger the size of the prior parameter space the better the goodness of fit but also the larger the size of the penalty [107].

4. The posterior (distribution) $p(\{\theta\}|D, I)$ is the probability distribution of the unknown parameters $\{\theta\}$, conditional on the observed data D and our background information I. The posterior is constructed by combining (6.21) and (6.22), by way of the product rule:

$$p(\{\theta\}|D, I) = \frac{p(D, \{\theta\}|I)}{p(D|I)} = p(\{\theta\}|I)\,\frac{p(D|\{\theta\})}{p(D|I)}. \qquad (6.23)$$

The posterior is the general solution to the problem of *inverse probabilities*, which has as its specific solution the beta distribution. This specific solution was given by Bayes in 1763, who derived the beta distribution as the posterior that results from a binomial likelihood and a constant prior for the probability p of a success. However, it was Laplace who in his memoir of 1774 perceived the general principle behind Bayes' specific solution of the problem of inverse probabilities and who proposed "Bayes' theorem" (6.23) as a general principle of inference [45].

We now proceed to take a more in-depth look into these four constructs of the Bayesian data analysis.

6.2.1   The Likelihood Model

We assume as an initial model that a given data point x_i is generated by a constant signal μ plus some white noise e_i:

$$x_i = \mu + e_i. \qquad (6.24)$$

The white noise is assumed to be normally distributed with a mean of zero and a spread of σ:

$$p(e_i|\sigma)\, de_i = \frac{1}{\sqrt{2\pi}\sigma}\, \exp\left(-\frac{e_i^2}{2\sigma^2}\right) de_i. \qquad (6.25)$$

We may rewrite (6.24) as

$$e_i = x_i - \mu. \tag{6.26}$$

It then follows that

$$de_i = dx_i. \tag{6.27}$$

So we may substitute (6.26) and (6.27) into (6.25), and so obtain the probability of a given data point:

$$p(x_i | \mu, \sigma) \, dx_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right] dx_i. \tag{6.28}$$

If we have n data points which are generated by the same process, independently from each other, then we have that the probability of the observed data set,

$$D = (x_1, \ldots, x_n). \tag{6.29}$$

given the unknown parameters μ and σ, or, equivalently, the likelihood L of μ and σ, is given as:

$$L(\mu, \sigma) = p(D | \mu, \sigma)$$

$$= p(x_1, \ldots, x_n | \mu, \sigma)$$

$$= \prod_{i=1}^{n} p(x_i | \mu, \sigma) \tag{6.30}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

Let, (6.1) and (6.5),

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{6.31}$$

Then we may rewrite the exponential in (6.30) as

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\mu n\bar{x} + n\mu^2 + n\bar{x}^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n\left(\mu^2 - 2\mu\bar{x} + \bar{x}^2\right) \qquad (6.32)$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\left(\mu - \bar{x}\right)^2$$

$$= (n-1)\,s^2 + n\left(\mu - \bar{x}\right)^2$$

Substituting (6.32) into (6.30), we obtain the compact likelihood, which takes into account the data D by way of the sufficient statistics $\bar{x}$ and s^2, (6.31), and the sample size n:

$$L(\mu, \sigma) = p(D|\,\mu, \sigma)$$

$$\qquad (6.33)$$

$$= \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \left[(n-1)\,s^2 + n\left(\mu - \bar{x}\right)^2 \right] \right\}.$$

Note that in the Bayesian analysis sufficient statistics like (6.31), or, equivalently, (6.1) and (6.5), flow forth naturally from likelihoods like, for example, (6.30).

6.2.2   The Prior Model

The unknown parameters are the strength of constant signal μ and the spread of the white noise σ. The parameter μ is a *location parameter*, as are all regression coefficients, and the parameter σ is a *scale parameter*.

The uninformative prior for a location parameter is the uniform distribution [42]:

$$p(\mu|\,I) = C_\mu, \qquad \text{or, equivalently,} \qquad p(\mu|\,I) \propto \text{constant}, \qquad (6.34)$$

where C_μ is the normalizing constant of the uniform prior and '$\propto$' is the proportionality sign, which absorbs any constant which is not dependent upon the parameter of interest μ and which allows for a Bayesian short-hand for those

who are familiar with the Bayesian algebra. The uninformative prior for a scale parameter is the Jeffreys' prior [42]:

$$p(\sigma|I) = \frac{C_\sigma}{\sigma}, \qquad \text{or, equivalently,} \qquad p(\sigma|I) \propto \frac{1}{\sigma}, \qquad (6.35)$$

where C_σ is the normalizing constant of the Jeffreys' prior. Assuming logical independence between μ and σ, we obtain, by way of the product rule, the following prior model for both μ and σ.

$$p(\mu,\sigma|I) = \frac{C_\mu C_\sigma}{\sigma}, \qquad \text{or, equivalently,} \qquad p(\mu,\sigma|I) \propto \frac{1}{\sigma}. \qquad (6.36)$$

To make the uninformative priors (6.34) and (6.35) intuitive, we will proceed to give consistency arguments of the kind which are so typical for the Bayesian paradigm.

The Jeffreys' Prior for Location Parameters

Suppose that under our first state of ignorance I_1 we assign, for some given coordinate system, the prior distribution $p(\mu|I_1)$ to the location parameter μ. Then we are informed that the origin of our initially assumed coordinate system is actually lying c units to the left, *but nothing more*. Under this new state of ignorance I_2 we may assign and updated prior distribution $p(\mu'|I_2)$, where μ and μ' both point to different coordinate systems that express the same state of ignorance.

The unknown parameters μ and μ' are mathematically related as follows:

$$\mu' = \mu + c \qquad \text{and} \qquad d\mu' = d\mu. \qquad (6.37)$$

Since we are equally ignorant about μ under I_1 as we are about μ' under I_2, *consistency* demands that the following functional equation should hold [42]:

$$p(\mu|I_1)\,d\mu = p(\mu'|I_2)\,d\mu' = p(\mu+c|I_2)\,d\mu, \qquad (6.38)$$

whose general solution is given by (6.34), as only for the constant function $f(x) = c$ one will have that $f(x_1) = f(x_2)$, for general arguments $x_1 \neq x_2$.

The Jeffreys' Prior for Scale Parameters

Suppose that under our first state of ignorance I_1 we assign, for some given coordinate system, the prior distribution $p(\sigma|I_1)$ to the scale parameter σ. Then we are informed that the scale of our initially assumed coordinate system is actually off by a factor c, *but nothing more*. Under this new state of ignorance I_2 we may assign an updated prior distribution $p(\sigma'|I_2)$, where σ and σ' both point to different coordinate systems that express the same state of ignorance.

The unknown parameters σ and σ' are mathematically related as follows:

$$\sigma' = c\,\sigma \qquad \text{and} \qquad d\sigma' = c\,d\sigma. \qquad (6.39)$$

Since we are equally ignorant about σ under I_1 as we are about σ' under I_2, consistency demands that the following functional equation should hold [42]:

$$p(\sigma|I_1)\,d\sigma = p(\sigma'|I_2)\,d\sigma' = p(c\,\sigma|I_2)\,c\,d\sigma, \qquad (6.40)$$

whose general solution is given by (6.35). And it may be checked that, (6.35) and (6.39),

$$p(\sigma|I_1)\,d\sigma \propto \frac{d\sigma}{\sigma} = \frac{c\,d\sigma}{c\,\sigma} = \frac{d\sigma'}{\sigma'} \propto p(\sigma'|I_2)\,d\sigma'. \qquad (6.41)$$

6.2.3   The Evidence

Combining the likelihood (6.33) with the prior (6.36), by way of the product rule, we obtain the probability distribution:

$$p(D,\mu,\sigma|I) = p(\mu,\sigma|I)\,L(\mu,\sigma)$$

$$= p(\mu,\sigma|I)\,p(D|\mu,\sigma) \qquad (6.42)$$

$$= \frac{C_\mu C_\sigma}{(2\pi)^{n/2}\,\sigma^{n+1}}\,\exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\,(\mu-\bar{x})^2\right]\right\}.$$

Integrating out the unknown parameters μ and σ, we obtain the evidence [112, 107]:

$$p(D|I) = \int\int p(D,\mu,\sigma|I)\,d\mu\,d\sigma$$

$$= \int_0^\infty \int_{-\infty}^\infty \frac{C_\mu C_\sigma}{(2\pi)^{n/2}\,\sigma^{n+1}}\,\exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\,(\mu-\bar{x})^2\right]\right\}d\mu\,d\sigma$$

$$= \frac{C_\mu C_\sigma}{\sqrt{n}\,(2\pi)^{(n-1)/2}}\int_0^\infty \frac{1}{\sigma^n}\,\exp\left[-\frac{(n-1)\,s^2}{2\sigma^2}\right]\left\{\int_{-\infty}^\infty \frac{\sqrt{n}}{(2\pi)^{1/2}\,\sigma}\,\exp\left[-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right]d\mu\right\}d\sigma$$

$$= \frac{C_\mu C_\sigma}{\sqrt{n}\,(2\pi)^{(n-1)/2}}\int_0^\infty \frac{1}{\sigma^n}\,\exp\left[-\frac{(n-1)\,s^2}{2\sigma^2}\right]d\sigma \qquad (6.43)$$

$$= \frac{C_\mu C_\sigma}{\sqrt{n}\,\pi^{(n-1)/2}}\,\frac{\Gamma[(n-1)/2]}{2}\left[(n-1)\,s^2\right]^{-(n-1)/2}$$

Note that the evidence is both the marginal probability of the data D, given the likelihood and prior models, (6.33) and (6.36), as well as the term which will transform (6.42) into the bivariate posterior of μ and σ.

6.2.4 The Posterior

By way of (6.42), (6.43), and the product rule (4.1), we obtain the bivariate posterior:

$$p(\mu, \sigma \mid D, I) = \frac{p(D, \mu, \sigma \mid I)}{p(D \mid I)}$$

$$= \frac{2\,\pi^{(n-1)/2} \sqrt{n}\, \left[(n-1)\,s^2\right]^{(n-1)/2}}{C_\mu C_\sigma\, \Gamma[(n-1)/2]} \frac{C_\mu C_\sigma}{(2\pi)^{n/2}\,\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\,(\mu - \bar{x})^2\right]\right\}$$

$$= \frac{\sqrt{n}\, \left[(n-1)\,s^2\right]^{(n-1)/2}}{2^{(n-2)/2}\, \Gamma[(n-1)/2]\,\sqrt{\pi}} \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\,(\mu - \bar{x})^2\right]\right\}.$$

$$(6.44)$$

Integrating σ out of (6.44), we obtain the marginalized posterior distribution of μ [112]:

$$p(\mu \mid D, I) = \int_0^\infty p(\mu, \sigma \mid D, I)\, d\sigma$$

$$= \frac{\sqrt{n}\, \left[(n-1)\,s^2\right]^{(n-1)/2}}{2^{(n-2)/2}\, \Gamma[(n-1)/2]\,\sqrt{\pi}} \int_0^\infty \frac{1}{\sigma^{n+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\,(\mu - \bar{x})^2\right]\right\} d\sigma$$

$$= \frac{\sqrt{n}\, \left[(n-1)\,s^2\right]^{(n-1)/2}}{2^{(n-2)/2}\, \Gamma[(n-1)/2]\,\sqrt{\pi}} \frac{\Gamma(n/2)}{2} \frac{2^{n/2}}{\left[(n-1)\,s^2 + n\,(\mu - \bar{x})^2\right]^{n/2}}$$

$$= \frac{\sqrt{n}\, \Gamma(n/2)}{\Gamma[(n-1)/2]\, \Gamma(1/2)} \frac{\left[(n-1)\,s^2\right]^{(n-1)/2}}{\left[(n-1)\,s^2 + n\,(\mu - \bar{x})^2\right]^{n/2}}$$

$$= \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]\, \Gamma(1/2)} \sqrt{\frac{n}{(n-1)\,s^2}} \left[1 + \frac{n\,(\mu - \bar{x})^2}{(n-1)\,s^2}\right]^{-n/2} \qquad (6.45)$$

where we have used the identity $\sqrt{\pi} = \Gamma(1/2)$. In the Bayesian (6.45) we then have the general Student-t distribution (6.19) which is implied by (6.16).

6.3   Discussion

The Bayesian approach is highly algorithmic in nature: (i) assign your likelihood and prior, (ii) compute the evidence and posterior. In stark contrast we have the orthodox approach, where one has to guess beforehand the pertinent statistics whose probability distributions may provide the necessary building blocks, (6.4) and (6.7), to come to the probability distribution of interest, (6.16) or (6.19). Stated differently, the Bayesian algorithm side-steps in its derivation of the Student-t distribution the *ad hoc* step of choosing a statistic and the ensuing mathematical problem of finding its sampling distribution [44]. Moreover, as is demonstrated in [13], if no sufficient statistics like $\overline{X}$ and S^2, (6.1) and (6.5), are to be had, then the orthodox approach will come to a grinding halt, whereas the Bayesian approach just keeps on going [13].

Chapter 7

The Ellsberg Paradox

Ellsberg found that the willingness to bet on an uncertain event depends not only on the degree of uncertainty but also on its source. He observed that people prefer to bet on an urn containing equal numbers of red and black balls, rather than on an urn that contains red and black balls in unknown proportions. This phenomenon of source dependence is seen as evidence that people are not probabilistically sophisticated. It will be argued in this chapter that source dependence does not necessarily imply a lack of probabilistic sophistication. Also, it is stated by Ellsberg that the unknown quantities in the Ellsberg urns represent examples of non-quantifiable Knightian uncertainty, whereas the known quantities are examples of probability [24]. It will be shown in this chapter that Bayesian probability theory may be used to come to probability distributions of the number of colored balls in the Ellsberg urns with unknown quantities.

7.1   The First Ellsberg Experiment

Ellsberg's first urn experiment involves the two urns [24]:

- Urn 1 - 100 balls, 50 red, 50 black,

- Urn 2 - 100 balls, red and black with proportions not specified,

with payoffs defined as:

I. "Payoff on Red1": Draw from Urn 1, receive $100 if Red, $0 if Black,

II. "Payoff on Black1": Draw from Urn 1, receive $100 if Black, $0 if Red,

III. "Payoff on Red2": Draw from Urn 2, receive $100 if Red, $0 if Black,

IV. "Payoff on Black2": Draw from Urn 2, receive $100 if Black, $0 if Red.

The gambles posed are:

1. Which do you prefer: "Payoff on Red1" vs. "Payoff on Black1"?,

2. Which do you prefer: "Payoff on Red2" vs. "Payoff on Black2"?,

3. Which do you prefer: "Payoff on Red1" vs. "Payoff on Red2"?,

4. Which do you prefer: "Payoff on Black1" vs. "Payoff on Black2"?

Results according to Ellsberg, from introspection and non-experimental surveying of colleagues, are:

a. Majority will be indifferent in gambles (1) and (2), which indicates that subjective probabilities of Red and Black are 1/2 for both Urns 1 and 2.

b. Majority prefers Red1 in (3) and Black1 in (4) - in other words most people prefer Urn 1 (known fifty-fifty split) over Urn 2 (unknown split between red and black).

The Ellsberg results (a) and (b) are paradoxical if we interpret the preference for Urn 1 over Urn 2 in the gambles (3) and (4) to mean that the probabilities for Red1 and Black1 are greater than the corresponding probabilities Red2 and Black2. Since in such an interpretation observation (b) is in contradiction with observation (a), which states that the probabilities of Red1, Red2, Black1, and Black2 are all 1/2.

This paradox, however, may be trivially resolved if one realizes that *all things being equal*, probability-wise, one is perfectly free to prefer one Ellsberg urn over the other without being inconsistent. Stated differently, the inconsistency in the Ellsberg paradox lies not with the experimental subjects, but, rather, with the interpretation that the preference for the first Ellsberg urn necessarily imply subjective probabilities for Red2 and Black2 which are smaller than the corresponding known probabilities of 1/2 of Red1 and Black1. And it should be noted that this interpretation is also problematic in that it implies a sum of subjective probabilities of Red2 and Black2 less than 1, or, equivalently, a perception that the occurrence of either Red2 or Black2 is not a certainty should we draw from the second Ellsberg urn.

If we witness a clear preference for the urn with known proportions over the urn with unknown proportions, then we may conclude that people prefer gambles with crisp and clear probabilities to gambles with "fuzzy" probabilities, even if these probabilities are ultimately equivalent. In this conclusion there is nothing paradoxical. And it will materialize in the next sections that this preference for the urn with known proportions corresponds with the least amount of uncertainty should we decide to again partake in the Ellsberg bet.

7.2   An Analysis of the First Ellsberg Experiment

We now proceed to give a probability theoretical analysis of the pay-offs for n draws with replacement from the respective Ellsberg urns. This will allow us to

formally demonstrate the equivalency between these urns for $n = 1$, as in the Ellsberg experiment.

The probability distribution of r, the number of red balls, say, in n draws with replacement, for an urn having N balls of which R are red, is given by the binomial distribution

$$p(r \,|\, n, R, N) = \frac{n!}{r! \,(n - r)!} \left(\frac{R}{N}\right)^r \left(\frac{N - R}{N}\right)^{n-r}. \tag{7.1}$$

So the expected value and standard deviation of the payoff of r 100-dollar bills are given as

$$E(r \,|\, n, R, N) = \sum_{r=0}^{n} r \, p(r \,|\, n, R, N) \tag{7.2}$$

$$= n\frac{R}{N}$$

and

$$\mathrm{std}(r \,|\, n, R, N) = \sqrt{n\frac{R\,(N - R)}{N^2}}, \tag{7.3}$$

since we have that

$$\mathrm{var}(r \,|\, n, R, N) = \sum_{r=0}^{n} \left[r - E(r \,|\, n, R, N)\right]^2 p(r \,|\, n, R, N)$$

$$= \sum_{r=0}^{n} \left(r - n\frac{R}{N}\right)^2 p(r \,|\, n, R, N) \tag{7.4}$$

$$= n\frac{R\,(N - R)}{N^2}.$$

where both expectation value and standard deviation are understood[1] to be in units of \$100.

If the number of red balls R is unspecified, the Bayesian thing to do is to weigh the probability of drawing r red balls over all plausible values of R in (7.1). Based on the available background information I that the number of R must lie somewhere in the range from 0 to N, we assign as an uninformative prior the uniform probability distribution to the unknown number of red balls in the urn [47]:

$$p(R \,|\, I) = \frac{1}{N + 1}, \tag{7.5}$$

---

[1]If we compute the expected value $E(100r)$, rather than $E(r)$, and the standard deviation $\mathrm{std}(100r)$, rather than $\mathrm{std}(r)$, the units will be in dollars. But this will greatly increase the clutter in the derivations that will follow, which is why we chose to derive our results in 100-dollar bill units.

where $R = 0, 1, \ldots, N$. By way of the product and the generalized sum rules, (4.1) and (4.4), the probability of drawing r red balls from an urn having N balls of which an unspecified number are red translates to, (7.1) and (7.5),

$$p(r \mid n, N, I) = \sum_{R=0}^{N} p(r, R \mid n, N, I)$$

$$= \sum_{R=0}^{N} p(R \mid I) \, p(r \mid n, R, N) \tag{7.6}$$

$$= \sum_{R=0}^{N} \frac{1}{N+1} \frac{n!}{r! \, (n-r)!} \left(\frac{R}{N}\right)^{r} \left(\frac{N-R}{N}\right)^{n-r} .$$

The expectation value and standard deviation of (7.6) are given as

$$E(r \mid n, N, I) = \sum_{r=0}^{n} r \, p(r \mid n, N, I)$$

$$= \frac{1}{N+1} \sum_{R=0}^{N} \sum_{r=0}^{n} r \, \frac{n!}{r! \, (n-r)!} \left(\frac{R}{N}\right)^{r} \left(\frac{N-R}{N}\right)^{n-r}$$

$$\tag{7.7}$$

$$= \frac{1}{N+1} \sum_{R=0}^{N} n \frac{R}{N}$$

$$= \frac{n}{2}$$

and

$$\mathrm{std}(r \mid n, N, I) = \sqrt{n \frac{2N-2}{12N} + n^{2} \frac{N+2}{12N}} ,$$

since we have that

$$\mathrm{var}(r \,|\, n, N, I) = \sum_{r=0}^{n} [r - E(r \,|\, n, N, I)]^2 \, p(r \,|\, n, N, I)$$

$$= \sum_{r=0}^{n} \left(r - \frac{n}{2}\right)^2 p(r \,|\, n, N, I)$$

$$= \frac{1}{N+1} \sum_{R=0}^{N} \left[n \frac{R\,(N-R)}{N^2} + n^2 \frac{N - 2R}{4N^2} \right] \qquad (7.8)$$

$$= n \frac{2N - 2}{12N} + n^2 \frac{N + 2}{12N},$$

and where both expectation value and standard deviation are understood to be in units of \$100.

For $n = 1$, $N = 100$, $R = 50$, we have that the probabilities of the first and second urn reduce to, respectively, (7.1),

$$p(r = 0 \,|\, n = 1, R = 50, N = 100) = \frac{1}{2}$$

$$p(r = 1 \,|\, n = 1, R = 50, N = 100) = \frac{1}{2} \qquad (7.9)$$

and, for any N, (7.6),

$$p(r = 0 \,|\, n = 1, N, I) = \frac{1}{N\,(N+1)} \sum_{R=0}^{N} (N - R) = \frac{1}{2}$$

$$p(r = 1 \,|\, n = 1, N, I) = \frac{1}{N\,(N+1)} \sum_{R=0}^{N} R = \frac{1}{2}. \qquad (7.10)$$

So, for the special case of $n = 1$ both Ellsberg urns are indeed equivalent probability-wise.

It follows that the Ellsberg observations (a) and (b) of the previous section would have been paradoxical only for the alternative sets of urns:

- Urn 1 - 100 balls, red > 50,

- Urn 2 - 100 balls, red and black with proportions not specified,

Since only for these sets of urns do we have that a preference of Red2 over Red1 and Black1 over Black2 will work against the taker of the bet.

Note that the equivalency of the Ellsberg urns does not hold for repeated bets (i.e., $n > 1$). The standard deviation of the payout of the second Ellsberg

urn (7.8) will exceed the standard deviation of the payout from the first Ellsberg urn (7.3) for $n > 1$, whereas the expectations values of the payouts of both urns remain identical to each other, (7.2) and (7.7).

So, all things being equal probability-wise, there is an observed preference for the urn which in the repeated-draw case results in the least uncertainty.

7.3   The Second Ellsberg Experiment

Ellsberg's second urn experiment involves just the one urn [24]:

- Urn - 90 balls, 30 red, and 60 some combination of black and yellow with proportions not specified,

with payoffs defined as:

    I. "Payoff on Red": Receive $100 if Red, $0 if Black or Yellow,

   II. "Payoff on Black": Receive $100 if Black, $0 if Red or Yellow,

  III. "Payoff on Red or Yellow": Receive $100 if Red or Yellow, $0 if Black,

  IV. "Payoff on Black or Yellow": Receive $100 if Black or Yellow, $0 if Red.

The gambles posed are:

1. Which do you prefer: "Payoff on Red" vs. "Payoff on Black" (i.e., I vs. II)?,

2. Which do you prefer: "Payoff on Red or Yellow" vs. "Payoff on Black or Yellow" (i.e., III vs. IV)?

Results according to Ellsberg, from introspection and non-experimental surveying of colleagues, are:

a. Majority will prefer "Payoff on Red" over "Payoff on Black" (i.e., I over II).

b. Majority will prefer "Payoff on Black or Yellow" over "Payoff on Red or Yellow" (i.e., IV over III).

The Ellsberg observations **a** and **b** become paradoxical if we interpret the preference for payoff I over II to mean that the subjective probabilities for Black are smaller than the known probability of Red of $1/3$, while at the same interpreting the preference for payoff IV over III to mean that the subjective probabilities for Black are greater than the known probability of Red of $1/3$, as this would be inconsistent.

Again, this paradox may be trivially resolved if one realizes that, all things being equal probability-wise, one is perfectly free to prefer both payoff I over II and payoff IV over III without being inconsistent. Stated differently, the inconsistency in the Ellsberg paradox lies not with the experimental subjects, but, rather, with the premise that the observed preferences necessarily imply differing subjective probabilities.

7.4 An Analysis of the Second Ellsberg Experiment

We now proceed to give a probability theoretical analysis of the pay-offs for n draws with replacement from the urn in the second Ellsberg experiment.

In order to construct the probability distribution of r and b, the number of red and black balls in n draws with replacement, for an urn having N balls of which R are red, and the number of black balls B is left unspecified, we may make use of the product rule (4.1):

$$p(r, b|\, n, R, N, I) = p(r|\, n, R, N)\, p(b|\, r, n, R, N, I),\qquad (7.11)$$

where the probability distribution of the number r of red balls in n draws is given as

$$p(r|\, n, R, N) = \frac{n!}{r!\,(n-r)!}\left(\frac{R}{N}\right)^r \left(\frac{N-R}{N}\right)^{n-r}\qquad (7.12)$$

and the marginal probability distribution of the number b of blacks balls in the remaining $n - r$ draws is given as

$$p(b|\, r, n, R, N, I) = \sum_{B=0}^{N-R} p(B, b|\, r, n, R, N, I)$$

$$(7.13)$$

$$= \sum_{B=0}^{N-R} p(B|\, R, N, I)\, p(b|\, r, n, R, B, N).$$

We may assign an uniform prior probability distribution to the unknown number of black balls in the urn B [47]:

$$p(B|\, N, R, I) = \frac{1}{N - R + 1},\qquad (7.14)$$

for $B = 0, 1, \ldots, N - R$, while we may assign a binomial probability distribution to the number b of blacks balls in the remaining $n - r$ draws for known proportions of black balls:

$$p(b|\, r, n, R, B, N) = \frac{(n-r)!}{b!\,(n-r-b)!}\left(\frac{B}{N-R}\right)^b \left(\frac{N-R-B}{N-R}\right)^{n-r-b}.\quad (7.15)$$

By way of (7.11) through (7.15) and some algebraic reshuffling, we find that the probability distribution of the second Ellsberg urn is given as

$$p(r, b|\, n, R, N) = \frac{1}{N - R + 1}\sum_{B=0}^{N-R}\frac{n!}{r!b!\,(n-r-b)!}\left(\frac{R}{N}\right)^r \left(\frac{B}{N}\right)^b \left(\frac{N-R-B}{N}\right)^{n-r-b}.$$

$$(7.16)$$

It may be checked from (7.18) that for $n = 1$ the probabilities of a red, black, and yellow draw are given, respectively, as

$$p(r = 1, b = 0 | n = 1, R, N) = \frac{R}{N}$$

$$p(r = 0, b = 1 | n = 1, R, N) = \frac{N - R}{2N} \qquad (7.17)$$

$$p(r = 0, b = 0 | n = 1, R, N) = \frac{N - R}{2N},$$

which for $R = 30$, and $N = 90$, as in the second Ellsberg experiment, gives specific probabilities of

$$p(r = 1, b = 0 | n = 1, R = 30, N = 90) = \frac{1}{3}$$

$$p(r = 0, b = 1 | n = 1, R = 30, N = 90) = \frac{1}{3} \qquad (7.18)$$

$$p(r = 0, b = 0 | n = 1, R = 30, N = 90) = \frac{1}{3}.$$

There are several ways by which to compute the expected values and standard deviations of payoffs I through IV. The easiest of those ways is to observe first that the payoffs I and IV are complementary in that it requires the probability distribution of r and its complement not-r and then to observe that the payoffs II and III are also complementary in that it requires the probability distribution of b and its complement not-b.

The probability distribution of r is simply the binomial distribution (7.12):

$$p(r | n, R, N) = \frac{n!}{r! \, (n - r)!} \left(\frac{R}{N}\right)^r \left(\frac{N - R}{N}\right)^{n - r}$$

which has a expected value for the payoff on red (7.2)

$$E(r | n, R, N) = n\frac{R}{N} \qquad (7.19)$$

and an expected value for the payoff on black and yellow, or, equivalently, not-r, of

$$E(n - r | n, R, N) = \sum_{r=0}^{n} (n - r) \; p(r | n, R, N)$$

$$= n - n\frac{R}{N} \qquad (7.20)$$

$$= n\frac{N - R}{N}.$$

The standard deviation for both payoffs I and IV is (7.3)

$$\text{std}(r|\,n, R, N) = \sqrt{n\frac{R\,(N-R)}{N^2}} = \text{std}(n - r|\,n, R, N).\qquad (7.21)$$

The probability distribution of b for an unknown number B of black balls in the urn is given as

$$p(b|\,n, R, N) = \sum_{B=0}^{N-R} p(B, b|\,n, R, N)$$

$$\qquad (7.22)$$

$$= \sum_{B=0}^{N-R} p(B|\,R, N)\,p(b|\,n, B, N),$$

where the prior distribution of the number B of black balls in the urn is given as (7.14)

$$p(B|\,R, N) = \frac{1}{N - R + 1}$$

and the probability distribution of b for a known number B of black balls in the urn is given as

$$p(b|\,n, B, N) = \frac{n!}{b!\,(n-b)!} \left(\frac{B}{N}\right)^{b} \left(\frac{N-B}{N}\right)^{n-b}.\qquad (7.23)$$

It follows from (7.14), (7.22), and (7.23) that

$$p(b|\,n, R, N) = \frac{1}{N - R + 1} \sum_{B=0}^{N-R} \frac{n!}{b!\,(n-b)!} \left(\frac{B}{N}\right)^{b} \left(\frac{N-B}{N}\right)^{n-b},\qquad (7.24)$$

which has an expected value for the payoff on black of (7.2)

$$E(b|\,n, R, N) = \sum_{b=0}^{n} b\,p(b|\,n, R, N)$$

$$= \sum_{B=0}^{N-R} \frac{1}{N - R + 1} \sum_{b=0}^{n} b\frac{n!}{b!\,(n-b)!} \left(\frac{B}{N}\right)^{b} \left(\frac{N-B}{N}\right)^{n-b}$$

$$= \frac{1}{N - R + 1} \sum_{B=0}^{N-R} n\frac{B}{N}\qquad (7.25)$$

$$= n\frac{N - R}{2N}$$

and an expected value for the payoff on red and yellow, or, equivalently, not-b, of, (7.25)

$$E(n - b|\, n, R, N) = \sum_{b=0}^{n} (n - b) \; p(b|\, n, R, N)$$

$$= \sum_{B=0}^{N-R} \frac{1}{N - R + 1} \sum_{b=0}^{n} (n - b) \frac{n!}{b!\,(n-b)!} \left(\frac{B}{N}\right)^{b} \left(\frac{N - B}{N}\right)^{n-b}$$

$$= \frac{1}{N - R + 1} \sum_{B=0}^{N-R} \left(n - n\frac{B}{N}\right)$$

$$= n - n\frac{N - R}{2N} \tag{7.26}$$

$$= n\frac{N + R}{2N},$$

The standard deviation for both payoffs II and III is

$$\mathrm{std}(b|\, n, R, N) = \sqrt{n\,(N - R)\left[\frac{2\,(N + 2R - 1)}{12N^2} + n\frac{(N - R + 2)}{12N^2}\right]} = \mathrm{std}(n - b|\, n, R, N), \tag{7.27}$$

since we have that

$$\mathrm{var}(b|\, n, R, N) = \sum_{b=n}^{R} \left[b - E(b|\, n, R, N)\right]^2 p(b|\, n, R, N)$$

$$= \sum_{B=0}^{N-R} \frac{1}{N - R + 1} \sum_{b=0}^{n} \left(b - n\frac{N - R}{2N}\right)^2 \frac{n!}{b!\,(n-b)!} \left(\frac{B}{N}\right)^{b} \left(\frac{N - B}{N}\right)^{n-b}$$

$$= \frac{1}{N - R + 1} \sum_{B=0}^{N-R} n\frac{B\,(N - B)}{N^2} + n^2\frac{(N - R - 2B)^2}{N^2} \tag{7.28}$$

$$= n\,(N - R)\left[\frac{2\,(N + 2R - 1)}{12N^2} + n\frac{(N - R + 2)}{12N^2}\right]$$

It follows from (7.18) that for the special case of $n = 1$ the payoffs I and II are equivalent probability-wise, as are payoffs III and IV. This probabilistic equivalency for these sets of payoffs does not hold for repeated bets.

The standard deviation of payout II (7.27) will exceed the standard deviation of payout I (7.21) for repeated bets , while the expected values of both payouts remain identical to each other, (7.19) and (7.25). Likewise, the standard deviation of payout IV (7.27) will exceed the standard deviation of payout II (7.21) for repeated bets, while the expected values of both payouts remain identical to each other, (7.20) and (7.26).

So, all things being equal probability-wise, there is an observed preference for the urn which in the repeated-draw case results in the least uncertainty.

7.5 Discussion

It is stated that the unknown quantities in the Ellsberg urns are examples of non-quantifiable Knightian uncertainty, whereas the known quantities are examples of probability [24]. But it has been shown that also for unknown quantities probability distributions can be assigned to the number of balls; i.e., (7.6) and (7.23).

The Ellsberg observations have also been interpreted as evidence that people are generally not probabilistically sophisticated [34]:

> "Experimental results indicate that many DMs [i.e. decision makers] are indifferent between (urn 1, red) and (urn 1, blue) but they strictly prefer either of these choices to (urn 2, red) and (urn 2, blue). If the DMs were probabilistically sophisticated and assigned probability p to choosing a red ball from urn 1 and q to choosing a red ball from urn 2, the preferences above would indicate that $p = 1 - p$, $p > q$, and $p > 1 - q$, a contradiction. Hence, many DMs are not probabilistically sophisticated."

In other words, it is argued that if people were probabilistically sophisticated, then in the first Ellsberg experiment they ought to act as if they assign a probability of $1/2$ to the probability of drawing a red ball from the urn with unknown quantities of red balls and for the second experiment they ought to act as if they assign a probability of $1/3$ to the probability of drawing a black ball.

But, as stated before, one is perfectly free in the Ellsberg experiments to choose one option over the other without being inconsistent, as both options are so constructed as to be equivalent probability-wise; i.e., (7.9) and (7.10) for the first experiment, and (7.19) and (7.18) for the second experiment.

More importantly even, people seem to be able to intuit this probabilistic equivalency all on their own, without the aid of a formal Bayesian analysis (i.e., the derivations that lead us to (7.10) and (7.18)), as they let themselves be persuaded of, or, alternatively, as they argue, the validity of a paradox that pivots around the correct understanding that a probability of $1/2$ ought to be assigned to the drawing of a red ball from the second urn in the first experiment and a probability of $1/3$ to the drawing of a black ball in the second experiment.

In closing, all things being equal probability-wise, there is an observed preference for the payoff which in the repeated-draw case would result in the least

uncertainty; i.e., (7.3) is preferred over (7.8) in the first Ellsberg experiment, and (7.21) is preferred over (7.27) in the second. If this is either just a coincidence, or a necessary corollary of ambiguity aversion, or the very definition of ambiguity aversion itself (i.e., second-order probability [15]), is a subject for further research.

Chapter 8

Common Sense Quantified

The working assumption in this thesis is that Bayesian probability is common sense quantified [40, 47]. We will now give an outline of the proofs by Cox and Knuth that both lead to the product and sum rules, (4.1) and (4.2), as the necessary and sufficient operators of consistent inference [17, 65].

8.1 The Cox Derivation

Cox found that if we try to represent degrees of plausibility by real numbers, then the conditions of consistency can be stated by functional equations whose general solutions can be found. The results are: out of all possible monotonic functions which might in principle serve our purpose, there exists a particular scale, that is, class of functions, on which to measure degrees of plausibility which we henceforth call 'probability'. The consistent rules on how to combine these probabilities take the form of Laplace's product and sum rules. Cox, thus, proved that any method of inference in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to Laplace's or inconsistent [17].

8.1.1 The Product Rule

The derivation of the product and sum rules starts with some desiderata of common sense we wish our system of inference to adhere to. The first of these desiderata is that we wish to order our plausibilities consistently, that is, transitively.

Let the symbol '$>$' stand for the statement 'more plausible than', then the desideratum of transitivity translates to

$$A > B \quad \text{AND} \quad B > C \quad \implies \quad A > C, \tag{8.1}$$

where '$\implies$' is the symbol for implication. Any transitive ranking can be mapped onto real numbers, by assigning numerical codes $w(A)$, $w(B)$, $w(C)$, etc., such that

$$w(A) > w(B) > w(C) > \text{etc} \ldots \tag{8.2}$$

So the most basic of desiderata, that is, transitivity, leads to the observation that we wish to express our probabilities as numbers on the real line, [90].

The second desideratum is the desideratum of consistency, that is, if we specify the plausibility of proposition A as well as the plausibility of proposition B after having observed A, then we have implicitly the plausibility of propositions A and B occurring together, that is

$$w(AB) = F\left[w(A), w(B|A)\right], \tag{8.3}$$

where F is some unknown function. Likewise, we have that the plausibility of proposition B as well as the plausibility of proposition A after having observed B, implicitly give us the plausibility of propositions A and B occurring together

$$w(AB) = F\left[w(B), w(A|B)\right]. \tag{8.4}$$

Note that only functional relations of the type (8.3) and (8.4) do not exhibit qualitative violations of common sense in some extreme case [100].

Using the Boolean rules obeyed by logical conjunction of propositions, Cox is able to manipulate (8.3) and (8.4) into the associativity functional equation [17]:

$$F\left[p, F(q, r)\right] = F\left[F(p, q), r\right]. \tag{8.5}$$

Assuming differentiability of the unknown functional F, the third and final desideratum, there exists some monotonically increasing non-negative function π, say, of the coded plausibilities w, in terms of which the unknown function F is just scaled multiplication [1]:

$$\pi[w(A)]\,\pi[w(B|A)] = \pi[w(AB)] = \pi[w(B)]\,\pi[w(A|B)]. \tag{8.6}$$

In Jaynes' derivation of the product and sum rules the desideratum of differentiability is called the desideratum of qualitative correspondence with common sense [47]. Since it is common sense that would suggest to us that infinitesimal small changes in either $w(A)$ or $w(B|A)$ can only lead to corresponding infinitesimal small changes in $w(AB)$, (8.3).

Some Boundary Conditions

Now, consistency, the second desideratum, also demands that propositions having equal truth values are equally plausible and, hence, should be assigned the same numerical values. So, if A implies AB, then we have the following equality of truth values

$$A = AB. \tag{8.7}$$

By way of the consistency desideratum we then have that

$$w(A) = w(AB). \tag{8.8}$$

Since π is some function of the numerically coded plausibilities, we also have, trivially, that

$$\pi[w(A)] = \pi[w(AB)]. \tag{8.9}$$

Furthermore, if A implies AB, then we know B to be true whenever A is true. Substituting (8.9) into (8.6), we then find the identity

$$\pi[w(A)] = \pi[w(A)]\,\pi[w(\,B|\,A)]\,. \tag{8.10}$$

It follows that, in order for the equality (8.10) to hold, the certainty of B given A should be represented by the number 1, that is,

$$\pi[w(\,B|\,A)] = 1 = \text{certainty}. \tag{8.11}$$

If A is impossible, then so are AB and A given B. By the desiderata of consistency, we then have that

$$w(A) = w(AB) = w(\,A|\,B)\,. \tag{8.12}$$

As π is some function of the numerically coded plausibilities, we also have, trivially, that

$$\pi[w(A)] = \pi[w(AB)] = \pi[w(\,A|\,B)]\,. \tag{8.13}$$

Substituting (8.13) into (8.6), we may find the identity

$$\pi[w(A)]\,\pi[w(\,B|\,A)] = \pi[w(A)] = \pi[w(B)]\,\pi[w(A)]\,. \tag{8.14}$$

In order for the equality (8.14) to hold, impossibility of A should be either represented by the number 0 or by ∞. Following historical precedence, we let impossibility be represented by 0 rather than by ∞:

$$\pi[w(A)] = 0 = \text{impossibility}. \tag{8.15}$$

It should be noted that the boundary conditions (8.11) and (8.15) are postulated axiomatically for Kolgomorov's probability measure on sets, whereas in the Cox derivation these boundary conditions may be derived from the primitive desiderata of transitivity, consistency, and differentiability.

8.1.2 The Sum Rule

We will now proceed to outline the derivation of the Cox analog of Kolgomorov's additivity axiom for propositions.

We assume that the number we assign to our plausibility of proposition A somehow is related to the number we assign to our plausibility of proposition A not being true, that is,

$$\pi\left[w(\overline{A})\right] = G\left\{\pi[w(A)]\right\}. \tag{8.16}$$

By way of (8.16) and some involved Boolean algebra, we may obtain the equality [47]:

$$x\,G\left[\frac{G(y)}{x}\right] = y\,G\left[\frac{G(x)}{y}\right]\,. \tag{8.17}$$

Using the boundary conditions, $0 \le x \le 1$ and $0 \le y \le 1$, (8.11) and (8.15), and assuming differentiability of the function G, the solution of (8.17) may be found to be [47]:

$$G(x) = (1 - x^m)^{1/m}, \quad m > 0. \tag{8.18}$$

It follows from (8.16) and (8.18) that

$$\pi\left[w\big(\overline{A}\big)\right]^m = 1 - \pi[w(A)]^m \tag{8.19}$$

We observe that (8.6), without any loss of generality, may equally well be written as

$$\pi[w(A)]^m \, \pi[w(\,B|\,A)]^m = \pi[w(AB)]^m = \pi[w(B)]^m \, \pi[w(\,A|\,B)]^m \,. \tag{8.20}$$

If we define the probability measure $P(x)$ to be the mth power of some monotonically increasing non-negative function of the plausibility of some proposition x, that is,

$$P(x) = \pi[w(x)]^m \,, \tag{8.21}$$

where $m > 0$, (8.18), and $0 \le \pi(y) \le 1$, (8.11) and (8.15). Then, substituting (8.21) into (8.19) and (8.20), we obtain the product and sum rules

$$P(A)\,P(\,B|\,A) = P(AB) = P(\,A|\,B)\,P(B) \tag{8.22}$$

and

$$P\big(\overline{A}\big) = 1 - P(A)\,. \tag{8.23}$$

This concludes the Cox derivation of the rules of probability theory. We now proceed the give the alternative derivation by Knuth that encompasses and generalizes both the Cox and Kolmogorov formulations, respectively, [17] and [71].

8.2   The Knuth Derivation

By introducing probability as a bi-valuation defined on a lattice of statements we can quantify the degree to which one statement implies another. This generalization from logical implication to degrees of implication not only mirrors Cox's notion of plausibility as a degree of belief, but includes it. The main difference is that Cox's formulation is based on a set of desiderata derived from his particular notion of plausibility; whereas here the symmetries of lattices in general form the basis of the theory and the meaning of the derived measure is inherited from the ordering relation, which in the case of statements is implication [64].

The fact that these lattices may also be derived from sets means that this works encompasses Kolmogorov's formulation of probability theory as a measure on sets. However, mathematically this theory improves on Kolmogorov's foundation by deriving, rather than assuming, summation. Furthermore, this foundation further extends Kolmogorov's measure-theoretic foundation by introducing the concept of context. This leads directly to probability necessarily being conditional, and Bayes' Theorem follows as a direct result of the chain rule in terms of a change in context [64].

8.2.1 Lattice Theory and Quantification

Two elements of a set are ordered by comparing them according to a binary ordering relation, that is, by way of '$\leq$', which may be read as 'is included by'. Elements may be comparable, in which case they form a chain, or they may be incomparable, in which case they form a antichain. A set consisting of both inclusion and incomparability are called partially ordered sets, or posets for short [62].

Given a set of elements in a poset, their upper bound is the set of elements that contain them. Given a pair of elements x and y, the least element of the upper bound is called the join, denoted[1] $x \vee y$. The lower bound of a pair of elements is defined dually by considering all the elements that the pair of elements share. The greatest elements of the lower bound is called the meet, denoted $x \wedge y$.

A lattice is a partially ordered set where each pair of elements has a unique meet and unique join. There often exist elements that are not formed from the join of any pair of elements. These elements are called join-irreducible elements. Meet-irreducible elements are defined similarly. We can choose to view and join and meet as algebraic operations that take any two lattices elements to a unique third lattice element. From this perspective, the lattice is an algebra.

An algebra can be extended to a calculus by defining functions that take lattice elements to real numbers. This enables one to quantify the relationships between the lattice elements. A valuation v is a function that takes a single lattice element x to a real number $v(x)$ in a way that respects the partial order, so that, depending on the type of algebra, either $v(x) \leq v(y)$ or $v(y) \leq v(x)$, if in the poset we have that $x \leq y$. This means that the lattice structure imposes constraints on the valuation assignments, which can be expressed as a set of constraint equations [64].

8.2.2 The General Sum rule

In what follows we will closely follow the exposition of [64]. This exposition is a beautiful re-telling of a tale already told by Cox [17]. Though Knuth's narration is much more abstract and, consequently, general. Point in case being that it will lead us, amongst other things [68], to the derivation of a truly Bayesian information theory, also called inquiry calculus[2] [59].

We begin by considering a special case of elements x and y with join $x \vee y$ and a null meet $x \wedge y = \emptyset$. The value we assign to the join $x \vee y$, written $v(x \vee y)$, must be a function of the values we assign to both x and y, $v(x)$ and $v(y)$. Since, if there did not exist any functional relationship, then the valuation could not possibly reflect the underlying lattice structure; that is, valuation must maintain ordering, in the sense that $x \leq x \vee y$ implies either $v(x) \leq v(x \vee y)$ or $v(x) \geq v(x \vee y)$. So, we write this functional relationship in terms of an

---

[1] Note that we over-load the symbol '$\vee$' here, which still stands for disjunction, though now in the general context of lattice theory.

[2] This inquiry calculus is the topic of the third part of this thesis.

unknown binary operator $\oplus$:

$$v(x \vee y) = v(x) \oplus v(y). \tag{8.24}$$

We now consider another case where we have three elements x, y, and z, such that their meets are again disjoint, Figure 8.1.
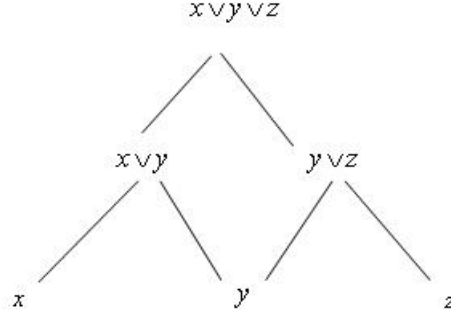
Figure 8.1: Sub-lattice of x, y, and z

Because of the *associativity* of the join, we have that the least upper bound of these three elements, $x \vee y \vee z$, can be obtained in three different ways, of which two are given below:

$$x \vee (y \vee z) \qquad \text{and} \qquad (x \vee y) \vee z. \tag{8.25}$$

By applying (8.24) to (8.25), it follows that the value that we assign to this join can also be obtained in different ways:

$$v(x) \oplus [v(y) \oplus v(z)] \qquad \text{and} \qquad [v(x) \oplus v(y)] \oplus v(z). \tag{8.26}$$

Consistency then demands that the equivalent assignments (8.26) have the same value:

$$v(x) \oplus [v(y) \oplus v(z)] = [v(x) \oplus v(y)] \oplus v(z). \tag{8.27}$$

This the functional equation for the operator $\oplus$, for which the general solution is given by [1]:

$$f[v(x \vee y)] = f[v(x)] + f[v(y)], \tag{8.28}$$

where f is an arbitrary invertible function, so that many valuations are possible. If we define the valuation u as

$$u(x) \equiv f[v(x)],$$

then we may rewrite (8.28) as

$$u(x \vee y) = u(x) + u(y). \tag{8.29}$$

Now that we have a constraint on the valuation for our simple example, we seek the general solution for the entire lattice. To derive the general case, we consider the lattice in Figure 8.2.
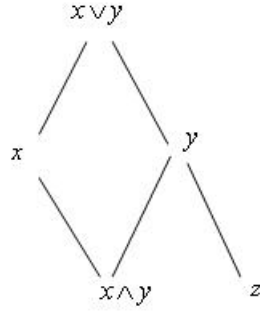
Figure 8.2: Extended lattice of x and y

In this lattice, the meet $x \wedge y$ and the element z, as well as the elements x and z, have a null meet. So, applying (8.29) to both cases, we get

$$u(y) = u(x \wedge y) + u(z), \tag{8.30}$$

and

$$u(x \vee y) = u(x \vee z) = u(x) + u(z),$$

since y is the part it shares with x joined to z. Substituting for $u(z)$ in (8.30) and in (8.31), we get the general sum rule:

$$u(x \vee y) = u(x) + u(y) - u(x \wedge y). \tag{8.31}$$

For bi-valuations, we have in general

$$w(x \vee y \,|\, t) = w(x \,|\, t) + w(y \,|\, t) - w(x \wedge y \,|\, t), \tag{8.32}$$

for any context t. Note that the sum rule is not focused solely on joins since it is symmetric with respect to interchange of joins and meets.

At this point we have *derived* additivity of the measure, which is considered to be an axiom of measure theory. This is significant in that associativity constrains us to have additive measures - there is no other option [65].

8.2.3 The Chain Rule

We now focus on bi-valuations and explore changes in context [64]. We begin with a special case and consider four ordered elements $x \leq y \leq z \leq t$.

The relationship $x \leq z$ can be divided into the two relations $x \leq y$ and $y \leq t$. In the event that z is considered to be the context, this sub-division implies that the context can be considered in parts. The bi-valuation we assign to x with respect to the context z, that is, $w(x|z)$, must be related to both the bi-valuation we assign to x with respect to the context y, that is, $w(x|y)$, and the bi-valuation we assign to y with respect to the context z, that is, $w(y|z)$.

So, there exists a binary operator $\otimes$ that relates the bi-valuations assigned to the two steps to the bi-valuation assigned to the one step:

$$w(x|z) = w(x|y) \otimes w(y|z). \tag{8.33}$$

By extending (8.33) to three steps, and considering the bi-valuation $w(x|t)$, relating x and t, via intermediate contexts y and z, Figure 8.3,
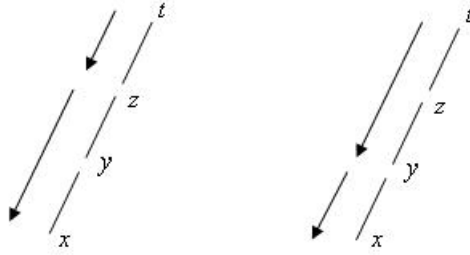
Figure 8.3: Context lattice of t

we obtain the associativity relationship:

$$[w(x|y) \otimes w(y|z)] \otimes w(z|t) = w(x|y) \otimes [w(y|z) \otimes w(z|t)]. \tag{8.34}$$

By way of the associativity theorem, (8.34) results in a constraint equation for non-negative bi-valuations involving changes in context [65]:

$$w(x|z) = w(x|y)\, w(y|z). \tag{8.35}$$

We call this the *chain rule*.

This completes the derivation of a valuation calculus. The associativity of the join gives rise to the sum rule, which is symmetric with respect to interchange of joins and meets, whereas the associativity of changes of context results in a chain rule for bi-valuations that dictates how valuations should be manipulated when changing context [64].

8.2.4 The Hypothesis Space

The state space is an enumeration of all the possible states that our system may be in. A given individual may not know precisely which state the system is in, but may have some information that rules out some states, but not others. So, the set of potential states defines what one can say about the state of the system. For this reason, we call a set of potential states a statement. A statement describes a state of knowledge about the state of the system. The set of all possible statements is called the *hypothesis space* [63].

The lattice of statements is generated by taking the power set, which is the set of all possible subsets of the set of all states, and ordering them according to set inclusion. For a system of n possible states, there are

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n$$

statements, including the null set. The bottom element is often omitted from the diagram due to the fact that it represents the logical absurdity. The statement at the top is the truism, generically called the top and denoted 'T', which represents the state of knowledge where one only knows that the system can be in one of n possible states.

The ordering relation of set inclusion naturally encodes logical implication, such that a statement implies all the statements above it. Logical deduction is straightforward in this framework since a statement in the lattice implies (i.e., is included by) every element above it with certainty. For example, x implies $x \vee y$, $x \vee y \vee z$, etc. In this sense the lattice of statements is an algebra.

Logical induction works backwards. One would like to quantify the degree to which one's current state of knowledge implies a statement of greater certainty below it. Since statements do not imply statements below them, this requires a generalization of the algebra representing ordering. In the previous section we have laid the groundwork for generalizing this algebra to a calculus by introducing quantification. In what follows we derive a measure, called probability, that quantifies the degree to which one statement implies another.

8.2.5 The Product Rule for the Lattice of Statements

We now focus on applying the sum and chain rule to the lattice of statements in Figure 8.4, where the elements x, y, z are understood to be statements. First we focus on the small diamond in Figure 8.4 which is defined by x, $x \vee y$, y, and $x \wedge y$, Figure 8.5. If we consider the context to be x, then the sum rule (8.32) for this diamond may be written down as:

$$w\left(x \vee y \mid x\right) = w\left(x \mid x\right) + w\left(y \mid x\right) - w\left(x \wedge y \mid x\right). \tag{8.36}$$

Since $x \leq x$ and $x \leq x \vee y$, we have that the statement x implies both statements x and $x \vee y$ with absolute certainty, that is,

$$w\left(x \mid x\right) = w\left(x \vee y \mid x\right) = 1. \tag{8.37}$$

$$x \vee y \vee z$$

$$x \vee y \qquad y \vee z$$

$$x \qquad y \qquad z$$

$$x \wedge y \qquad y \wedge z$$

$$x \wedge y \wedge z$$
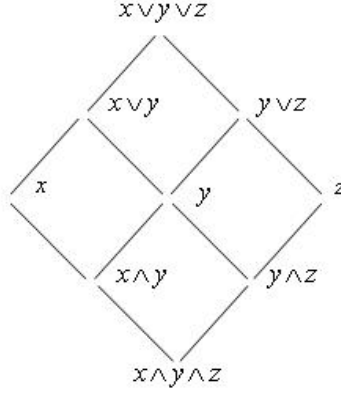
Figure 8.4: A lattice of statements

Substituting (8.37) into the sum rule (8.36), we obtain:

$$w\left(x \wedge y \mid x\right) = w\left(y \mid x\right). \tag{8.38}$$

This relationship is expressed by the equivalence of the arrows in Figure 8.5.

$$x \vee y$$

$$x \qquad y$$

$$x \wedge y$$
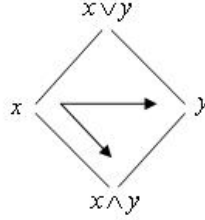
Figure 8.5: The diamond x, $x \vee y$, y, and $x \wedge y$

Consider the chain where the bi-valuation $w\left(x \wedge y \wedge z \mid x\right)$ with context x is decomposed into two parts, by introducing the intermediate context $x \wedge y$. The chain rule (8.35) gives

$$w\left(x \wedge y \wedge z \mid x\right) = w\left(x \wedge y \wedge z \mid x \wedge y\right) w\left(x \wedge y \mid x\right). \tag{8.39}$$

To simplify this relation, we first consider the parallelogram in Figure 8.4, defined by $x \wedge y$, $y \vee z$, z, and $x \wedge y \wedge z$.

If we consider the context to be $x \wedge y$, then the sum rule (8.32) for this parallelogram may be written down as:

$$w\left(y \vee z \mid x \wedge y\right) = w\left(x \wedge y \mid x \wedge y\right) + w\left(z \mid x \wedge y\right) - w\left(x \wedge y \wedge z \mid x \wedge y\right). \tag{8.40}$$

Since $x \wedge y \leq x \wedge y$ and $x \wedge y \leq y \vee z$, we have that the statement $x \wedge y$ implies both statements $x \wedge y$ and $y \vee z$ with absolute certainty, that is,

$$w\left(x \wedge y \mid x \wedge y\right) = w\left(y \vee z \mid x \wedge y\right) = 1. \tag{8.41}$$

Substituting (8.41) into the sum rule (8.40), we obtain:

$$w\left(x \wedge y \wedge z \mid x \wedge y\right) = w\left(z \mid x \wedge y\right). \tag{8.42}$$

We now have simplified the first term on the right hand side of (8.39). Note that we can also deduce equality (8.42), by noting that the right-hand parallelogram in Figure 8.6 has the same topology as the diamond in Figure 8.5.
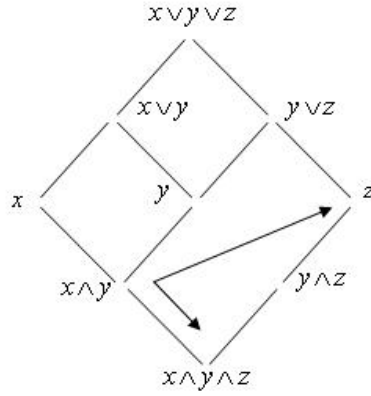
Figure 8.6: The parallelogram $x \wedge y$, $y \vee z$, z, and $x \wedge y \wedge z$

In order to simplify the left hand side of (8.39), we consider the parallelogram defined by x, $x \vee y$, $y \wedge z$, and $x \wedge y \wedge z$. If we consider the context to be x, then the sum rule (8.32) for this parallelogram may be written down as follows:

$$w\left(x \vee y \mid x\right) = w\left(x \mid x\right) + w\left(y \wedge z \mid x\right) - w\left(x \wedge y \wedge z \mid x\right). \tag{8.43}$$

Since $x \leq x$ and $x \leq x \vee y$, we have that the statement x implies both statements x and $x \vee y$ with absolute certainty, that is,

$$w\left(x \mid x\right) = w\left(x \vee y \mid x\right) = 1. \tag{8.44}$$

Substituting (8.44) into the sum rule (8.43), we obtain:

$$w\left(x \wedge y \wedge z \mid x\right) = w\left(y \wedge z \mid x\right). \tag{8.45}$$

We can also deduce equality (8.45), by noting that the left-hand parallelogram in Figure 8.7 has the same topology as the diamond in Figure 8.5.
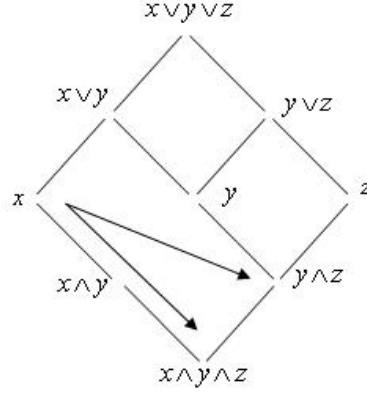
Figure 8.7: The diamond x, $x \vee y$, $y \wedge z$, and $x \vee y \vee z$

Substituting the simplifications (8.38), (8.42), and (8.45) into (8.39) results in the general *product rule for context change on the lattice of statements*:

$$w\left(y \wedge z \mid x\right) = w\left(y \mid x\right) w\left(z \mid x \wedge y\right). \tag{8.46}$$

In (8.36) and (8.46), we now have the two necessary rules of probability theory.

By relabeling the measure w to p, we get the sum and product rule of probability theory:

$$p\left(x \vee y \mid t\right) = p\left(x \mid t\right) + p\left(y \mid t\right) - p\left(x \wedge y \mid t\right) \tag{8.47}$$

and

$$p\left(x \wedge y \mid t\right) = p\left(x \mid t\right) p\left(y \mid x \wedge t\right),$$

where t is typically a context situated higher on the lattice. Note that for lattices, we have that

$$a \leq b \Rightarrow a \wedge b = a. \tag{8.48}$$

So, for compactness sake, we may also write the product rule of probability theory, (8.48), as

$$p\left(x \wedge y \mid t\right) = p\left(y \mid x\right) p\left(x \mid t\right). \tag{8.49}$$

This concludes the Knuth proof of the sum and product rules of the Bayesian probability. In part three of this thesis there are derived the sum and product rules of a Bayesian information theory by way of the lattice theoretical approach that was presented in this section.

Chapter 9

A Short Historical Overview

Laplace, an 18th century mathematical physicist, is considered by many to be the first Bayesian[1], in that he was to the first to propose and apply the product and sum rules with great success and generality to a wide range of problems of inference in applied physics.

Comparing experimental observations with existing theory, by way of his novel probability theory, Laplace would typically focus his attention only on those theories where discrepancies between observations and theory were so large that they indicated with high probability the existence of some still unknown systematic cause. This approach led him to some of his most important discoveries in celestial mechanics [45].

From Laplace we also have what is perhaps the first formal quantitative treatment of interval estimation. It had been estimated in Laplace's time, based on the mutual perturbations of Jupiter and Saturn, and the motion of their moons, that the mass of Saturn had to be a 1/3512th part of the mass of the sun. Laplace quantified the accuracy of this estimate by computing that the probability of $P = 0.0001$ for this estimation being in error with more than 1%. Indeed, in 1976 it was found that another 150 years of accumulation of data had only managed to increase this estimation by a mere 0.63%, well within the bounds predicted [44].

Laplace's successes in the field of celestial mechanics were such that he was called the French Newton by his contemporaries. Nonetheless, shortly after his death there started a series of increasingly violent attacks on his work, starting with Ellis [23]. The proponents of this counter-stream of thought were pure mathematicians who, totally disregarding the many pragmatic successes, felt that Laplace's definition of probability lacked the necessary rigor.

---

[1]Reverend Thomas Bayes derived the beta-distribution, thus, providing a solution to the problem of inverse probabilities for binomial ditributions. But Bayes did not formulate the product and sum rules [47].

These mathematicians rejected the notion of probability theory as describing a state of knowledge and insisted that by 'probability' one must mean only 'frequency in a random experiment'. The rational being that a state of knowledge is too 'subjective', whereas only frequencies are 'objective'. For a time this viewpoint dominated the field so completely that those who were students in the period 1930-1960 were hardly aware that any other conception had ever existed [45].

In 1939 the geophysicist Jeffreys produced a book in which the methods of Laplace were reinstated and applied to a mass of then current scientific problems, [48]. The applications worked out beautifully and yielded the same or demonstrably better results than those found by sampling theory methods. But, unfortunately, like Laplace, Jeffreys did not derive his principles as necessary consequences of compelling desiderata; and thus left room to continue the same old arguments over their justification. The sampling theorists, seizing eagerly upon the lack of rigour while again totally ignoring the practical successes, proceeded to give Jeffreys the same treatment as Laplace [45]. But Jeffreys work did succeed in sparking the Bayesian renaissance we are witnessing today. Jaynes, a physicist, and Zellner, an econometrist, both pioneers of this renaissance in their respective fields, took their direct inspiration from Jeffreys, [41, 112].

A further impetus for the reinstatement of Laplace's probability theory came from an article by Cox in which the problem under debate was turned around, [17]. Instead of making dogmatic assertions that it is or is not legitimate to use probability in the sense of degree of plausibility rather than frequency, Cox asked the question: Is it possible to construct a consistent set of mathematical rules for carrying out plausible, rather than deductive, reasoning?

He found that, if we try to represent degrees of plausibility by real numbers, then the conditions of consistency can be stated by functional equations whose general solutions can be found. The results were: out of all possible monotonic functions which might in principle serve our purpose, there exists a particular scale, that is, class of functions, on which to measure degrees of plausibility which we henceforth call 'probability'. The consistent rules on how to combine these probabilities take the form of Laplace's product and sum rules. Cox, thus, proved that any method of inference in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to Laplace's or inconsistent.

But this is not where our story ends. The Cox formulation encompasses and generalizes the Kolmogorov formulations, as it derives, rather than postulates, the properties of normalization, non-negativity, and additivity, and does so in the more general context of proposition logic, rather than in the more confined context of set theory. Knuth and Skilling, both direct descendants from the Laplace-Jeffreys-Jaynes line, have, in their turn encompassed and generalized the Cox formulations [65].

By introducing probability as a bi-valuation defined on a lattice of statements they quantify the degree to which one statement implies another. This generalization from logical implication to degrees of implication not only mirrors Cox's notion of plausibility as a degree of belief, but includes it. The main difference is that Cox's formulation is based on a set of desiderata derived from his

particular notion of plausibility. Whereas the symmetries of lattices in general form the basis of the theory and the meaning of the derived measure is inherited from the ordering relation, which in the case of statements is implication.

Furthermore, by introducing the measure of relevance as a bi-valuation defined on a lattice of questions it is possible to quantify the degree to which the answering of one question is relevant to the answering of another. This quantification gives rise to an extended information theory, also called inquiry calculus, which is intimately connected with probability theory and has its own 'Bayesian' product and sum rules for relevancies[2].

The lattice theoretic derivation of this new extended information theory, together with some information theoretic applications, are given in the third part of this thesis.

---

[2]As an aside, this new inquiry calculus took as its starting point the very last article Cox ever wrote [58].

Chapter 10

Discussion

If we wish our reasoning to adhere to the basic desiderata of transitivity and consistency then it may be derived that the product and sum rules are the sufficient and necessary operators for our given probabilities. But these rules stay mute on the subject on how we should assign probabilities to our initial propositions, [47].

While performing a data-analysis, a Bayesian may, depending on the structure of the problem under consideration, assign his probabilities on the basis of either indifference[1] or entropy principles, [42, 43, 45]. Nonetheless, the assignment of probabilities is still very much an open-ended research question, as there still are many instances in which the necessary problem structure is lacking for the application of either principle to our propositions, [47].

So, from a Bayesian perspective, all psychology literature on how we as humans assign probabilities to our initial propositions may, in principle, be a welcome help in the construction of a general model of human risk perception, the subject of this thesis. However, this is with the important caveat that the human reasoning process itself is understood to be inherently rational; that is, once we have assigned our probabilities, then rationality, or, equivalently, common sense, dictates us that we apply the product and sum rules, in some shape or form, to these probabilities.

So what about the role of affect in reasoning? Polya, a mathematician, demonstrates in great detail how mathematicians may derive their theorems by way of rational inference, that is, the product and sum rules, [84, 85]. In contrast, Slovic *et al.*, social scientists themselves, point to the intuitive goodness that those same mathematicians may feel once they have derived their elegant and simple theorems and suggest that this positive affect is what guides the mathematicians to their theorems, [97].

As human beings are not wholly infra-rational, groping their way through life by emotions alone, nor are they wholly rational, we propose a middle position. The mathematician deriving his theorems is a limit case of rationality with low

---

[1]The indifference principle is implicitly applied whenever we assign a fifty-fifty probability to the heads-or-tails proposition in a coin flip.

emotional content. In such limit cases of rationality the product and sum rules, that is, Bayesian probability theory, provides us with an excellent model of how we as humans navigate in a rational manner through the hypothesis space.

However, there are also instances where the states of nature under consideration have such a strong emotional content that they may distort our natural tendency towards rational reasoning. For example, denial is a state of mind in which one tries to override one's own rationality and assign a probability of zero to those states of nature which are too painful to conceive, even in the face of overwhelming evidence pointing to the contrary[2].

---

[2]A poignant example of irrational denial could be witnessed in the Chernobyl disaster [79]. After the initial Chernobyl explosion all available dosimeters had limits of 0.001 Roentgens per second (R/s). So the reactor crew could only state with certainty the lower bound of radiation levels of 0.001 R/s. The reactor crew chief Alexander Akimov then choose to assume that the reactor was intact and that this lower bound was the actual radiation level. He subsequently ignored the evidence of pieces of graphite and reactor fuel lying around the building, while also dismissing the reading of a new dosimeter that was brought in later and which had a limit of 1.0 R/s. It was later estimated that the actual radiation levels in the worst-hit areas of the reactor building were at 5.6 R/s at the time, where 0.028 R/s are deemed to be fatal after 5 hours.

Part II

Bayesian Decision Theory

Chapter 11

Introduction

The making of decisions is one of the most basic of human of activities. As humans, we are daily confronted by the necessity of having to choose between the alternative courses of actions that present themselves to us. It may be recognized that there are many instances where the decisions making process has a highly procedural quality about itself.

For example, when we are 'in doubt' we will typically enumerate the alternative courses of actions that are open to us, their possible consequences, the likelihood of these consequences, and the losses and gains which these consequences could entail, should they materialize.

As procedures imply structure, and structure may be captured in algorithms, we will offer here an algorithm of rational decision making. This algorithm is Bayesian in that rational judgment under uncertainty is modeled as the repeated application of the product rule and sum rules[1].

In the Bayesian decision theory each problem of choice is understood to consist of a set of decisions from which we must choose. Each possible decision, when taken, has its own set of possible outcomes, and each outcome, for a given decision, has its own plausibility of occurring relative to the other outcomes under that same decision. So, each decision in our problem of choice admits its own outcome probability distribution. The outcome probability distributions are the information carriers which represent our state of knowledge in regards to the consequences of our decisions.

Furthermore, each outcome may be mapped on either a single utility value or a range of utility values; where an utility value is understood to be the subjective worth which is assigned to a given objective outcome. The remapping of objective outcomes to their subjective values leaves us with the utility probability distributions. The utility probability distributions are the information carriers which represent the subjective valuation of the consequences of our decisions.

If our utilities are a monotonic increasing function of the perceived worth of the objective outcomes and if we have an utility axis which goes from mi-

---

[1]See Part I of this thesis

nus infinity to plus infinity, then the utility probability distribution which is 'more-to-the-right' will tend to be more profitable then the utility probability distribution which is 'more-to-the-left'. It follows that decisions may be made based on the comparison of their corresponding utility probability distributions.

Whenever the outcomes of our decision problems are monetary in nature, then we may assign utilities to monetary outcomes by way of the Weber-Fechner law which was proposed in 1860. The Weber-Fechner law is equivalent to Bernoulli's utility function. In other words, Bernoulli's logarithmic function for the subjective value of monetary outcomes, derived in 1738 from intuitive first principles, was a century later confirmed to hold for the perception of the subjective magnitudes of sensory stimuli in general.

In the Bayesian framework the utility probability distributions themselves are compared. This is done by comparing the positions of the utility probability distributions under the different decisions and, as a consequence, the Bayesian decision theory uses both the expected values and standard deviations of the utility probability distributions to differentiate between the different decisions.

In contrast, expected utility theory only uses the expected values of the utility probability distributions. It follows that the Bayesian decision theory presented in this thesis is an intuitive extension of expected utility theory. It will be demonstrated that this extension removes both the Ellsberg and the Allais[2] paradoxes.

Finally, the material in this thesis is presented in a loose narrative way. Important concepts are introduced in the first two chapters. Comprehensive decision theoretic examples are worked out in the last two 'Rationale of' chapters, whereas the more simple examples, as typically found in the decision theoretic literature, are sprinkled throughout the middle chapters.

---

[2]In fact, Allais constructed his paradoxes in order to demonstrate the need of variance preferences; that is, the need to also take into account the higher order moments of the utility probability distributions, [3].

Chapter 12

Bernoulli's Utility Function

In this chapter we will discuss the Bernoulli's utility function by which utilities may be assigned to increments in wealth. We will also here discuss the negative Bernoulli utility function by which utilities may be assigned to increments in debt.

12.1 Bernoulli's Original Derivation

Bernoulli derives his utility function by first observing that any increment of Δx in the initial wealth position x must correspond with some utility Δy [6].

This observation translates to the following trivial mathematical identity:

$$\Delta y = f(x + \Delta x) - f(x), \tag{12.1}$$

where f is some unknown function that assigns utilities to final wealth positions. And if we rewrite (12.1) as

$$\Delta y = \frac{f(x + \Delta x) - f(x)}{\Delta x} \, \Delta x, \tag{12.2}$$

then it follows that the increment Δy will tend to the differential

$$dy = \frac{f(x + dx) - f(x)}{dx} \, dx = f'(x) \, dx, \tag{12.3}$$

as the increment Δx tends to the differential dx.

Bernoulli then observes that an one ducat increment for someone who has a fortune of a hundred thousand ducats and a yearly income of five thousand ducats will have the same utility as one semi-ducat (i.e., a half-ducat) increment for someone who has a fortune of a hundred thousand semi-ducats and a yearly income of five thousand semi-ducats [6].

Stated differently, Bernoulli postulates that utilities of monetary increments Δx are invariant if we rescale the initial wealth position x and the monetary

increment Δx by the same factor c:

$$f(x + \Delta x) - f(x) = f(c\,x + c\,\Delta x) - f(c\,x)\,. \tag{12.4}$$

By way of , the right-hand side of (12.4) will tend to

$$\frac{f(c\,x + c\,dx) - f(c\,x)}{c\,dx}\,c\,dx = c\,f'(c\,x)\,dx, \tag{12.5}$$

as the increment Δx tends to the differential dx. Substituting (12.1) and (12.5) into (12.4), we obtain the functional equation

$$f'(x) = c\,f'(cx)\,, \tag{12.6}$$

which has its general solution [1]

$$f'(x) = q\,\frac{1}{x}, \tag{12.7}$$

where q is some constant. Substituting (12.7) into (12.1), we obtain the mathematical identity, as proposed by Bernoulli [6]:

$$dy = q\,\frac{dx}{x}, \tag{12.8}$$

where dy is an infinitesimally small utility and dx an infinitesimally small monetary increment.

Solving (12.8) for x, we find the general solution of the function that assigns utilities to final wealth positions:

$$y = f(x) = q\log x + c, \tag{12.9}$$

where c is some constant of integration. Substituting (12.9) into (12.1), we obtain Bernoulli's utility function:

$$\Delta y = q\log\frac{x + \Delta x}{x}. \tag{12.10}$$

And it should be noted that Bernoulli propose the difference function (12.10) as his utility function, as may be checked in Bernoulli's 1738 paper [6], rather than the utility function for final asset positions (12.9), as is erroneously stated in [54, 55].

In what follows, we rewrite (12.10) as

$$u(\Delta x\,|\,x) = q\log\frac{x + \Delta x}{x}, \tag{12.11}$$

where we let u be the function that assigns a utility to the monetary increment Δx conditional on the current wealth position x.

12.2 The Utility of Wealth and Loss Aversion

Let Δx be an increment, either positive or negative, in an initial wealth position x. Then we may define, by way of Bernoulli's utility function, the utility u of that monetary increment to be, (12.11),

$$u(\Delta x \,|\, x) = q \log \frac{x + \Delta x}{x}, \qquad \text{for } -x + x_0 < \Delta x < \infty, \tag{12.12}$$

where q is some positive scaling constant and x_0 is the threshold of wealth which is still significant.

The threshold of wealth Δx_0 has the following interpretation [47]. For everyone there will be some minimum amount of wealth that is still significant. If the loss of money results in a wealth that is smaller than this limit x_0, then we are left with an amount of money which, for all intents and purposes, is equivalent to financial ruin.

It will be found that in many instances the scaling constant q, also known as the Weber constant, will fall away in the decision theoretical (in)equalities. But if we wish to either give a graphical representation of (12.12) or compute the utiles of a monetary increment Δx conditional on a current wealth position x, then the Weber constant q must be set to some numerical value. This numerical value can be obtained by way of introspection and/or psychological experimentation.

For example, say, we have someone who has a monthly expendable income of a thousand euros for groceries and the like. Then introspection may suggest to us that a loss or gain of an amount less than ten euros would not move such a person that much; i.e., $\Delta x = \pm 10$ constitutes a just noticeable difference, or, equivalently, 1 utile, for an initial wealth of $x = 1000$, (12.12):

$$1 \text{ utile} = q \log \frac{1000 \pm 10}{1000}. \tag{12.13}$$

where utiles represent the utility of monetary outcomes, much like decibels represent the perceived intensity of sound. Solving for the unknown q, we then find a Weber constant for increments in wealth of

$$q = \frac{1}{\log\left(990/1000\right)} \approx \frac{1}{\log\left(1010/1000\right)} \approx 100. \tag{12.14}$$

Now, if we have a person who has three hundred euros per month to spend on groceries and the like and who stands to lose or to gain up to two hundred euros, then, by way of (12.12) and (12.14), we obtain the mapping of monetary outcomes to utilities given in Figure 12.1. Whereas for the case of the rich man who has one million euros to spend on groceries and the like and who stands to lose or to gain up to a hundred thousand euros, we obtain the alternative mapping given in Figure 12.2.

Loss aversion is the psychological phenomenon that losses may loom larger than gains [103]. Comparing Figures 12.1 and 12.2, we see that Bernoulli's
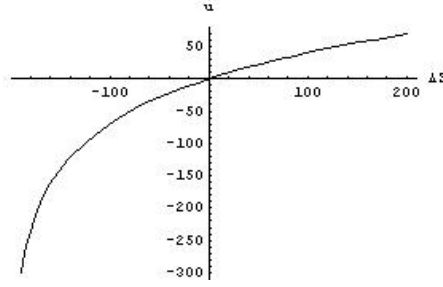
Figure 12.1: Utility plot for initial wealth 300 euros
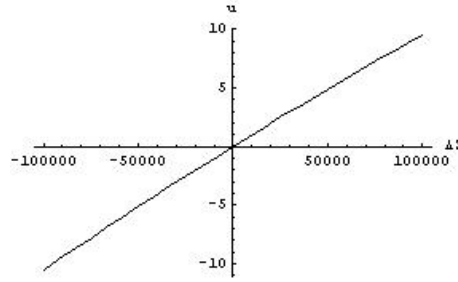
Figure 12.2: Utility plot for initial wealth 1,000,000 euros

utility function (12.12) captures both the loss aversion of the poor person, that is, asymmetry in the utility of equivalent gains and losses, as well as the linearity of the utility of relatively small equivalent gains and losses for the rich man[1].

In closing, note that different persons may have different Weber constants q for increments in wealth. For example, in a limit of detachment, one would expect a Buddhist monk to have a Weber constant that approaches zero. Also, it may be speculated that the Weber constant q is some monotonic decreasing function of time t, as it has been found that present gains and losses are felt more intensely than equivalent gains and losses at some future point in time [36]; i.e, $q(t_1) > q(t_2)$ for $t_1 < t_2$.

12.3   An Alternative Consistency Derivation

In what follows, we derive the Bernoulli utility function (12.11), or, equivalently, the Weber-Fechner law [77], or, equivalently in content, Stevens' Power law [98], using the desiderata of path independence and unit invariance [105].

Say, we have the positive quantities x and z, of some stimulus or commodity of interest. Then these quantities, being numbers on the positive real axis, admit

---

[1]As a matter of historical interest, it is to be noted the psychological phenomenon of loss aversion was already predicted and discussed by Bernoulli in his 1738 paper [6].

an ordering; i.e, we either have $x < z$, or $x = z$, or $x > z$. We now want to find the function f that quantifies the perceived distance between the quantities x and z.

This problem formulation gives us, by construction, our first functional equation. For we are implicitly saying that the perceived distance between x and z should be some function of the values x and z only, which is just another way of saying that the perceived decrease should be path independent [69]. So, the first consistency equation

$$f(x, z) = g[f(x, y), f(y, z)], \tag{12.15}$$

of which the solution is given in [65], is a direct consequence of the desideratum of path independence.

The second functional equation is based on the desideratum that the unknown function f should be invariant for a change of scale in our quantities:

$$f(x, z) = f(c\, x, c\, z), \tag{12.16}$$

where c is some positive constant. For example, if our quantities concern sums of money, then the perceived loss in going from ten euros to one euro should be the same perceived loss if we reformulate this scenario in euro cents, rather than euros.

The general solution to (12.16) is [1]

$$f(x, y) = h\left(\frac{y}{x}\right), \tag{12.17}$$

where h is some arbitrary function. The general solution to (12.15) is [65]

$$\Theta[f(x, z)] = \Theta[f(x, y)] + \Theta[f(y, z)], \tag{12.18}$$

where Θ is some arbitrary monotonic function. We may define Θ as [65]

$$\Theta(x) \equiv \log \Psi(x). \tag{12.19}$$

By way of (12.19), we then may rewrite (12.18), without any loss of generality, as

$$\log \Psi[f(x, z)] = \log \Psi[f(x, y)] + \log \Psi[f(y, z)], \tag{12.20}$$

or, equivalently,

$$\Psi[f(x, z)] = \Psi[f(x, y)] \ \Psi[f(y, z)]. \tag{12.21}$$

Substituting (12.17) into (12.18) and (12.21) and letting, respectively,

$$\theta\left(\frac{y}{x}\right) = \Theta\left[h\left(\frac{y}{x}\right)\right], \tag{12.22}$$

and

$$\psi\left(\frac{y}{x}\right) = \Psi\left[h\left(\frac{y}{x}\right)\right], \tag{12.23}$$

we obtain the equivalent functional equations

$$\theta\left(\frac{z}{x}\right) = \theta\left(\frac{y}{x}\right) + \theta\left(\frac{z}{y}\right) \tag{12.24}$$

and

$$\psi\left(\frac{z}{x}\right) = \psi\left(\frac{y}{x}\right) \, \psi\left(\frac{z}{y}\right). \tag{12.25}$$

We may use (12.24) to go from x to z via x itself:

$$\theta\left(\frac{z}{x}\right) = \theta\left(\frac{x}{x}\right) + \theta\left(\frac{z}{x}\right), \tag{12.26}$$

from which it follows that under θ the distance of x relative to itself must be zero:

$$f(x,x) = \theta\left(\frac{x}{x}\right) = 0. \tag{12.27}$$

By way of (12.27) and because of our intended use of the distance function f as an utility function of monetary increments, we may formulate the boundary condition that if y is our initial position, then a larger position x should correspond with a distance greater than zero:

$$f(x,y) = \theta\left(\frac{y}{x}\right) > 0 \quad \text{for } y > x. \tag{12.28}$$

Functional equation (12.24) and the boundary condition (12.28), together with the assumption of differentiability, are sufficient to find the function f that quantifies the perceived increase associated with going from the smaller quantity x to the greater quantity y. And this function turns out to be Bernoulli's utility function, or, equivalently, for the particular instance where x is the threshold value x_0, the Weber-Fechner law of sense perception[2]:

$$f(x,y) = q \log \frac{y}{x}, \qquad q \geq 0, \tag{12.29}$$

where x is our initial asset position, y the final asset position, and q is some arbitrary constant which has to be obtained by way psychological experimentation. If we substitute $x + \Delta x$ for y in (12.30), we again obtain (12.12):

$$u(\Delta x \,|\, x) = f(x, x + \Delta x) = q \log \frac{x + \Delta x}{x}. \tag{12.30}$$

So, Bernoulli's 1738 utility function (12.30) is the only function that is consistent with the desiderata of path independence and unit invariance, respectively, (12.15) and (12.16), and the boundary condition that an objective monetary gain should be assigned a positive subjective value, (12.28). Any other utility function will be in violation of these fundamental desiderata and specific boundary condition.

---

[2]See Section 12.6.1.

Note that Fechner re-derived Bernoulli's utility function in 1860 as the Weber-Fechner law that guides our sensory perception. In the years that followed (12.30) proved to be so successful that it established psychology as a legitimate experimental science, as this function, amongst other things, gave rise to our decibel scale [25].

Also note that there is one other consistent distance function apart from (12.30), which may be derived as follows. Again, we may use (12.25) to go from x to z via x itself:

$$\psi\left(\frac{z}{x}\right) = \psi\left(\frac{x}{x}\right)\ \psi\left(\frac{z}{x}\right), \tag{12.31}$$

from which it follows that under ψ the distance of x relative to itself must be one:

$$f(x,x) = \psi\left(\frac{x}{x}\right) = 1. \tag{12.32}$$

By way of (12.32) and because of our intended use of the distance function f as subjective stimulus perception function of increments in objective stimulus strengths, we may formulate the boundary condition that if x is our initial position, then a larger position y should correspond with a distance greater than one:

$$f(x,y) = \psi\left(\frac{y}{x}\right) > 1 \quad \text{for } y > x. \tag{12.33}$$

Functional equation (12.25) and the boundary condition (12.33), together with the assumption of differentiability, are sufficient to find the function f that quantifies the perceived increase associated with going from the smaller quantity x to the greater quantity y. And this function turns out to be Stevens' power law:

$$f(x,y) = \left(\frac{y}{x}\right)^q, \qquad q \geq 0, \tag{12.34}$$

where x is our initial position and y is the final asset position, and q is some arbitrary constant which has to be obtained by way of psychological experimentation.

We summarize, given the desiderata (12.15) and (12.16), the Bernoulli utility function (12.30), or, equivalently, the Fechner-Weber law, results from the boundary condition that positive increments result in positive utilities, (12.28). On the other hand, Stevens' power law (12.34) results from the boundary condition that positive increments result in subjective utilities (12.33). Stated more succinctly, the Bernoulli utility function, or, equivalently, the Fechner-Weber law, and Stevens' power law are equivalent in content, differing only in the proposed utility scale.

For a discussion of the Fechner-Weber law vs. Stevens' power law controversy, we refer the interested reader to [98].

12.4 The Utility of Debt and Debt Relief

Until now we have treated only the case were the maximal loss did not exceed the initial wealth m. However, in real life we may lose more than we actually

have, by way of debt. So, we now proceed to assign utilities to increments in debt.

According to Bernoulli's utility function, or, equivalently, the Weber-Fechner law, we cannot lose more money than we initially had. Otherwise we may have that the ratio in the logarithm in Bernoulli's utility function (12.12) may become negative, leading to a breakdown of the logarithm. But whenever we incur a debt we lose more money than we have. Furthermore, we can have both debt and wealth at same time. So, we now propose that there are two different monetary stimuli dimensions in play. The first dimension is an actual wealth dimension and the second dimension a debt dimension.

In order to derive the utility of increments in debt, we may follow the line of reasoning of the previous section. But since a decrease in debt must correspond with a positive utility, we now must replace the boundary condition (12.28) with the alternative boundary condition

$$f(x,y) = \theta\left(\frac{y}{x}\right) < 0 \quad \text{for } y > x. \tag{12.35}$$

If we assume differentiability, then (12.24) and (12.35) are sufficient to find the function f that quantifies the perceived decrease associated with going from an initial debt position x to the final debt position y. This function turns out to be the negative Bernoulli utility function:

$$f(x,y) = -q\log\frac{y}{x}, \qquad q \geq 0, \tag{12.36}$$

where q is some arbitrary constant which has to be obtained by way psychological experimentation.

In order to differentiate in our notation between the utility of wealth function (12.12), we let the utility of debt be in all Greek letters:

$$\upsilon\left(\Delta\chi \mid \chi\right) = -\gamma\log\frac{\chi + \Delta\chi}{\chi}, \qquad -\chi + \chi_0 < \Delta\chi < \infty, \tag{12.37}$$

where we let χ be the initial debt, $\Delta\chi$ the increment in debt, χ_0 the threshold of debt which is still significant, and γ the Weber constant of a monetary debt.

The threshold of debt has the following interpretation. If the repaying of one's debt results in a debt that is smaller than this limit χ_0, then we are left with an amount of debt which for all intents and purposes is equivalent to being debt-free.

If we want to give a graphical representation of (12.37) or compute the utiles of a debt increment $\Delta\chi$ conditional on a current debt position χ, then the Weber constant γ must be set to some numerical value. We will assume the Weber constant γ of increments in debt to be smaller or equal to the Weber constant q of increments in wealth, (12.14):

$$\gamma \leq q = 100. \tag{12.38}$$

The reason that we assume the Weber constant of debt to be smaller or equal to the Weber of constant of wealth is that the losing of actual monies, in the here

and now, is quite concrete, whereas the accruing of a debt can be somewhat more abstract, especially so, if the repayment of this debt can be postponed to some distant future.

In other words, it is speculated that the Weber constant of debt is a discounted constant; i.e, $\gamma(t_1) > \gamma(t_2)$ for $t_1 < t_2$. But we will leave the issue for future (experimental) research, as we proceed with our general discussion of the debt utilities.

Suppose that a student has a student loan which has accumulated to forty thousand euros. Then, by way of (12.37) and the upper bound of (12.38), we obtain the mapping of debt increments to utilities which is given in Figure 12.3.
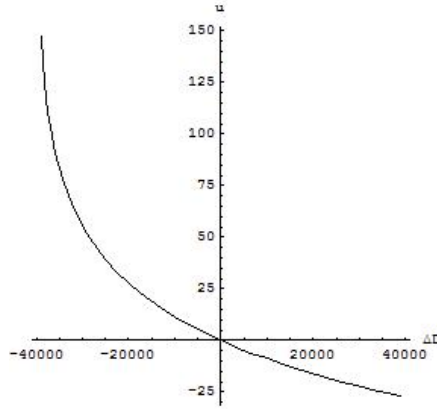
Figure 12.3: Utility plot for initial debt 40,000 euros

As stated previously, loss aversion is the phenomenon that the disutility of losses may loom larger than the utility of equivalent gains. In Figure 12.3 we see the mirror phenomenon that the utility of debt reduction may loom larger than the disutility of an equivalent debt increase. We will call this corollary of the negative Bernoulli utility function 'debt relief', the relief of paying off one's debts, or, alternatively, 'jubilee joy', the joy of being released of one's debts.

12.5 Debt Relief Examined

Now, does the phenomenon of debt relief correspond with a real psychological phenomenon? We belief that it actually does. If we have a debt of a thousand euros, then we can imagine ourselves feeling greatly relieved were we to be released of our debt. Now, if our debt, instead, is to be doubled to two thousand euros, then we can also imagine ourselves feeling unhappy about this. But introspection would suggest that this feeling of unhappiness about the doubling of our debt would be of a lesser intensity than the corresponding relief of having our debt acquitted.

We will now look at the practical implications of the negative Bernoulli utility function. A student loan initially represents a gain in the debt stimulus. This debt makes itself felt, in terms of actual loss of wealth, only after graduation, the moment the monthly payments have to be paid and take a considerable chunk out of one's actual income.

If the student of Figure 12.3, having become a PhD, and having a net income of fifteen hundred euros, is called upon to make good on his loan, by way of monthly payments of five hundred euros. Then these payments represent both a loss in income, having a negative utility of, (12.12) and (12.14):

$$u\left(-500|\,1500\right) = 100\log\frac{1500-500}{1500} = -41.5 \text{ utile}, \qquad (12.39)$$

as well as a decrements in debt, having a positive utility of, (12.37) and the upper bound of (12.38):

$$v\left(-500|\,40{,}000\right) = -100\log\frac{40{,}000-500}{40{,}000} = 1.3 \text{ utile}, \qquad (12.40)$$

which makes for a net disutility of

$$u\left(-500|\,1500\right) + v\left(-500|\,40{,}000\right) = -40.2 \text{ utile} \qquad (12.41)$$

for these payments.

And it follows that this PhD student can find little to no comfort in the fact that he is paying of his debt, as he acutely feels the sting of his loss of income (with a disutility of 41.5 utile) much more severely than the relief (with a utility of 1.3 utile) of reducing his debt. This is because his utility function for income is highly non-linear in the neighborhood of the increment, whereas his utility function for debt is highly linear in that region, as there is much more debt y then there is wealth x.

Now, say that we have another PhD student, who during his student days lived a more frugal life style and, consequently, only has a debt of two thousand euros. When called upon to make good on the loan, the loss of income will be felt by this PhD student just as keenly, with a negative utility of $u = -41.5$, (12.39). He, nonetheless, will find more satisfaction in the fact that he is paying of his debts, (12.37) and the upper bound of (12.38):

$$v\left(-500|\,2000\right) = -100\log\frac{2000-500}{2000} = 28.8 \text{ utile}, \qquad (12.42)$$

seeing that he has a more curved utility function for debt than our previous PhD student.

Note that the positive and negative Bernoulli utility functions, (12.12) and (12.37), respectively, predict that for the poor, having a minimum monthly wage of only seven hundred euros together with a large debt of twenty thousand euros, for example, a loss of income of five hundred euros will be perceived to be so much more devastating than a two-fold increase in debt of a thousand euros.

Since the loss of five hundred euros on an spendable income of seven hundred euros has a negative utility of -125 utile, whereas the increase of a debt of twenty thousand euros by a thousand euros has a negative utility of only -5 utile.

Alternatively, a loan of a thousand euros which constitutes a direct increase of a thousand euros in an spendable income of seven hundred euros will have a positive utility of 89 utile, whereas the negative utility of an increase of a thousand euros in an already existing debt of twenty thousand euros only will have a negative utility of -5 utile, which makes for a net positive utility of 84 utile for such a loan. So, the two Bernoulli utility functions, taken together, also underline the need for consumer protection in regards to predatory lending, as they predict that the temptation for the poor to take out large loans will be quite great [38].

12.6 The Fechner and Stevens' Derivations

In order to underline the ubiquitousness of Bernoulli's utility function, we now give the alternative derivations of this function by the psycho-physicists (i.e., experimental psychologists) Fechner and Stevens.

12.6.1 Fechner's Derivation

Let y signify the subjective stimulus intensity and let x signify the objective stimulus strength. Fechner takes as the starting point of his derivation of the Weber-Fechner law the empirical finding by Weber [111] that the increment in objective stimulus strength Δx needed to elicit a judgment that the increase $x + \Delta x$ is just noticeably different from x is proportional to the ratio between Δx and x.

Stated differently, the increment in stimulus strength Δx needs to be a constant fraction of the baseline stimulus strength x, in order for it to be just noticeably different:

$$\Delta x = \frac{x}{q}. \tag{12.43}$$

where q is the Weber constant which is dependent upon the specific type of sensory stimulus offered, or, equivalently, [26]:

$$\text{Just Noticeable Difference (JND)} = q\frac{\Delta x}{x}. \tag{12.44}$$

Fechner generalizes this Weber law (12.44) by postulating the mathematical auxiliary principle [77], which states that all small differences in subjective stimulus intensity Δy, and not only the ones that are just noticeably different, are proportional to the ratio between Δx and x:

$$\Delta y = q\frac{\Delta x}{x}. \tag{12.45}$$

Dividing both sides of (12.45) by Δx gives

$$\frac{\Delta y}{\Delta x} = q\frac{1}{x}. \tag{12.46}$$

In order to come to the Weber-Fechner law, Fechner then needs to take his mathematical auxiliary principle to its limit, as he must make the assumption that just as a small quantity of the physical Δx can be reduced without limit to the differential dx, so a small quantity of the subjective Δy can be reduced without limit to the differential dy [77]. By way of this assumption, we may let (12.46) tend to the differential equation

$$\frac{dy}{dx} = q\frac{1}{x}, \tag{12.47}$$

The general solution of this differential equation gives the following sensation strength function:

$$y = q\log x + c, \tag{12.48}$$

where c is some constant of integration.

Fechner then introduces the boundary condition which states that a non-perceptible objective stimulus strength x_0 should correspond with an experienced stimulus intensity of zero:

$$0 = q\log x_0 + c. \tag{12.49}$$

This boundary condition solves for the constant of integration as

$$c = -q\log x_0. \tag{12.50}$$

Substituting (12.50) into (12.48), we obtain the Weber-Fechner law [77]:

$$y = q\log \frac{x}{x_0}. \tag{12.51}$$

However, Fechner could have justified his boundary condition (12.49), with even greater generality, by stating that an objective stimulus strength x that corresponds with the reference stimulus strength x_0 should correspond an with experienced increment in stimulus intensity of zero. All the more so, since the Weber law, from which (12.48) is derived, is about differences Δx in general objective stimulus strengths x which are just noticeable [77], rather than the just noticeable difference relative to the threshold value x_0 only.

Also, Fechner erroneously assumed that in Bernoulli's formulation x_0 stands for the threshold value in wealth that is no longer perceivable [77], rather than an arbitrary reference asset position [6]. Consequently, Fechner thought his Weber-Fechner law (12.51) to be identical to the utility function (12.9) which had been proposed a century earlier by Bernoulli [77].

Moreover, Fechner believed his derivation to be the more general, as he argued that Bernoulli's derivation only applied to the special case of monetary

utilities, whereas his law, though identical, applied to all sensations, as it invokes Weber's law. However, as pointed out in [77], Fechner failed to provide any compelling reason why the principles employed in Bernoulli's derivation of the subjective value of objective monies should not be extendible to sensations in general.

But it may have been that Fechner might have felt that a law that assigned subjective values to objective monies was too arbitrary and sordid a foundation for the lofty purpose he wished it to serve[3]. In contrast, the initial Weber law allowed Fechner to forgo of the money argument and derive a law, which though in form identical to Bernoulli's, differed in that it had been constructed by way of sensory stimulus argument.

12.6.2 Stevens' Derivation

Let y signify the subjective stimulus intensity and let x signify the objective stimulus strength. Then Stevens' power law [83, 12], which was pushed to the forefront some hundred years after the Weber-Fechner law [98], is based on the observation that it is the ratio $\Delta y/y$, rather than the difference Δy, that is proportional to $\Delta x/x$. This observation translates to the mathematical identity

$$\frac{\Delta y}{y} = q\frac{\Delta x}{x}. \tag{12.52}$$

Letting the differences in y and x go to differentials, we may rewrite (12.52) as

$$\frac{dy}{y} = q\frac{dx}{x}. \tag{12.53}$$

This equation has its general solution

$$\log y = q\log x + c'. \tag{12.54}$$

Taking the exponent of both sides of (12.54), we get the power law for stimulus perception

$$y = c\,x^q, \tag{12.55}$$

where $c = \exp(c')$. The power law is applied by letting subjects compare the subjective stimulus intensity ratio of y_2 to y_1 for corresponding objective stimuli strengths x_2 and x_1:

$$\frac{y_2}{y_1} = \left(\frac{x_2}{x_1}\right)^q. \tag{12.56}$$

Stevens finds the power law to hold for several sensations; binaural and monaural loudness, brightness, lightness, smell, taste, temperature, vibration duration,

---

[3]Fechner felt a deep need to find some kind of harmony between the physical and mental universes, and the Weber-Fechner law provided him with this harmony, for this law pointed to the basic oneness of the physical and mental universes, as it demonstrated that both universes adhere to seemingly mechanistic laws. It then followed that the freedom of the latter universe, in terms of free will and volition, implied, by way of analogy, a commensurate freedom of the former; thus, opening the way for the possibility of a be-souled physical universe [25].

repetition rate, finger span, pressure on palm, heaviness, force of hand grip, autophonic response, and electric shock [98].

Let $x_1 = x$ and $x_2 = x + \Delta x$ where Δx is some increment, then we may rewrite (12.56) as

$$\frac{y_2}{y_1} = \left(\frac{x + \Delta x}{x} \right)^q . \qquad (12.57)$$

For an increment of $\Delta x = 0$, the ratio of perception stimuli will be $y_2/y_1 = 1$. Taking the log of the ratio (12.57) we may map the ratio of perceived stimuli to a corresponding scale where an zero increment in objective stimulus strength Δx corresponds with a zero increment in subjective stimulus intensity Δy:

$$u\left(\Delta x \,|\, x \right) = \log \frac{y_2}{y_1} = q \log \frac{x + \Delta x}{x}, \qquad (12.58)$$

which is just the Bernoulli's utility function (12.11), or, equivalently, if we take as our baseline x_1 in (12.56) the minimal threshold value x_0, the Weber-Fechner law (12.51). This finding is commensurate with the consistency derivation in Section 12.3, in which it was demonstrated that the Bernoulli utility function, or, equivalently, the Fechner-Weber law, is equivalent in content to Stevens' power law, differing only in the proposed utility scale.

Chapter 13

The Bayesian Decision Theory

In this chapter a theoretical discussion of the Bayesian decision theory is given [105]. This is done by relating the Bayesian decision theory to the expected outcome theory and Bernoulli's expected utility theory.

13.1 Expected Outcome Theory

Expected outcome theory has been around since the 17th century, when the rich merchants of Amsterdam sold and bought expectations as if they were tangible goods. And it would seem to many that a person acting in pure self-interest should always behave in such a way as to maximize his expected profit [47].

Let each possible decision D_j have associated with it an outcome probability distribution:

$$p(x_i | D_j) = P(X = x_i | D_j) = \begin{cases} \theta_1, & X = x_1, \\ \theta_2, & X = x_2, \\ \vdots & \vdots \\ \theta_n, & X = x_n, \end{cases} \tag{13.1}$$

and a corresponding expected value:

$$E(X | D_j) = \sum_{i=1}^{n} x_i \, p(x_i | D_j), \tag{13.2}$$

where the x_i are monetary outcomes (i.e., increments in wealth) and the θ_i their corresponding probabilities. Then the algorithmic steps of expected outcome theory are very simple:

1. For each possible decision construct an outcome probability distribution; i.e., for each possible decision, assign to every conceivable contingency both an estimated net-monetary outcome and a probability.

2. Choose that decision which maximizes the expected values (i.e., means) of the outcome probability distributions.

In closing, it is to be noted that the notion of the expectation of profit was very intuitive to the first workers in probability theory, even more so than the notion of the probability of profit [47]. For example, in Bernoulli's 1738 paper the following can be read [6]:

> "Ever since mathematicians first began to study the measurement of risk there has been general agreement on the following proposition: *Expected values are computed by multiplying each possible gain by the number of ways in which it can occur, and the dividing the sum of these products by the total number of possible cases where, in this theory, the consideration of cases which are all of the same probability is insisted upon.* If this rule be accepted, what remains to be done within the framework of this theory amounts to the enumeration of all alternatives, their breakdown into equi-probable cases and, finally, their insertion into corresponding classifications."

So, if k_1, k_2, ..., k_m, be the numbers which indicate the number of ways in which m distinct equi-probable classifications of gains in wealth x_1, x_2, ..., x_m, can occur, then it was proposed by the first workers in probability theory that the expected value be computed by way of

$$E(X) = \frac{k_1 \, x_1 + k_2 \, x_2 + \cdots + k_m \, x_m}{k_1 + k_2 + \cdots + k_m}, \tag{13.3}$$

rather than by way of the (currently) more customary (13.2).

13.2   Bernoulli's Expected Utility Theory

In the 18th century Bernoulli provided a fundamental contribution to expected outcome theory in that he proposed that it were not the actual gains and losses, but rather the utility of these gains and losses that move us to action. Moreover, Bernoulli offered up (12.12) as the function by which to translate these gains and losses to their corresponding utilities, as discussed in the previous chapter:

$$u_i = q \log \frac{m + x_i}{m}, \tag{13.4}$$

where m is the initial wealth position of the decision maker and q is some scaling constant that falls away in the decision theoretical (in)equalities.

Let the decision maker have an initial wealth position of m and let each possible decision D_j have a associated with it an outcome probability distribution

$p(x_i| D_j)$. Then each outcome probability distribution has associated with it an utility probability distribution, (13.1) and (13.4),

$$p(u_i| D_j) = P(U = u_i| D_j) = \begin{cases} \theta_1, & U = u_1 = q \log \frac{m+x_1}{m}, \\ \theta_2, & U = u_2 = q \log \frac{m+x_2}{m}, \\ \vdots & \vdots \\ \theta_n, & U = u_n = q \log \frac{m+x_n}{m}, \end{cases} \tag{13.5}$$

and an expected utility:

$$E(U| D_j) = \sum_{i=1}^{n} u_i \, p(u_i| D_j) . \tag{13.6}$$

The algorithmic steps of expected utility theory are then as follows [6]:

1. For each possible decision construct an outcome probability distribution; i.e., for each possible decision, assign to every conceivable contingency both an estimated net-monetary outcome and a probability.

2. Transform outcome probability distributions to their corresponding utility probability distributions; i.e., convert the outcomes of the outcome probability distributions to their corresponding utilities, using the utility function (13.4).

3. Choose that decision which maximizes the expected values (i.e., means) of the utility probability distributions.

Bernoulli, having provided both the concept and the quantification of utilities, proposed his expected utility theory as a straightforward generalization of expected outcome theory and, in doing so, identified the initial wealth position m as a new and important decision theoretical parameter, next to the monetary gains and losses x_i and their probabilities θ_i of the then predominant expected outcome theory.

And it is to be noted that in expected utility theory the initial wealth m functions as a reference point in the following sense. For increments x_i which are large relative to the initial wealth m the utility function (13.4) becomes non-linear, as losses are weighted more severely than like gains, whereas for increments x_i which are small relative to the initial wealth m the utility function (13.4) becomes linear, as losses and like gains are approximately weighted the same; see Figures 12.1 and 12.2, respectively.

13.3   The Bayesian Decision Theory

The Bayesian decision theory is a neo-Bernoullian decision theory in that it proposes that Bernoulli's utility function is the most appropriate function by which to translate, for a given initial wealth, gains and losses to their corresponding

utilities and it does so both on the strength of the consistency derivation presented in Section 12.3, as well as the proven empirical track record of the utility function, in its Fechner-Weber law re-incarnation, as a general model for human sense perception of increments in sensory stimuli.

But the Bayesian decision theory deviates from both the expected outcome and the expected utility theories in that it questions the appropriateness of the criterion of choice where one has to choose that decision that maximizes the expected values (i.e., means) of the outcome probability distributions under the different decisions. And by doing so, it takes a cue from the behavioral economists who have shown, by way of hypothetical betting experiments, that expected utility theory, which takes as its implied position measure the expected value, may suggest to us decisions which are forcefully rejected by our common sense [103].

The basic tenet of the Bayesian decision theory is that what we wish to maximize in our decision making is the positions of our utility probability distributions. It then follows from the observed discrepancy between the predictions made by expected utility theory and the observed betting preferences in psychological laboratory experiments, the very bedrock upon which the behavioral economy paradigm of the non-rational chooser is founded [55], that the expected value may be a suboptimal position measure for at least some probability distributions.

Moreover, in the absence of a formal (consistency) derivation of what should constitute a position measure for a given probability distribution[1], this position measure constitutes a degree of freedom. In what follows we will take advantage of the this degree of freedom in the third algorithmic step of Bernoulli's expected utility theory, as we search for an alternative, less problematic position measure which will allow us to retain the hypothesis of a decision maker.

13.3.1   The Criterion of Choice as a Degree of Freedom

Let D_1 and D_2 be two decisions we have to choose from. Let x_i, for $i = 1, \ldots, n$, and x_j, for $j = 1, \ldots, m$, be the monetary outcomes associated with, respectively, decisions D_1 and D_2. Then we first construct the two outcome distributions that correspond with these decisions:

$$p(x_i | D_1), \qquad \text{and} \qquad p(x_j | D_2), \tag{13.7}$$

where we note that in the Bayesian decision theory the outcome probability distributions are considered to be the information carriers which represent our state of knowledge regarding the consequences of our decisions.

We then proceed to map the monetary outcomes x_i in (13.7), for a given initial asset position m, to their corresponding utilities u_i, by way of the Bernoulli utility function (13.4), and we do the same for the monetary outcomes x_j under decision D_2. This mapping, which for continuous probability distributions

---

[1]See also Section (13.4.4).

will be by way of a change of variable, leaves us with the utility probability distributions:

$$p(u_i | D_1), \qquad \text{and} \qquad p(u_j | D_2). \tag{13.8}$$

Now, our most primitive intuition regarding the utility probability distributions (13.8) is that the decision which corresponds with the utility probability distribution which lies most to the right will also be the decision that promises to be the most advantageous. So, when making a decision we ought to compare the positions of the utility probability distributions on the utility axis and then choose that decision which maximizes the position of these utility probability distributions.

This all sounds intuitive enough. But how do we define the position of a probability distribution?

13.3.2 Worst-, Likely-, and Best-Case Scenarios

The expected value,

$$E(U) = \sum_{i=1}^{n} u_i \, p(u_i), \tag{13.9}$$

is a measure of the location of the center of mass of a given probability distribution. As such, it may give us a probabilistic indication of the most likely scenario[2]. From the introduction of expected outcome theory in the 17th century and expected utility theory in the 18th century onwards, the implicit assumption has been that the expected value of a given probability distribution is an optimal criterion of choice [6, 47].

Alternatively, in the Value at Risk (VaR) methodology used in the financial industry the probabilistic worst-case scenarios are taken as a criterion of choice, rather than the likely scenarios (13.9). In the VaR methodology the probabilistic worst-case scenario is operationalized as the first, or up to the fifth, percentile of an outcome probability distribution [50]. But instead of percentiles one may also use the confidence lower bound to operationalize a probabilistic worst-case scenario.

We now proceed to introduce the concept of the undershoot corrected lower confidence bound. The absolute worst-case scenario is

$$a = \min(u_1, \ldots, u_n), \tag{13.10}$$

which is also known as the minimax criterion of choice [74]. The k-sigma lower bound of a given probability distribution is given as

$$lb(k) = E(U) - k \, \mathrm{std}(U), \tag{13.11}$$

---

[2]Note that we say 'probabilistic indication' in order to point to the fact that the expected value, or, equivalently, the mean, need not give a value that one would necessarily expect, seeing that centers of mass of discrete probability distributions, more often than not, are located at 'impossible' values [47].

where k is the sigma level of the lower bound and where, (13.9),

$$\text{std}(U) = \sqrt{\sum_{i=1}^{n} u_i^2\, p(u_i) - [E(U)]^2} \qquad (13.12)$$

is the standard deviation. The probabilistic worst-case scenario then may be quantified as an undershoot corrected lower bound, (13.10) and (13.11):

$$LB(k) = \begin{cases} a, & lb(k) < a, \\ E(U) - k\,\text{std}(U), & lb(k) \geq a. \end{cases} \qquad (13.13)$$

Note that the probabilistic worst-case scenario (13.13) holds the minimax criterion of choice (13.10) for k sufficiently large to ensure $lb(k) < a$ in (13.11). For $k = 1$, the criterion of choice (13.13) constitutes a highly likely worst-case scenario (in the probabilistic sense).

Now, we may also imagine, in principle, a decision problem in which we are interested only in the probabilistic best-case scenarios. The absolute best-case scenario is

$$b = \max(u_1, \ldots, u_n), \qquad (13.14)$$

which is also known as the maximax criterion of choice [74]. The k-sigma upper bound of a given probability distribution is given as

$$ub(k) = E(U) + k\,\text{std}(U), \qquad (13.15)$$

where k is the sigma level of the upper bound. The probabilistic best-case scenario then may be quantified as an overshoot corrected upper bound, (13.14) and (13.15):

$$UB(k) = \begin{cases} E(U) + k\,\text{std}(U), & ub(k) \leq b, \\ b, & ub(k) > b. \end{cases} \qquad (13.16)$$

Note that the probabilistic best-case scenario (13.16) holds the maximax criterion of choice (13.14) for sufficiently large k to ensure $ub(k) > b$ in (13.15). For $k = 1$, the criterion of choice (13.16) constitutes a highly likely best-case scenario (in the probabilistic sense).

If we use the criterion of choice (13.9), then we will neglect what may happen in the worst and the best of all worlds. If we use the criterion of choice (13.13), then we will neglect what may happen in the most likely and the best of all worlds. If we use the criterion of choice (13.16), then we will neglect what might happen in the worst and the most likely of all worlds.

An exclusive commitment to any one of the criteria of choice (13.9), or (13.13), or (13.16), will necessarily lead us to leave out some pertinent information in our decision theoretical considerations. So, how do we untie this Gordian knot?

13.3.3 A Balanced Probabilistic Hurwicz Criterion of Choice

In Hurwicz's criterion of choice the absolute worst- and best-case scenarios are both taken into account. For a balanced pessimism coefficient of $\alpha = 1/2$, we have [39]

$$\text{Hurwicz's criterion} = \frac{a + b}{2}, \tag{13.17}$$

where a is the minimax criterion (13.10) and b is the maximax criterion (13.14). Now, if we replace the absolute worst- and best-case scenarios in (13.17) with their corresponding probabilistic counterparts, (13.13) and (13.16), then we obtain

$$\text{probabilistic Hurwicz criterion} = \frac{LB(k) + UB(k)}{2}. \tag{13.18}$$

Under the criterion of choice (13.18), indecision between D_1 and D_2 translates to the decision theoretical equality:

$$\frac{LB(k|D_1) + UB(k|D_1)}{2} = \frac{LB(k|D_2) + UB(k|D_2)}{2}, \tag{13.19}$$

or, equivalently,

$$LB(k|D_1) - LB(k|D_2) = UB(k|D_2) - UB(k|D_1), \tag{13.20}$$

a trade-off between the gains/losses in the probabilistic worst-case scenarios and the corresponding losses/gains in the probabilistic best-case scenarios; i.e., if $LB(k|D_1)$ is greater than $LB(k|D_2)$ by some value, then $UB(k|D_1)$ must be smaller than $UB(k|D_2)$ with that same value, in order for the decision theoretical equalities (13.19) and (13.20) to hold.

It follows, seeing that (13.17) is a limit case of (13.18), as k grows large, that, for a balanced pessimism coefficient of $\alpha = 1/2$, Hurwicz's criterion of choice provides a balanced trade-off between the differences in the absolute worst-case scenarios and the differences in the absolute best-case scenarios.

The probabilistic Hurwicz criterion of choice (13.18) translates to the *locus* (i.e., position measure)

$$\text{loc}(k) = \frac{LB(k) + UB(k)}{2}, \tag{13.21}$$

where, because of the under- and overshoot corrections, (13.13) and (13.16),

$$\frac{LB(k) + UB(k)}{2} = \begin{cases} E(U), & lb(k) \geq a, \ ub(k) \leq b, \\ \frac{1}{2}\left[a + E(U) + k\,\text{std}(U)\right], & lb(k) < a, \ ub(k) \leq b, \\ \frac{1}{2}\left[E(U) - k\,\text{std}(U) + b\right], & lb(k) \geq a, \ ub(k) > b, \\ \frac{1}{2}\left(a + b\right), & lb(k) < a, \ ub(k) > b. \end{cases} \tag{13.22}$$

It follows that the alternative criterion of choice (13.18), which takes into account what may happen in the worst- and best-case scenarios, holds both the traditional expected value criterion of choice (13.9) as a special case, when neither a lower bound undershoot nor an upper bound overshoot occurs, as well as Hurwicz's criterion of choice with a balanced pessimism factor (13.17), when both a lower bound undershoot and an upper bound overshoot occurs.

13.3.4   Weaver's Criterion of Choice

It may be found that the criterion of choice (13.18), and by implication also the Hurwicz criterion of choice (13.17), is vulnerable to a simple counter-example.

Imagine two utility probability distributions having equal lower and upper bounds $LB(k)$ and $UB(k)$, but one distribution being right-skewed and the other being left-skewed. Then the criterion of choice (13.18) will leave its user undecided between the two decisions. Intuition, however, will give preference to the decision corresponding with the left-skewed distribution, as the bulk of the probability distribution of the left-skewed distribution will be more to the right than that of the right-skewed distribution [109].

It follows that a criterion of choice, in order to be as universal as possible, should not only take into account the trade-off between the probabilistic worst- and best-case scenarios, as is done in (13.21), but also the location of the probabilistic bulk of the probability distribution.

Weaver's criterion of choice[3] is a position measure for a probability distribution which not only takes into account not the trade-off between the probabilistic worst- and best-case scenarios, but also the location of the bulk of the probability density in a unimodel probability distribution, thus accommodating the intuitive preference for the left-skewed distribution of the just mentioned counter-example,

$$\text{loc}(k) = \frac{LB(k) + E(U) + UB(k)}{3}, \qquad (13.23)$$

where, because of the under- and overshoot corrections, (13.13) and (13.16),

$$\frac{LB(k) + E(U) + UB(k)}{3} = \begin{cases} E(U)\,, & lb(k) \geq a,\ ub(k) \leq b, \\ \frac{1}{3}\left[a + 2E(U) + k\operatorname{std}(U)\right], & lb(k) < a,\ ub(k) \leq b, \\ \frac{1}{3}\left[2E(U) - k\operatorname{std}(U) + b\right], & lb(k) \geq a,\ ub(k) > b, \\ \frac{1}{3}\left[a + E(U) + b\right], & lb(k) < a,\ ub(k) > b. \end{cases}$$
$$(13.24)$$

It is to be noted that this compound position measure holds the traditional expected value criterion of choice (13.9) as a special case when the k-sigma lower confidence bound (13.11) does not undershoot the absolute minimum (13.10) and the upper confidence bound (13.15) does not overshoot the absolute maximum (13.14).

The sigma level k in the probabilistic criterion of choice (13.23) is the parameter that controls the extremeness of the probabilistic worst- and best-case scenarios (13.13) and (13.16). The higher the sigma level k, the more extreme the worst-case scenario $LB(k)$ will be in terms of its utility and, consequently, the smaller the probability that the actual utility will lie below (i.e., be worse than) this k-sigma utility lower confidence bound. Likewise, the higher the sigma level k, the more extreme the best-case scenario $UB(k)$ in terms of its

---

[3]In dedication to the SP Indian Lake trooper Randy Weaver who helped us to make it to the MaxEnt 2015, Potsdam, New York in time.

utility and, consequently, the smaller the probability that the actual utility will lie above (i.e., be better than) this k-sigma utility upper confidence bound.

The balanced position measure (13.23) can be generalized, trivially, by allowing for imbalanced sigma bounds (13.13) and (13.16):

$$\text{loc}(k_1, k_2) = \frac{LB(k_1) + E(U) + UB(k_2)}{3},\qquad (13.25)$$

where k_1 and k_2 are the sigma levels of the corrected the lower and upper bounds, respectively, and where, because of the under- and overshoot corrections, (13.13) and (13.16),

$$\frac{LB(k_1) + E(U) + UB(k_2)}{3} = \begin{cases} E(U) + \frac{1}{3}(k_2 - k_1)\ \text{std}(U), & lb(k_1) \geq a,\ ub(k_2) \leq b, \\ \frac{1}{3}[a + 2E(U) + k_2\ \text{std}(U)], & lb(k_1) < a,\ ub(k_2) \leq b, \\ \frac{1}{3}[2E(U) - k_1\ \text{std}(U) + b], & lb(k_1) \geq a,\ ub(k_2) > b, \\ \frac{1}{3}[a + E(U) + b], & lb(k_1) < a,\ ub(k_2) > b. \end{cases}$$
$$(13.26)$$

This then brings us to the following question: To what values do we set the sigma levels in the confidence bounds of the general position measure (13.25)?

From a prescriptive (i.e., normative) perspective, it is proposed that one ought, as a rule of thumb, use a balanced 1-sigma position measure; i.e., $k_1 = k_2 = 1$, or, equivalently, $k = 1$ in (13.23). Since the probabilistic worst- and best-case scenarios are then given equal weight, while the 1-sigma level ensures that the probabilistic worst- and best-case scenarios will be temperate in their pessimism and optimism, respectively.

But whenever it is better to be safe than sorry, then the decision maker can be called upon to put a premium on caution while being on guard against any unwarranted optimism; i.e., $k_1 > k_2 = 1$. For example, when deciding on the safety level of our flood defenses and our nuclear facilities, we, the public, would like the decision makers to be extra attentive to the possible down-sides of their decisions without any undue focus on the potential up-sides.

From a descriptive (i.e., positivist) perspective, it is found that the balanced 1-sigma position measure, with $k_1 = k_2 = 1$, or, equivalently, $k = 1$ in (13.23), gives general good predictions when it comes to the modeling of non-linear preferences (i.e., probability weighting functions), risk-seeking in the positive domain (i.e., lotteries), and risk-aversion in the negative domain (i.e., insurances).

However, for risk-aversion in the positive domain (i.e., the Allais paradox) and risk-seeking in the negative domain (i.e., trying to avert large losses by taking a chance on even larger losses) the experimental data point to a setting of $k_1 > k_2 = 1$ and $k_2 > k_1 = 1$, respectively[4].

---

[4]These data points are discussed in Chapters 14, 15, and 16.

13.3.5   The Algorithmic Steps of the Bayesian Decision Theory

In any problem of choice one will endeavor to choose that action which has a corresponding utility probability distribution that is lying most the right on the utility axis; i.e., one will choose that action that maximizes the position of the utility probability distributions. In this there is little freedom.

But one is free, in principle, to choose the measures of the positions of one's utility probability distributions any way one see fit. Nonetheless, it is held to be self-evident that it is always a good policy to take into account all the pertinent information at hand.

If we maximize only the expected values of the utility probability distributions, then we will, by definition, neglect the information that the standard deviations of the utility probability distributions provide regarding our problem of choice, by way of the symmetry breaking in the case of an overshoot of one of the confidence bounds.

Likewise, we are free to maximize only one of the confidence bounds of our utility probability distributions, while neglecting the other. But in doing so, we will be performing probabilistic minimax or maximax analyses, and, consequently, neglect the possibilities of both potentially astronomical gains in the upper bound and potentially catastrophic losses in the lower bound.

If we only maximize the sum of the lower and upper bound, or a scalar multiple thereof, then we will make a trade-off between the probabilistic worst- and best-case scenarios, but in the process, we will, for unimodal distributions, be neglecting the location of the bulk of our probability distributions.

This is why, in our minds, the mean of the sum of the lower confidence bound, expected value, and upper bound bound, currently is the best all-round position measure for a given probability distribution, as it reflects the position of the probabilistic worst- and best-case scenarios, (13.13) and (13.16), as well as the position of the expected outcome (13.9).

So, the algorithmic steps of the Bayesian decision theory are as follows [105]:

1. For each possible decision construct an outcome probability distribution; i.e., for each possible decision, assign to and/or derive, by way of the product and sum rules, for every conceivable contingency both an estimated net-monetary consequence and a probability.

2. Transform the outcome probability distributions to their corresponding utility probability distributions; i.e., convert the outcomes of the outcome probability distributions to their corresponding utilities, using Bernoulli's utility function (13.4).

3. Maximize the position of the resulting utility probability distributions; i.e., choose that decision which maximizes the general position measure (13.25) for the utility probability distributions.

Note that for balanced k-sigma levels and neither a lower bound undershoot occurs in (13.13) nor an upper bound overshoot in (13.16) the criterion

of choice (13.25) collapses to the expected value (13.9), as the Bayesian decision theory collapses to Bernoulli's expected utility theory. It follows that the Bayesian decision theory, relative to Bernoulli's original expected utility theory [6], is nothing more than a mathematically trivial readjustment of the proposed position measure which is to be maximized.

It will be demonstrated in the remainder of this part of the thesis that this trivial readjustment, however, has some non-trivial decision theoretical implications, both in terms of actual numerical results and the resolution of paradoxes that have plagued expected utility theory for a long time.

13.4 Some Miscellanea

In this section we give some technical miscellanea, which are pertinent to the discussion of the Bayesian decision theory, but which would have interrupted the flow of exposition had we discussed them at their appropriate place.

13.4.1 Constructing Outcome Probability Distributions

The construction of the outcome probability distributions (i.e., the first step in the Bayesian decision theory) will be the most challenging part in any (non-trivial) probabilistic cost-benefit analysis, as these outcome probability distributions need to incorporate the available expert knowledge and data. In this section we give a first, rough, non-exhaustive outline on how to construct outcome probability distributions by way of the product and sum rules.

In its most abstract form, we have that each problem of choice consists of a set of potential decisions

$$D_k = \{D_1, \ldots, D_l\}.$$

Each decision D_k we make may give rise to a set of possible events

$$E_j^{(k)} = \left\{ E_1^{(k)}, \ldots, E_{m_k}^{(k)} \right\}.$$

These events $E_j^{(k)}$ are associated with the decisions D_k by way of the conditional probabilities $P\left(E_j^{(k)} \middle| D_k \right)$. Furthermore, each event $E_j^{(k)}$ allows for a set of potential monetary outcomes

$$x_i^{(jk)} = \left\{ x_i^{(jk)}, \ldots, x_{n_{jk}}^{(jk)} \right\}.$$

These outcomes $x_i^{(jk)}$ are associated with the events $E_j^{(k)}$ by way of the conditional probabilities $P\left(x_i^{(jk)} \middle| E_j^{(k)} \right)$.

By way of the product rule (4.1), we compute the bivariate probability distribution of an event and an outcome conditional on the decision taken:

$$P\left(E_j^{(k)}, x_i^{(jk)} \middle| D_k \right) = P\left(E_j^{(k)} \middle| D_k \right) P\left(x_i^{(jk)} \middle| E_j^{(k)} \right). \qquad (13.27)$$

The outcome probability distribution is then obtained by marginalizing, by way of the sum rule (4.2), over all the possible events

$$
P\left(x_i^{(jk)}\middle|\, D_k\right) = \sum_{j=1}^{m_k} P\left(E_j^{(k)}, x_i^{(jk)}\middle|\, D_k\right) = \sum_{j=1}^{m_k} P\left(E_j^{(k)}\middle|\, D_k\right) P\left(x_i^{(jk)}\middle|\, E_j^{(k)}\right).
$$
(13.28)

where the probability distribution of the $x_i^{(jk)}$ is the probability weighted sum (i.e., mixture) of the conditional probability distributions $P\left(x_i^{(jk)}\middle|\, E_j^{(k)}\right)$.

The outcome probability distribution (13.28) is the information carrier which captures our state of knowledge in regards to the plausibility of the various consequences under decision D_k. If we want to collect all the possible outcomes under a given decision in one probability distribution, then we may use a divide-and-conquer strategy in which we first enumerate and assign conditional probabilities to the different events that may follow a given decision, after which we enumerate and assign conditional probabilities to all the outcomes that may follow a given event.

13.4.2 Mapping Outcomes to Utilities

Outcomes x of outcome probability distributions $p(x|\, D_i)$ may be mapped to their corresponding utilities u in the following manners; for discrete outcome probability distributions, by way of a simple manual relabeling or, if the number of possible outcomes is too large, the use of a Dirac delta function; and for continuous outcome probability distributions, by way of a change of variable.

If we have the simple outcome probability distribution

$$
p(\, x_i|\, D_j) = \begin{cases} \theta, & x_1, \\ 1 - \theta, & x_2, \end{cases}
$$
(13.29)

where the x_i are scalars (i.e., numbers), then we may do a simple manual mapping, by way of (13.4),

$$
x_i \mapsto u_i = q \log \frac{m + x_i}{m},
$$
(13.30)

where m is the reference asset position of the decision maker. This gives the corresponding utility probability distribution

$$
p(\, u_i|\, D_j) = \begin{cases} \theta, & u_1 = q \log \frac{m+x_1}{m}, \\ 1 - \theta, & u_2 = q \log \frac{m+x_2}{m}. \end{cases}
$$
(13.31)

The kth-order moments of utility probability distributions like may be evaluated by way of the sum:

$$
E\left(U^k\middle|\, D_j\right) = \sum_i u_i^k\, p(\, u_i|\, D_j)\, du.
$$
(13.32)

In case that the number of possible outcomes is too large we may make of the Dirac delta function [47]. For example, if we have the outcome probability distribution

$$p(x_i | D_j), \quad i = 0, 1, \ldots, n, \tag{13.33}$$

then we may we introduce the conditional probability distribution $p(u | x_i)$ in order to map the outcomes x_i in (13.33) to utilities u. By way of the product rule (4.1), we then have

$$p(u, x_i | D_j) = p(u | x_i) p(x_i | D_j), \tag{13.34}$$

where the conditional probability distribution $p(u | x_i)$ takes us from the x dimension, which is the dimension of the outcomes, to the u dimension, which is the dimension of the corresponding utilities. If we marginalize (13.34) over all the possible outcomes x_i, by way of the generalized sum rule (4.4), then we may get the utility probability distribution of interest:

$$p(u | D_j) = \sum_i p(u, x_i | D_j). \tag{13.35}$$

Now, if every outcome x_i admits only one utility value u:

$$p(u | x_i) = \begin{cases} 1, & u = q \log \frac{m + x_i}{m}, \\ 0, & u \neq q \log \frac{m + x_i}{m}, \end{cases} \tag{13.36}$$

for $i = 0, 1, \ldots, n$, then we may rewrite (13.36) as

$$p(u | x_i) = \delta\left(u - q \log \frac{m + x_i}{m} \right), \tag{13.37}$$

where δ is the delta-function for which we have that

$$\delta(u - c) \, du = \begin{cases} 1, & u = c, \\ 0, & u \neq c. \end{cases} \tag{13.38}$$

Because of property (13.38), we have that

$$\int \delta(u - c) f(u) \, du = f(c). \tag{13.39}$$

This property of the delta-function enables us to make a one-on-one mapping from outcomes to utilities.

Substituting (13.37) into (13.34) and (13.35), we obtain the utility probability distribution we are looking for:

$$p(u | D_j) = \sum_i \delta\left(u - q \log \frac{m + x_i}{m} \right) p(x_i | D_j). \tag{13.40}$$

The kth-order moments of the utility probability distribution (13.40) may be evaluated by way of the integral:

$$E\big(U^k\big|\,D_j\big) = \int u^k\, p(\,u|\,D_j)\,du. \tag{13.41}$$

In the case where we have a continuous outcome probability distribution

$$p(\,x|\,D_j)\ dx = f(x)\ dx, \tag{13.42}$$

then we may make the change of variable

$$u = q\log\frac{m+x}{m}, \qquad \frac{du}{dx} = \frac{q}{m+x}, \tag{13.43}$$

from which it follows that

$$x = m\left[\exp\left(\frac{u}{q}\right) - 1\right], \qquad dx = \frac{m+x}{q}du. \tag{13.44}$$

Substituting (13.44) into (13.42), we obtain the transformed utility probability distribution

$$p(\,u|\,D_j)\ du = f\left(m\left[\exp\left(\frac{u}{q}\right) - 1\right]\ \right)\frac{m\exp\left(u/q\right)}{q}du \tag{13.45}$$

The kth-order moments of the utility probability distribution (13.45) may be evaluated by way of the integral (13.41).

Alternatively, in order to compute the moments of the utility probability distribution, we may make use of the short-cut that for a given stochastic X, or, equivalently, a probability distribution of x, we may compute the kth-order moments of a function g of that stochastic for discrete outcome probability distributions as [74]

$$E\left([g\,(X)]^k\,\big|\,D_j\right) = \sum_i [g\,(x_i)]^k\ p(\,x_i|\,D_j)\,, \tag{13.46}$$

and for continuous outcome probability distributions as [74]

$$E\left([g\,(X)]^k\,\big|\,D_j\right) = \int [g\,(x)]^k\ p(\,x|\,D_j)\ dx. \tag{13.47}$$

If we substitute

$$u = g(x) = q\log\frac{m+x}{m}, \tag{13.48}$$

into (13.46) or (13.47), we may obtain the kth-order moments of the corresponding utility probability distributions:

$$E\left(U^k\big|\,D_j\right) = E\left([g\,(X)]^k\,\big|\,D_j\right), \tag{13.49}$$

without having to explicitly construct the actual utility probability distribution itself.

13.4.3 Probabilistic Utility Functions

All the outcome probabilities used in this thesis are discrete and all the utilities used are deterministic; i.e., in principle, the delta-function (13.37)

$$p(u \mid x_i) = \delta\left(u - q \log \frac{m + x_i}{m}\right),$$

will suffice to map monetary outcomes, that is, positive or negative increments in wealth x_i, to corresponding utilities. Now, we also may envisage decision problems in which we are uncertain regarding the actual utility of a given outcome x_i in (13.33). In those cases we will want to assign probability distributions $p(u \mid x_i)$ less dogmatic than the Dirac delta distribution (13.37) to our utilities.

This may be the case if the initial wealth m in (13.4) allows for a probability distribution under some outcome. In these cases the deterministic utility (13.37) can be replaced by its probabilistic counterpart. For example, if we only have knowledge about the first two moments of the reference asset position m of the decision maker, then we may use those two moments to assign a normal probability distribution distribution to that asset reference position m:

$$p(m \mid \mu, \sigma) = \frac{C}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(m - \mu)^2}{2\sigma^2}\right], \tag{13.50}$$

where $0 \le m \le \infty$, and

$$C^{-1} = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(m - \mu)^2}{2\sigma^2}\right] dm. \tag{13.51}$$

Then by setting, (13.4),

$$u = q \log \frac{m + x_i}{m} \tag{13.52}$$

and solving for m we obtain

$$m = \frac{x_i}{e^{u/q} - 1} \tag{13.53}$$

and a corresponding Jacobian

$$J = \left| -\frac{e^{u/q} x_i}{q \left(e^{u/q} - 1\right)^2} \right|. \tag{13.54}$$

Then by substituting (13.53) into (13.50) and multiplying it by (13.54), we obtain the corresponding probabilistic utility function which is conditional on the outcome x_i and the μ and σ of the uncertain reference asset position m:

$$p(u \mid x_i, \mu, \sigma) = \frac{C \, |x_i|}{q \left(e^{u/q} - 1\right)^2 \sqrt{2\pi}\sigma} \exp\left[\frac{u}{q} - \frac{\left(\frac{x_i}{\exp^{u/q} - 1} - \mu\right)^2}{2\sigma^2}\right]. \tag{13.55}$$

If we substitute (13.55), instead of the deterministic (13.37), into (13.34) and (13.35), we obtain the utility probability distribution we are looking for:

$$p(u|\mu, \sigma, D_j) = \sum_i p(u|x_i, \mu, \sigma) \; p(x_i|D_j). \qquad (13.56)$$

13.4.4   Some Additional Footing for the Criterion of Choice

Ideally we would have some consistency derivation of what constitutes a position measure of a probability distribution, say,

$$H_n(p_1, \ldots, p_n, u_1, \ldots, u_n) \qquad (13.57)$$

where p_i are the probabilities of the values u_i, for $i = 1, \ldots, n$. However, in the absence of such a consistency derivation we have to take our recourse to *ad hoc* common sense considerations, as was done in Section 13.3. Stated differently, the criterion of choice in our decision theory still formally constitutes a degree of freedom.

We will provide here some additional footing for the position measure H_n, (13.57), which may be obtained from three simple common-sense considerations. These considerations concern the effects of a translation and a change of scale by some constant of the outcomes of a probability distribution, as well as the effect of one of the outcomes becoming certain.

Firstly, we have to assume that some numerical position measure H_n, (13.57), exists, i.e. that it is possible to set up some kind of association between the position of a probability distribution and real numbers. Secondly, we have to assume that H_n is continuous in the probabilities p_i and the outcomes u_i. Otherwise, an arbitrary small change in the outcome probability distribution would lead to a big change in the position of the outcome distribution. Thirdly, we require that this measure should correspond qualitatively with common-sense.

Qualitative correspondence with common-sense translates to a set of minimal properties we would like our position measure H to exhibit. The first of these properties is that when we translate all the outcomes with a positive quantity c, the quantity

$$h(c) = H_n(p_1, \ldots, p_n, u_1 + c, \ldots, u_n + c) \qquad (13.58)$$

should behave as

$$h(c) = h(0) + c. \qquad (13.59)$$

We also require that for a rescaling of the outcomes by a positive factor γ, the quantity

$$g(\gamma) = H_n(p_1, \ldots, p_n, \gamma \, u_1, \ldots, \gamma \, u_n) \qquad (13.60)$$

should behave as

$$g(\gamma) = \gamma \, g(1). \qquad (13.61)$$

Combining the quantities (13.58) and (13.60), we obtain the quantity

$$f(c, \gamma) = H_n[p_1, \ldots, p_n, \gamma \, (u_1 + c), \ldots, \gamma \, (u_n + c)], \qquad (13.62)$$

which, because of (13.59) and (13.61), should behave as

$$f(c, \gamma) = \gamma \, f(0, 1) + \gamma \, c. \tag{13.63}$$

The expected value, the minima and maxima, the lower and upper confidence bounds, and the weighted sums thereof, all adhere to the strong requirement (13.63).

We now give a common-sense consideration for the effect of probabilities on the value of a position measure. It seems reasonable to require that the position of the Dirac delta probability distribution $p(u \,|\, c) = \delta \, (u - c)$ is c, or, equivalently,

$$H_n(\ldots, p_k, \ldots, u_k, \ldots) \to u_k, \qquad \text{as } p_k \to 1. \tag{13.64}$$

This requirement then dictates, under the continuity of H_n, that the position measure should be bounded as

$$\min(u_1, \ldots, u_n) \le H_n \le \max(u_1, \ldots, u_n), \tag{13.65}$$

which gives us the formal rationale for the undershoot corrected lower confidence bound (13.13) and the overshoot corrected upper confidence bound (13.16).

Moreover, (13.64) dictates that the weights of the separate location elements in (13.17), (13.18), (13.21), and (13.23) should all sum to one; that is, if we take our position measure to be the mean of n separate location elements, then the weights of these elements are $1/n$ and the sum of these weights is 1.

Chapter 14

An Alternative to Probability Weighting

In prospect theory it is postulated that the decision maker is irrational in that small probabilities are overweighted and large probabilities are underweighted. Prospect theory models this irrational over- and underweighting by way probability weighting functions which map objective probabilities p in an inverted S-curve to subjective probability π. The inverted S-curve model is the working part of prospect theory as it allows prospect theory to make predictions regarding betting behavior. And the probability weighting functions of prospect theory are explicitly chosen for their ability to fit (i.e., mimic) the observed inverted S-shape of the ratios of the certain and uncertain outcomes in certainty bets.

The Bayesian decision theory [105], relative to Bernoulli's original expected utility theory [6], is nothing more than a mathematical trivial readjustment of the proposed position measure which is to be maximized. It will be demonstrated in this chapter that this trivial readjustment has some non-trivial decision theoretical implications. Since it leads us to predict, from first principles, rather than by construction, the observed inverted S-shape of the ratios of the certain and uncertain outcomes in certainty bets, while at the same time allowing us to retain the traditional hypothesis of the rational *homo oeconomicus* of economical theory.

So, the Bayesian decision theory, with its ease of implementation and its ability to accommodate the violations of expected utility theory, has the potential to bridge the divide between the decision theoretical academic literature, wherein it is widely acknowledged that (von Neumann and Morgensterns) expected utility theory is not valid as a descriptive theory of choice under risk, because of its inability to predict the observed inverted S-shape of the ratios of the certain and uncertain outcomes in certainty bets [9], and the practice of cost-benefit analysis, wherein expected utility maximizations for linear utility

functions are performed as a matter of course[1].

14.1   The Issue

Von Neumann and Morgenstern proposed in 1944 an expected utility theory, distinctly different from Bernoulli's original utility theory [6], in which the (dis)utility (i.e., subjective worth) of a risky choice, having n possible outcomes x_i with corresponding probabilities p_i, is modelled as

$$E(U) = \sum_{i=1}^{n} u(x_i) \; p_i, \qquad (14.1)$$

where u is some continuous, monotonic increasing utility function (i.e., more gain can only lead to more utility) which is to be inferred experimentally from the decision maker, based upon the decision maker's observed (betting) preferences[2] [110].

Von Neumann and Morgenstern's expected utility theory initially received a favorable reception [22]. But in the 1950s and 1960s evidence began to accumulate which suggested that von Neumann and Morgenstern's expected utility theory failed as a general descriptive theory of risky choice. The most damning evidence was by way of Allais' counter-examples, [3, 4], which showed that a number of patterns of responses to risky choices that systematically violated the axioms underlying the von Neumann and Morgenstern theorem of expected utility maximization [75].

14.2   Prospect Theory's Resolution

In reaction to the failure of von Neumann and Morgenstern's expected utility model, a new theory of choice was eventually proposed in 1979 by the psychologists Daniel Kahneman and Amos Tversky, the so-called prospect theory. Prospect theory was later on revised into cumulative prospect theory [103], as this second version of prospect theory[3] repaired the potential for violations of stochastic dominance[4].

The paper 'Prospect Theory: An Analysis of Decision Under Risk' [53] was ranked 10 years ago as the second most frequently cited paper published in economic journals since 1970 [57] and as of January 2017 held 43.453 citations in

---

[1]Guide to Cost-Benefit Analysis of Investment Projects (EU Directorate Guide Regional Policy, 2008), Project appraisal check-list (p.72): "Has the expected value criterion been used?"

[2]It is to be noted that Von Neumann and Morgenstern's utility theory is markedly different, in terms of implementation, from Bernoulli's original expected utility theory; see also Chapter 12 and Section 13.2.

[3]We shall refer to cumulative prospect theory as prospect theory. In this we follow Booij et al. [11].

[4]Stochastic dominance states that for monetary outcomes x and y both greater than zero, the bet of x with probability p and $x + y$ with probability $1 - p$ should always be preferred over the bet where we obtain x with certainty [75].

Google Scholar. And Kahneman was awarded a Nobel Prize in economics in 2002 in part for this prospect theory[5] [10], a theory which still figures prominently within the field of behavioral economics [7, 21, 55].

Prospect theory's first fundamental breakaway from von Neumann and Morgenstern's decision theory is that instead of defining preferences over wealth, preferences are defined in principle over changes with respect to a flexible reference point, often taken as the current asset position (i.e., current wealth). In prospect theory the phenomena of diminishing sensitivity and loss aversion are modeled by way of a two-part power value function [103]:

$$v(x) = \begin{cases} -\lambda \left(-x\right)^{\beta}, & x \leq 0 \\ x^{\alpha}, & x > 0 \end{cases} \tag{14.2}$$

for positive λ, α, and β. The power function (14.2) is often chosen by prospect theorists as a utility function[6] because of its simplicity and good fit to experimental data [11].

It may be noted at this point that this first breakaway of prospect theory from the von Neumann and Morgenstern's expected utility theory [110], with its introduction of the explicit utility function (14.2), in a sense is a return to the Bernoulli's expected utility theory [6], as the latter theory also proposes the explicit utility function (13.4),

$$u(x \,|\, m) = q \log \frac{m + x}{m}, \tag{14.3}$$

where m is the initial asset position and q is the Weber-constant of money [25].

Now, it is in Bernoulli's 1738 paper, paragraph 13 of [6], that one may find one of the first, if not the very first, mathematical prediction of the psychological phenomenon of loss aversion, by way of (14.3). Also, contrary to what is stated, at great length, in both [54] and [55], it is Bernoulli's utility function which treats value as a function in two arguments: the asset position that serves as a reference point, and the magnitude of the change from that reference point, as may readily be checked in [6].

Moreover, in relation to the value function (14.2), it may be read in [53] that

> "The emphasis on changes as the carriers of value should not be taken to imply that the value of a particular change is independent of initial position. Strictly speaking, value should be treated as a function in two arguments: the asset position that serves as a reference point, and the magnitude of the change (positive or negative) from that reference point."

Stated differently, Bernoulli's utility function adheres to an ideal that is also acknowledged by prospect theory itself. This point is worth belaboring, as the

---

[5]Amos Tversky died in 1996.

[6]We shall refer from time to time to the two-part power value function (14.2) as a utility function. In this we follow Booij et al. [11].

alternative prospect theoretical utility function (14.2), in contrast, treats value as a function in only one argument: the magnitude of the change (positive or negative) from that reference point. A fact that prospect theory justifies by arguing that [53]

> "[...] the preference order of prospects is not greatly altered by small or even moderate variations in asset position. [...] Consequently, the representation of value as a function in one argument generally provides a satisfactory approximation."

Prospect theory's second fundamental breakaway from the von Neumann and Morgenstern model is that linearity in probability is replaced by a subjective probability weighting function that is assumed to have an inverse-S shape, reflecting increased sensitivity toward changes in probabilities near zero and one. The replacement of linearity in the probabilities by a subjective probability weighting function accommodates anomalies like the Allais 1953 counterexamples [11]. And the very act of probability-weighting is said to fall outside the strictures of standard economic 'rationality' [103]:

> "The idealized assumption of rationality in economic theory is commonly justified [on the] grounds [...] that only rational behavior can survive in a competitive environment [. But] evidence indicates that people can spend a lifetime in a competitive environment without acquiring a general ability to avoid [...] to apply linear decision weights."

Let x_C be a certainty outcome, either positive or negative, which has a probability of one and let x_U be an uncertainty outcome with the same sign as x_C which has a probability of p. Then it is postulated by prospect theory that the following decision theoretical equality constitutes fairness for this certainty bet:

$$v(x_C) = \pi_1^{\pm} v(0) + \pi_2^{\pm} v(x_U)$$

$$= \left[1 - w^{\pm}(p) \right] v(0) + w^{\pm}(p) \, v(x_U) \qquad (14.4)$$

$$= w^{\pm}(p) \, v(x_U)$$

where '$\pm$' is the plus-minus sign, and where π^- and π^+ are the probability weights for negative and positive risky prospects, respectively, that are constructed by way of the corresponding cumulative probability weighting functions w^- and w^+, [103].

It follows from (14.4) that for the specific case of certainty bets, in which we have to chose between a certain x_C outcome and an uncertain outcome x_U, the probability weighting functions $w^{\pm}$ admit the following identity:

$$w^{\pm}(p) = \frac{v(x_C)}{v(x_U)}, \qquad (14.5)$$

where $|x_C| \leq |x_U|$. For a linear utility function v, or, equivalently, $\alpha = \beta = \lambda = 1$ in (14.3), the prospect theoretical probability weighting functions $w^{\pm}$ in (14.5) simplify to [103]:

$$w^{\pm}(p) = \frac{x_C}{x_U}. \tag{14.6}$$

If certainty outcomes x_C are elicited for varying combinations of probabilities p and uncertain outcomes x_U in fair certainty bets, then by plotting for each p the observed ratios x_C/x_U there is found an inverse-S shape for $w^{\pm}(p)$.

It is postulated by prospect theorists that this inverse-S shape reflects an increased sensitivity toward changes in probabilities near zero and one, in which small probabilities are over-weighted and moderate and high probabilities are under-weighted [103].

Moreover, the observed inverse-S shape of the ratio x_C/x_U as a function of p constitutes a serious challenge to expected utility theory [103]. Since expected utility theory, by way of its implicit assumption of linearity in the probabilities [6, 110], would predict for a linear utility function u, or, equivalently, $x/m \to 0$ in (14.4), the fair-bet relation

$$x_C = x_U \, p. \tag{14.7}$$

But this fair-bet relation results in the prediction that the ratio x_C/x_U should be linear in the probability p, (14.7):

$$\frac{x_C}{x_U} = p, \tag{14.8}$$

which is in contradiction with the experimentally observed non-linear inverse-S shape.

In order to accommodate the observed inverse-S shape of the ratios x_C/x_U as a function of the probability p of the uncertain outcome, it was initially proposed to use the function

$$w^{\pm}(p) = \frac{p^{\gamma}}{[p^{\gamma} + (1-p)^{\gamma}]^{1/\gamma}}, \tag{14.9}$$

as [103]

> "[t]his form [of the probability weighting function (14.9)] has several useful features: it has only one parameter; it encompasses weighting functions with both concave and convex regions; it does not require $w(.5) = .5$; and most important, it provides a reasonably good approximation to both the aggregate and the individual data for probabilities [p of the uncertain outcome x_U] in the range between 0.05 and 0.95."

But the now commonly used function is the one proposed in [30]:

$$w^{\pm}(p) = \frac{\delta \, p^{\gamma}}{\delta \, p^{\gamma} + (1-p)^{\gamma}}, \tag{14.10}$$

and the popularity of this function stems (for a large part) from [11] "its empirical tractability."

14.3   The Bayesian Decision Theory's Resolution

The basic premise of the Bayesian decision theory is that what is maximized in every decision concerning risky prospects is the position of the utility probability distributions corresponding with these prospects [105].

So, if it is found that the Allais counter-examples provide evidence against expected utility theory. Then, instead of proposing probability distortions and, thus, abandoning the assumption of rationality in economic theory, we may, alternatively, also entertain the possibility that the expectation value (i.e., mean) in some instances might be a sub-optimal position measure for a probability distribution.

And we may quote from Allais' 1998 Nobel Prize lecture (italics his own) [5]:

> "In the Theory of Games, von Neumann and Morgenstern presented both a method for determining cardinal utility and a rational rule of behaviour. [...] According to them, in order to be rational, any operator must maximize the mathematical expectation [...]. This stance struck me as being unacceptable because *it amounts to neglecting the probability distribution of psychological values around their mean*, which precisely represents the fundamental psychological element of the theory of risk. I illustrated my argumentation through counter-examples; one of them became famous as the 'Allais Paradox'."

So, the Bayesian decision theory proposes an alternative position measure, the so-called *locus*, for probability distributions which takes into account the position of the probabilistic worst- and best-case scenarios on the utility axis, as well as the position of the probabilistic most likely scenario, (13.25), or, equivalently, for $k_1 = k_2 = k$, (13.23):

$$\mathrm{loc}(D_i \,|\, k) = \frac{LB(k) + E(U \,|\, D_i) + UB(k)}{3}, \tag{14.11}$$

where $E(U \,|\, D_i)$ is the expected utility under decision D_i and $LB(k)$ and $UB(k)$ are the corresponding undershoot and overshoot corrected k-sigma lower and upper bounds, (13.13) and (13.16),

$$LB(k) = \begin{cases} a, & lb(k) < a, \\ lb(k), & lb(k) \geq a, \end{cases} \tag{14.12}$$

and

$$UB(k) = \begin{cases} ub(k), & ub(k) \leq b, \\ b, & ub(k) > b, \end{cases} \tag{14.13}$$

where a and b are the absolute worst- and best-case scenarios (i.e., the minimax and maximax criteria of choice), (13.10) and (13.14), respectively, and

where $lb(k)$ and $ub(k)$ are the (traditional) uncorrected k-sigma lower and upper bounds, (13.12) and (13.15),

$$lb(k) = E(U \mid D_i) - k \operatorname{std}(U \mid D_i), \tag{14.14}$$

and

$$ub(k) = E(U \mid D_i) + k \operatorname{std}(U \mid D_i), \tag{14.15}$$

where $\operatorname{std}(U \mid D_i)$ is the standard deviation of the utility probability distribution under decision D_i.

And it is through the undershoot and overshoot corrected k-sigma confidence bounds $LB(k)$ and $UB(k)$ that the spread $\operatorname{std}(U \mid D_i)$ around the expected utility $E(U \mid D_i)$, or, equivalently, "the probability distribution of psychological values around their mean, which precisely represents the fundamental psychological element of the theory of risk" [5], is brought to bear in the Bayesian criterion of choice (14.11).

It will now be demonstrated that the alternative criterion of choice (14.11) together with a sigma level of $k = 1$ imply the observed inverse-S shape of the ratios x_C/x_U as a function of the probability p.

Again, let x_C be a certainty outcome, either positive or negative, which has a probability of one and let x_U be an uncertainty outcome with the same sign as x_C which has a probability of p. Then we have for the certainty decision D_1 the outcome probability distribution[7]:

$$f(x \mid D_1) = \Big\{ 1, \quad x = x_C, \tag{14.16}$$

and a corresponding utility probability distribution:

$$f(u \mid D_1) = \Big\{ 1, \quad u = u(x_C \mid m), \tag{14.17}$$

where (14.3)

$$u(x_C \mid m) = q \, \log \frac{m + x_C}{m}. \tag{14.18}$$

The trivial expected value and standard deviation of the utility probability distribution (14.17) are given as

$$E(U \mid D_1) = u(x_C \mid m) \qquad \text{and} \qquad \operatorname{std}(U \mid D_1) = 0, \tag{14.19}$$

from which it follows that, (14.12) and (14.13),

$$LB(k \mid D_1) = E(U \mid D_1) = UB(k \mid D_1), \tag{14.20}$$

for any k. So, the locus of the utility probability distribution (14.17), for any sigma-level k, is given as, (14.11) and (14.20),

$$\operatorname{loc}(D_1 \mid \text{any } k) = u(x_C \mid m), \tag{14.21}$$

---

[7]We denote the probability distribution with the symbol f, rather than the symbol p. This is because we have given to the uncertain outcome x_U the probability p, thus, 'using-up' this symbol for the time being.

which is in correspondence with the common sense requirement that the position of a scalar should be that scalar itself[8].

For the uncertainty decision D_2 we have the following outcome probability distribution

$$f(x \mid D_2) = \begin{cases} p, & x = x_U, \\ 1 - p, & x = 0. \end{cases} \tag{14.22}$$

The corresponding utility distribution may be obtained by a simple relabeling, wherein the outcomes x in (14.22) are transformed into their corresponding utilities by way Bernoulli's utility function (14.3):

$$f(u \mid D_2) = \begin{cases} p, & u = u(x_U \mid m), \\ 1 - p, & u = u(0 \mid m) = 0, \end{cases} \tag{14.23}$$

where

$$u(x_U \mid m) = q \, \log \frac{m + x_U}{m}. \tag{14.24}$$

Since the probability distribution (14.23) is a Bernoulli distribution, having one zero outcome, we have that the expected value and standard deviation are given as

$$E(U \mid D_2) = u(x_U \mid m) \, p \tag{14.25}$$

and

$$\mathrm{std}(U \mid D_2) = |u(x_U \mid m)| \, \sqrt{p \, (1 - p)}. \tag{14.26}$$

For positive prospects $u(x_U \mid m) > 0$, the 1-sigma uncorrected lower and upper bounds for positive uncertain outcomes x_U are given as, (14.14), (14.15), (14.25) and (14.26),

$$lb^+(k = 1) = u(x_U \mid m) \, p - |u(x_U \mid m)| \, \sqrt{p \, (1 - p)}$$
$$= u(x_U \mid m) \left(p - \sqrt{p \, (1 - p)} \right) \tag{14.27}$$

and

$$ub^+(k = 1) = u(x_U \mid m) \, p + |u(x_U \mid m)| \, \sqrt{p \, (1 - p)}$$
$$= u(x_U \mid m) \left(p + \sqrt{p \, (1 - p)} \right), \tag{14.28}$$

where the '+' in the super-scripts denotes the fact that we are dealing with positive prospects. Since we have positive $u(x_U \mid m)$, the minimax and maximax values, (13.10) and (13.14), may be read from (14.23) as

$$a^+ = 0 \quad \text{and} \quad b^+ = u(x_U \mid m). \tag{14.29}$$

---

[8]See also (13.64) in the previous chapter.

For a sigma level of $k = 1$, lower bound undershoot occurs for $p < 1/2$, (14.27) and (14.29), as the lower bounds becomes negative and the minimax value $a^+ = 0$ is undershot; upper bound overshoot occurs for $p > 1/2$, (14.28) and (14.29), as the last term in (14.28) exceeds one and the maximax value $b^+ = u(x_U|m)$ is overshot; and neither undershoot nor overshoot occurs for $p = 1/2$, (14.27), (14.28) and (14.29):

$$
\begin{cases}
lb^+(k=1) < a^+ \text{ and } ub^+(k=1) < b^+, & p < \frac{1}{2}, \\
lb^+(k=1) = a^+ \text{ and } ub^+(k=1) = b^+, & p = \frac{1}{2}, \\
lb^+(k=1) > a^+ \text{ and } ub^+(k=1) > b^+, & p > \frac{1}{2}.
\end{cases}
\tag{14.30}
$$

So, the criterion of choice (14.11) translates for $u(x_U|m) > 0$ and a sigma level of $k = 1$ to, (14.12), (14.13), (14.14), (14.15), (14.25), (14.26), (14.29), and (14.30),

$$
\text{loc}^+(D_2|k=1) =
\begin{cases}
\frac{1}{3}\left[a^+ + 2E(U|D_2) + \text{std}(U|D_2)\right], & p < \frac{1}{2}, \\
E(U|D_2), & p = \frac{1}{2}, \\
\frac{1}{3}\left[2E(U|D_2) - \text{std}(U|D_2) + b^+\right], & p > \frac{1}{2},
\end{cases}
$$

$$
=
\begin{cases}
\frac{1}{3}\left[2\,u(x_U|m)\,p + |u(x_U|m)|\,\sqrt{p(1-p)}\right], & p < \frac{1}{2}, \\
u(x_U|m)\,p, & p = \frac{1}{2}, \\
\frac{1}{3}\left[2\,u(x_U|m)\,p - |u(x_U|m)|\,\sqrt{p(1-p)} + u(x_U|m)\right], & p > \frac{1}{2},
\end{cases}
$$

$$
=
\begin{cases}
u(x_U|m)\,\frac{1}{3}\left[2p + \sqrt{p(1-p)}\right], & p < \frac{1}{2}, \\
u(x_U|m)\,p, & p = \frac{1}{2}, \\
u(x_U|m)\,\frac{1}{3}\left[2p - \sqrt{p(1-p)} + 1\right], & p > \frac{1}{2},
\end{cases}
\tag{14.31}
$$

or, equivalently,

$$
\text{loc}^+(D_2|k=1) = u(x_U|m)\,g(p),
\tag{14.32}
$$

where

$$
g(p) =
\begin{cases}
\frac{2p + \sqrt{p(1-p)}}{3}, & p < \frac{1}{2}, \\
p, & p = \frac{1}{2}, \\
\frac{2p - \sqrt{p(1-p)} + 1}{3}, & p > \frac{1}{2}.
\end{cases}
\tag{14.33}
$$

For negative prospects $u(x_U|m) < 0$, the 1-sigma uncorrected lower and upper bound for negative uncertain outcomes x_U are given as, (14.14), (14.15), (14.25) and (14.26),

$$
lb^-(k=1) = u(x_U|m)\,p - |u(x_U|m)|\,\sqrt{p(1-p)}
$$

$$
\tag{14.34}
$$

$$
= u(x_U|m)\left(p + \sqrt{p(1-p)}\right)
$$

and

$$ub^-(k=1) = u(x_U|m)\ p + |u(x_U|m)|\ \sqrt{p\,(1-p)}$$

$$= u(x_U|m)\left(p - \sqrt{p\,(1-p)}\right),$$

(14.35)

where the '$-$' in the super-scripts denotes the fact that we are dealing with positive prospects. Since we have negative $u\,(x_U|m)$, the minimax and maximax values, (13.10) and (13.14), may be read from (14.23) as

$$a^- = u(x_U|m) \qquad \text{and} \qquad b^- = 0.$$

(14.36)

For a sigma level of $k = 1$, upper bound overshoot occurs for $p < 1/2$, (14.35) and (14.36), as the upper bound becomes positive and the maximax value $b^- = 0$ is overshot; lower bound undershoot occurs for $p > 1/2$, (14.34) and (14.36), as the last term in (14.34) exceeds one and the minimax value $a^- = u(x_U|m)$ is undershot; and neither undershoot nor overshoot occurs for $p = 1/2$, (14.34), (14.35) and (14.36):

$$\begin{cases} lb^-(k=1) > a^- \text{ and } ub^-(k=1) > b^-, & p < \frac{1}{2}, \\ lb^-(k=1) = a^- \text{ and } ub^-(k=1) = b^-, & p = \frac{1}{2}, \\ lb^-(k=1) < a^- \text{ and } ub^-(k=1) < b^-, & p > \frac{1}{2}. \end{cases}$$

(14.37)

So, the criterion of choice (14.11) translates for $u\,(x_U|m) < 0$ and a sigma level of $k = 1$ to, (14.12), (14.13), (14.14), (14.15), (14.25), (14.26), (14.36), and (14.37):

$$\text{loc}^-(D_2|k=1) = \begin{cases} \frac{1}{3}\left[2E(U|D_2) - \text{std}(U|D_2) + b^-\right] & p < \frac{1}{2}, \\ E(U|D_2), & p = \frac{1}{2}, \\ \frac{1}{3}\left[a^- + 2E(U|D_2) + \text{std}(U|D_2)\right], & p > \frac{1}{2}, \end{cases}$$

$$= \begin{cases} \frac{1}{3}\left[2\,u(x_U|m)\ p - |u(x_U|m)|\ \sqrt{p\,(1-p)}\right], & p < \frac{1}{2}, \\ u(x_U|m)\ p, & p = \frac{1}{2}, \\ \frac{1}{3}\left[u\,(x_U|m) + 2\,u(x_U|m)\ p + |u(x_U|m)|\ \sqrt{p\,(1-p)}\right], & p > \frac{1}{2}, \end{cases}$$

$$= \begin{cases} u(x_U|m)\ \frac{1}{3}\left[2p + \sqrt{p\,(1-p)}\right], & p < \frac{1}{2}, \\ u(x_U|m)\ p, & p = \frac{1}{2}, \\ u(x_U|m)\ \frac{1}{3}\left[1 + 2p - \sqrt{p\,(1-p)}\right], & p > \frac{1}{2}. \end{cases}$$

(14.38)

or, equivalently, (14.33),

$$\text{loc}^-(D_2|k=1) = u(x_U|m)\ g(p).$$

(14.39)

If we compare the loci of positive and negative prospects under the uncertainty decision D_2, (14.31) and (14.39), respectively, then it can be seen that both loci are the same:

$$\mathrm{loc}^+(D_2|\,k=1) = \mathrm{loc}^-(D_2|\,k=1)\,. \tag{14.40}$$

It follows that we may drop the '+' and '−' super-scripts in (14.32) and (14.39), as the general locus of the utility probability distribution (14.23), for both positive and negative prospects, is given as

$$\mathrm{loc}(D_2|\,k=1) = u(x_U|\,m)\ g(p)\,, \tag{14.41}$$

where, (14.33),

$$g(p) = \begin{cases} \frac{2p+\sqrt{p(1-p)}}{3}, & p < \frac{1}{2}, \\ p, & p = \frac{1}{2}, \\ \frac{2p-\sqrt{p(1-p)}+1}{3}, & p > \frac{1}{2}. \end{cases}$$

The certainty and uncertainty bets will be fair relative to each other whenever they are in a decision theoretical equilibrium, or, equivalently, have identical loci:

$$\mathrm{loc}(D_1|\,\mathrm{any}\ k) = \mathrm{loc}(D_2|\,k=1)\,, \tag{14.42}$$

or, equivalently, (14.3), (14.21) and (14.41),

$$q\log\frac{m+x_C}{m} = q\log\frac{m+x_U}{m}\ g(p)\,, \tag{14.43}$$

where we note that the Weber-constant q of monies falls away from our decision theoretical equality. If we solve (14.44) for the ratio x_C/x_U, we find the following prediction, (14.33):

$$h(p) = \frac{m}{x_U}\left[\left(\frac{m+x_U}{m}\right)^{g(p)} - 1\right] = \left(\frac{x_C}{x_U}\right)^{(\mathrm{elicited})}\,, \tag{14.44}$$

for both negative and positive prospects.

Now, the following meta-parameter estimates from the prospect theory literature are offered up in [11] for the probability weighting function (14.10) for positive prospects:

$$\left(\delta^+,\gamma^+\right) = (0.76, 0.69)\,. \tag{14.45}$$

Assuming a reference asset position of $m = 300$ euros, in terms of expendable monthly income, for our research subject (i.e., a graduate student), we may plot the probability weighting function, (14.10) and (14.45), and the Bayesian decision implication (14.44) for an uncertain outcome of $x_U = 200$ euros together, Figure 14.1.

For negative prospects, the following meta-parameter estimates from the prospect theory literature are offered up for the probability weighting function (14.10) in [11]:

$$\left(\delta^-,\gamma^-\right) = (1.09, 0.72)\,. \tag{14.46}$$
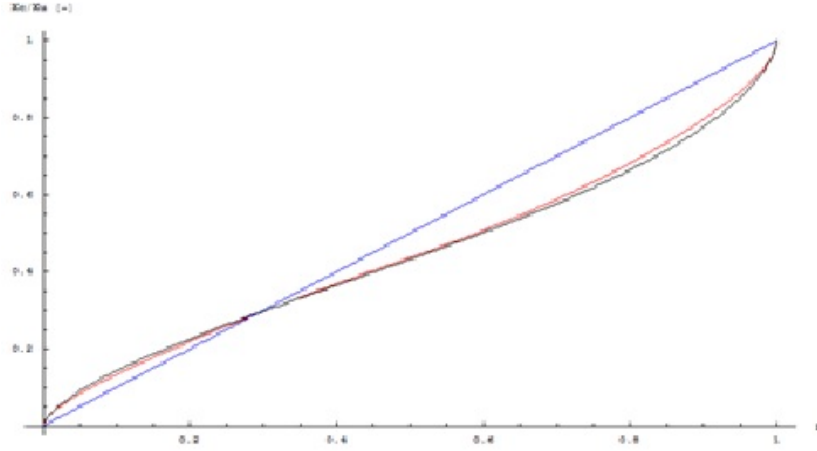
Figure 14.1: probability weighting functions: prospect theory (black); Bayesian decision theory (red); expected utility theory under linear utility (blue)

For the same reference asset position of $m = 300$ euros, we may plot the probability weighting function, (14.10) and (14.45), and the Bayesian decision implication (14.44) for an uncertain outcome of $x_U = -50$ euros together, Figure 14.2.

It can be seen in Figures 14.1 and 14.2 that the experimentally observed inverse S-shape in the elicited ratios x_C/x_U is accommodated by the Bayesian decision theory. Moreover, the equivalence of the functions (14.10) and (14.44), on a practical level, is quite surprising, as both functions have wildly differing mathematical expressions.

And we remind the reader that the prospect theoretical (14.10) is proposed because of its ability to fit the data, whereas (14.44) is a consequence of the alternative criterion of choice, (14.11) through (14.44), which the Bayesian decision theory proposes on the basis of a position maximization and fairness argument, as discussed in Section 13.3.

14.4 Discussion

Prospect theory needs to postulate its non-parsimonious hypothesis of the irrationality of the decision maker, by way of a psychological phenomenon of probability weighting, in order to be able to accommodate the observed inverse S-shape in preference ratios in certainty bets. The Bayesian decision theory is just as effective in its accommodation of the data, Figures 14.1 and 14.2, but much more parsimonious in its hypothesis. Since it only postulates that the expected value may be an insufficient position measure for general probability distributions, seeing that the expected value neglects the tail information of the probability distribution.

So, under prospect theory probability weighting is a general mechanism by
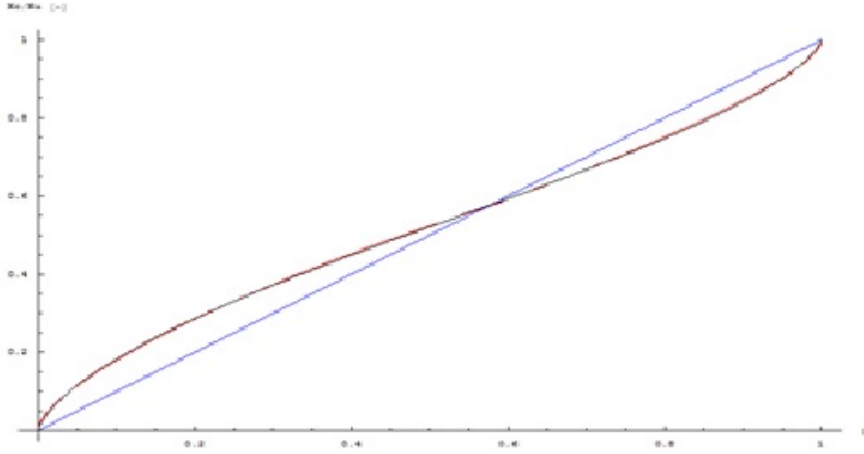
Figure 14.2: probability weighting functions: prospect theory (black); Bayesian decision theory (red); expected utility theory under linear utility (blue)

which the probability perception of small and large probabilities is distorted. But under the Bayesian decision theory probability weighting, or, equivalently, the observed inverse S-shape in preference ratios in certainty bets, is a consequence of the fact that in our risk assessments we take into account both the expected value and the tail information of the undistorted utility probability distributions, as opposed to only the expected value.

Prospect theory uses the observed preference ratios in certainty bets in order to come to the core of their theory, i.e., the use of probability weighting functions, as the probability weighting functions are proposed with these observations explicitly in mind. So, even though prospect theory is compatible with the observed violations of expected utility theory in the fair bet certainty equivalents, it does not imply them, and hence it does not explain them [80].

The Bayesian decision theory, in contrast, does imply the observed violations of expected utility theory, (14.11) through (14.44), and it explains these violations by observing that expected utility maximization ignores Allais' [5] "fundamental psychological element of the theory of risk" in the case of skewed utility probability distributions. Moreover, the Bayesian decision theory proposes the core of its theory, i.e., the use of an alternative criterion of choice, on the basis of a position maximization argument, as discussed in Section 13.3.1, which is unrelated to the observed preference ratios in certainty bets.

So, if both prospect theory and the Bayesian decision theory are able to accommodate the observed inverse S-shape in preference ratios in certainty bets, then prospect theory does so by construction, whereas the Bayesian decision theory does so by implication. It follows that only the latter Bayesian decision theory may claim this accommodation of the observed S-shape as an important and unexpected supporting contact for its hypothesis [85].

Chapter 15

The Allais Paradox

The Allais paradox shows a pattern of response to a risky choice which systematically violates the axioms underlying the von Neumann and Morgenstern theorem of expected utility maximization [75]. It will be demonstrated in this chapter how the Allais paradox can be accommodated by way the neo-Bernoullian Bayesian decision theory.

It is stated in [55] that the 'certainty effect' increases the desirability of certain large gains. We, however, will argue in this chapter that the certainty effect decreases the desirability of the uncertain large gains, rather than increasing the desirability of the certain large gains. This distinction may seem to be superfluous, as both interpretations lead to a relative increase/decrease in the attractiveness of the certain/uncertain gain.

But it is found that the Allais paradox points to an imbalanced risk aversion in which the probabilistic worst-case scenario is taken more forcefully into account, by way of a sigma level k_1 of the undershoot corrected lower confidence bound in Weaver's criterion of choice that is greater than the (normative) default value of one.

Stated differently, the Allais paradox is an example of risk aversion in the positive domain in which the observed betting preferences seem to be under the influence of a certainty effect wherein the undesirability of uncertain large gains increases as the the certain gain increases in size relative to the asset reference position (i.e., the initial wealth.) And it is proposed that the sigma level k_1 of Weaver's criterion of choice is the parameter that modulates the extent of this imbalanced risk aversion.

In this chapter we also give a rough outline on how to experimentally elicit the sigma level k_1 through the use of decision theoretical equilibrium (i.e., fair) values; i.e., where in Von Neumann and Morgenstern's expected utility one has to determine the unknown utility function u experimentally by way of the elicitation of fair values, there is proposed for the Bayesian decision theory an analogous program by which to determine the used confidence sigma level k_1 in Weaver's criterion of choice.

15.1   An Allais-Like Paradox

Allais offered up his paradox because he felt that the exclusive focus of the expected utility theory on the means (i.e., expected values) of the utility probability distributions neglects the spread of the probabilities of the utility values around their mean; a spread which according to Allais represents the fundamental psychological element of the theory of risk [5].

We will now discuss a slightly simplified version Allais' original paradox [3], where one has to choose between the following two options:

1. 10% chance of winning 100 million euros, and 90% chance of winning nothing,

2. 9% chance of winning 500 million euros, and 91% chance of winning nothing,

then most of us in correspondence with expected utility theory will prefer option 2, which has the greater expected utility. However, if one has to choose between the two options:

I. Absolute certainty of winning 100 euros,

II. 90% chance of winning 500 million euros, and 10% chance of winning nothing,

then most of us will prefer the secure option I, even though the uncertain option II has the greater expected utility, as we opt for security in the neighborhood of certainty.

15.2   The First Problem of Choice

In the first problem of choice we must choose between two bets that correspond with the following outcome probability distribution:

$$p(x|D_1) = \begin{cases} 0.10, & x = 100{,}000{,}000, \\ 0.90, & x = 0, \end{cases} \tag{15.1}$$

and

$$p(x|D_2) = \begin{cases} 0.09, & x = 500{,}000{,}000, \\ 0.91, & x = 0, \end{cases} \tag{15.2}$$

where the x are in euros.

15.2.1   The Choosing of Option 2

By way of Bernoulli's utility function (13.4),

$$u(x|m) = q \log \frac{m + x}{m}, \tag{15.3}$$

where q is the Weber-contant of money, m is the reference asset position of the decision maker, we may transform these outcome probability distributions to their corresponding utility probability distributions:

$$p(u|D_1) = \begin{cases} 0.10, & u = q\log\frac{m+100,000,000}{m}, \\ 0.90, & u = 0, \end{cases} \tag{15.4}$$

and

$$p(u|D_2) = \begin{cases} 0.09, & u = q\log\frac{m+500,000,000}{m}, \\ 0.91, & u = 0. \end{cases} \tag{15.5}$$

The 1-sigma loci (i.e., positions) of two-outcome utility probability distribution like (15.4) and (15.5) were in Chapter 14 derived as, (14.33), (14.41), and (15.3),

$$\text{loc}(D_i|k=1) = q\log\left(\frac{m+x_{Ui}}{m}\right)g(p_i), \tag{15.6}$$

where

$$g(p_i) = \begin{cases} \frac{2p_i+\sqrt{p_i(1-p_i)}}{3}, & p_i < \frac{1}{2}, \\ p_i, & p_i = \frac{1}{2}, \\ \frac{2p_i-\sqrt{p_i(1-p_i)}+1}{3}, & p_i > \frac{1}{2}, \end{cases} \tag{15.7}$$

and where x_{Ui} and p_i are the uncertain gains and their corresponding probability under decision D_i.

Since we have that the p_i of the uncertain outcomes in the utility distributions (15.3) and (15.4) both are smaller than $1/2$, we have, by way of a simplification of (15.7), that (15.6) simplifies to

$$\text{loc}(D_i|k=1) = q\log\left(\frac{m+x_{Ui}}{m}\right)\frac{2p_i+\sqrt{p_i(1-p_i)}}{3}. \tag{15.8}$$

If we assume a reference asset position of a yearly net-income of $m = 32,000$ per year, for reasons that will become apparent soon, then we find the respective loci, (15.3), (15.4), (15.5), (15.8), and a Weber constant of $q = 100$, (12.14),

$$\text{loc}(D_1|k=1) = 134.13 \text{ utile} \tag{15.9}$$

and

$$\text{loc}(D_2|k=1) = 150.06 \text{ utile}. \tag{15.10}$$

It follows that, in correspondence with the observed preferences, that the Bayesian decision theory picks the second option as the most profitable one with

$$\text{loc}(D_2|k=1) - \text{loc}(D_1|k=1) = 15.93 \text{ utile}, \tag{15.11}$$

where one utile corresponds with a just noticeable difference in utility.

So, the Bayesian decision theory solution aligns itself with the basic intuition to 'reach for the stars', that is, the choosing of option 2 in the first problem of choice, by taking into account the probabilistic worst- and best-case scenarios, as well as the most likely scenario.

15.2.2   A Deconstruction

From the first row of (14.30) it can be read that for $p < 1/2$ and a sigma level of $k = 1$ a lower bound undershoot will occur with no upper bound overshoot. So, in the presence of lower bound undershoot the probabilistic worst-case scenarios (13.13) of (15.4) and (15.5) evaluate to

$$LB(\,k = 1|\,D_i) = a_i, \tag{15.12}$$

where a_i is the minimax value (13.10). The most likely scenario is the expected value (14.6), which for this instance evaluates as, (14.25) and (15.3),

$$E(U|\,D_i) = q \log \left(\frac{m + x_{Ui}}{m} \right) \, p_i. \tag{15.13}$$

In the absence of upper bound overshoot the probabilistic best-case scenario (13.16) is the uncorrected k-sigma confidence bound (13.15), which for this instance evaluates as, (14.28) and (15.3),

$$UB(\,k = 1|\,D_i) = q \log \left(\frac{m + x_{Ui}}{m} \right) \left(p_i + \sqrt{p_i \,(1 - p_i)} \right). \tag{15.14}$$

Since the minimax value (13.10) is 0 for both options 1 and 2, (15.4) and (15.5), we have that both these options share the same probabilistic worst-case scenario, (15.12):

$$LB(\,k = 1|\,D_1) = 0 \text{ utile} \quad \& \quad LB(\,k = 1|\,D_2) = 0 \text{ utile}. \tag{15.15}$$

The most likely scenarios correspond with utilities of, (15.13),

$$E(U|\,D_1) = 80.48 \text{ utile} \quad \& \quad E(U|\,D_2) = 86.91 \text{ utile}, \tag{15.16}$$

whereas the probabilistic best-case scenarios (15.14) correspond with utilities of

$$UB(\,k = 1|\,D_1) = 321.90 \text{ utile} \quad \& \quad UB(\,k = 1|\,D_2) = 363.27 \text{ utile}. \tag{15.17}$$

We summarize, the second bet has both an expected utility that is 6.43 utiles more desireable than the first bet, and a probabilistic best-case scenario that is more desireable with 41.37 utiles, whereas the worst-case scenario for both bets is expected to be the same in a probabilistic sense. So the choice for the second bet in the first problem of choice is a win-win choice, all other things being equal.

Note that the sum of (15.15), (15.16), and (15.17) divided by three will give the loci (15.9) and (15.10).

15.3   The Second Problem of Choice

In the second problem of choice we must choose between two bets that correspond with the following outcome probability distributions:

$$p(\,x|\,D_{\mathrm{I}}) = \begin{cases} 1.0, \quad\quad x = 100{,}000{,}000, \end{cases} \tag{15.18}$$

and

$$p(x|D_{\mathrm{II}}) = \begin{cases} 0.90, & x = 500{,}000{,}000, \\ 0.10, & x = 0, \end{cases} \tag{15.19}$$

where the x are in euros.

15.3.1   The Choosing of Option I, Part I

By way of Bernoulli's utility function (13.4), we may transform these outcome probability distributions to their corresponding utility probability distributions:

$$p(u|D_{\mathrm{I}}) = \left\{ 1.0, \qquad u = q \log \tfrac{m+100{,}000{,}000}{m}, \right. \tag{15.20}$$

and

$$p(u|D_{\mathrm{II}}) = \begin{cases} 0.90, & u = q \log \tfrac{m+500{,}000{,}000}{m}, \\ 0.10, & u = 0. \end{cases} \tag{15.21}$$

The locus (i.e., position) of certainty utility probability distributions like (15.20) was previously derived as, (14.21) and (15.3),

$$\mathrm{loc}(D_{\mathrm{I}}|\,\text{any } k) = q \log \frac{m + x_C}{m}. \tag{15.22}$$

The balanced 1-sigma locus of utility probability distributions like (15.21) for an uncertain outcome probability of $p > 1/2$ is given as, (15.6) and (15.7),

$$\mathrm{loc}(D_{\mathrm{II}}|\,k=1) = q \log \left(\frac{m + x_U}{m} \right) \frac{2p - \sqrt{p(1-p)} + 1}{3}, \tag{15.23}$$

where x_U and p are the uncertain outcome and its the probability, respectively.

So, if we again assume a reference asset position of $m = 32{,}000$ per year, then we find for $x_C = 100{,}000{,}000$, $x_U = 500{,}000{,}000$, $p = 0.90$, (15.20) and (15.21), and a Weber constant of $q = 100$, (12.14), the respective loci, (15.22) and (14.34),

$$\mathrm{loc}(D_{\mathrm{I}}|\,\text{any } k) = 804.75 \text{ utile} \tag{15.24}$$

and

$$\mathrm{loc}(D_{\mathrm{II}}|\,k=1) = 804.72 \text{ utile}. \tag{15.25}$$

It follows that for this reference asset position of 32,000 neither option dominates, as the difference in utiles for both options does not cross the just noticeable difference threshold of one utile:

$$\mathrm{loc}(D_{\mathrm{I}}|\,k=1) - \mathrm{loc}(D_{\mathrm{II}}|\,k=1) = 0.03 \text{ utile}. \tag{15.26}$$

So, in light of the overwhelming (introspective) preference for the certainty decision D_I [3, 4], we are now forced to invoke the postulate of decision theoretical imbalance, in that there seems to be a tendency towards a more pronounced risk-aversion than we would normally expect in problems of choice involving high-stake (near) certainty gains.

15.3.2   Imbalanced Risk Aversion in the Positive Domain

In problems of choice involving large (near) certain gains there seems to be a tendency towards a more pronounced risk-avoiding than we would normally expect. If we examine our intuitions regarding risk aversion in the positive domain, then it would seem that this behavior is on the side of caution, as are compelled to 'take the money and run', in compliance with the adage that 'one bird in the hand is better than two in the bush.'

Now, as we make a small excursion into the utility of time, there are situations in which it is imperative that we not be late. For example, when going for a job interview, or when we have to attend an important project meeting, or when catching a flight. In such situations there will be an extreme high utility attached to the travel time t being smaller or equal to some value t_0.

So, when planning our trip to the job interview site, project meeting location, or airport, we will generously allow for excess travel time due to unforeseen contingencies, as it is 'better to be safe than sorry', and the more sorry we stand to be, i.e., the more imperative the need to be on time, the more safety measures we will be willing to take, i.e., the more excess travel time we will be willing to add to our travel plans.

What is the mechanism by which we determine the excess travel time needed to remain on the side of caution? Introspection would suggest that it is by entertaining worst-case travel scenarios that are more severe than the standard worst-case travel scenarios we normally would entertain. For example, when planning for regular train commutes we typically only allow for regular delays, whereas when planning for a train trip to the airport we typically will also allow for out of the ordinary train delays; i.e., to be cautious is to forcefully take into account all the things that can go wrong.

Now, as we again return to the utility of monetary gains, that which can go wrong as we choose for the uncertain gain is that we may not obtain this gain. So, if in [55] it is stated that the 'certainty effect' increases the desirability of certain large gains, then it is argued here that, rather than increasing the desirability of certain large gains, the certainty effect decreases the desirability of the uncertain large gains, as it puts to the fore the worst-case scenario of not obtaining that gain.

Also, it is expected that the larger the certain gain, the more undesirable the not obtaining of that gain will become. Stated differently, the larger the certain gain relative to our asset reference position, the more extreme will be the worst-case scenarios that impose themselves upon our consciousness.

In the case of no lower bound undershoot, the probabilistic worst-case scenario is given as a function of probabilities, outcomes, and the used k-sigma level, (13.9), (13.10), (13.11), (13.12) and (13.13),

$$lb(k_1) = E(U) + k_1 \operatorname{std}(U)\,. \tag{15.27}$$

For k-sigma levels greater than

$$k_1 \geq \frac{E(U) - a}{\operatorname{std}(U)}, \tag{15.28}$$

the minimax criterion a becomes the accounted for worst-case scenario, as opposed to the less pessimistic probabilistic worst-case scenario $lb(k_1)$.

It follows that the sigma level k_1 is the parameter that modulates the forcefulness in which the worst-case scenario is taken into account, as we are under the sway of the certainty effect [55], or some variation thereof, with a limit of pessimism that tends to the minimax criterion (13.10):

$$LB(k_1) \to a, \qquad \text{as } k_1 \to \frac{E(U) - a}{\text{std}(U)}. \tag{15.29}$$

So, the postulate of imbalanced risk-seeking translates to the general criterion of choice (13.25),

$$\text{loc}(k_1, k_2) = \frac{LB(k_1) + E(U) + UB(k_2)}{3}, \tag{15.30}$$

where the forcefulness of the worst-case scenario and, as a consequence, the unattractiveness of that choice, is modulated upward as the lower bound sigma level k_1 is set to some value greater than one, while the weighting of the best-case scenario remains the same, as k_2 is set to its default value of one.

It may be derived, as is done in Section 15.3.6, that the locus (15.30) for the utility probability distribution (16.14) simplifies to[1]

$$\text{loc}^+(D_1|, k_1, k_2) = q \log\left(\frac{m + x_U}{m}\right) g^+(p, k_1, k_2), \tag{15.31}$$

where

$$g^+(p, k_1, k_2) = \begin{cases} \frac{2p + k_2\sqrt{p(1-p)}}{3}, & p \leq \frac{1}{1+k_2^2}, \\ \frac{1+p}{3}, & \frac{1}{1+k_2^2} < p < \frac{k_1^2}{1+k_1^2}, \\ \frac{2p - k_1\sqrt{p(1-p)}+1}{3}, & p \geq \frac{k_1^2}{1+k_1^2}, \end{cases} \tag{15.32}$$

for general k_1 and k_2.

15.3.3   The Choosing of Option I, Part II

It follows from (15.31) that the general locus of uncertainty bets in the positive domain for $p = 0.9$ and $1 \leq k_1 \leq 3$ is given as,

$$\text{loc}^+(D_1| k_1, k_2) = q \log\left(\frac{m + x_U}{m}\right) \frac{2p - k_1\sqrt{p(1-p)}+1}{3}. \tag{15.33}$$

If we set the cautious probabilistic worst-case scenario lower bounds to $k_1 = 1, 2, 3$, with a default $k_2 = 1$ probabilistic best-case scenario in the locus (15.33),

---

[1]In order to keep the flow of exposition going, we have moved the derivation of the imbalanced locus under risk aversion to the end of this section.

then we obtain for $x_U = 500{,}000{,}000$, $p = 0.9$, $m = 32{,}000$, and a Weber constant of $q = 100$, (12.14), the following imbalanced loci:

$$\begin{cases} \mathrm{loc}^+(D_{II}|\,k_1 = 1, k_2 = 1) = 804.72 \text{ utile,} \\ \mathrm{loc}^+(D_{II}|\,k_1 = 2, k_2 = 1) = 708.16 \text{ utile,} \\ \mathrm{loc}^+(D_{II}|\,k_1 = 3, k_2 = 1) = 611.59 \text{ utile.} \end{cases} \qquad (15.34)$$

And it can be seen that the strength of the certainty effect, i.e., the extent of the decrease in the desirability of the uncertain large gain, is modulated by the lower bound sigma level k_1.

So, as we ever more forcefully take into account the the probabilistic worst-case scenario option II, the more forcefully will we be swayed go with option I, (15.24) and (15.34):

$$\begin{cases} \mathrm{loc}(D_I|\,\text{any } k) - \mathrm{loc}^+(D_{II}|\,k_1 = 1, k_2 = 1) = 0.03 \text{ utile,} \\ \mathrm{loc}(D_I|\,\text{any } k) - \mathrm{loc}^+(D_{II}|\,k_1 = 2, k_2 = 1) = 96.59 \text{ utile,} \\ \mathrm{loc}(D_I|\,\text{any } k) - \mathrm{loc}^+(D_{II}|\,k_1 = 3, k_2 = 1) = 193.16 \text{ utile.} \end{cases} \qquad (15.35)$$

15.3.4   A Deconstruction

From the fact that the lower bound, expected value, and upper bound of an certainty outcome are that self-same certainty outcome, (14.19) and (14.20), (15.3), as well as, (15.55) together with $k_1 \leq 3$ and $p = 0.90$,

$$LB^+(k_1) = q \log\left(\frac{m + x_U}{m}\right)\left(p - k_1\,\sqrt{p\,(1-p)}\right)$$

and, (16.30),

$$E(U|\,D_{II}) = q \log\left(\frac{m + x_U}{m}\right) p,$$

and, (15.56) together with $k_2 = 1$ and $p = 0.90$,

$$UB^+(k_2) = q \log\left(\frac{m + x_U}{m}\right),$$

we have for a Weber constant of $q = 100$, (12.14),

$$LB(\text{any } k|\,D_I) = 804.75 \text{ utile} \quad \& \quad \begin{cases} LB(k_1 = 1|\,D_{II}) = 579.40 \text{ utile,} \\ LB(k_1 = 2|\,D_{II}) = 289.70 \text{ utile,} \\ LB(k_1 = 3|\,D_{II}) = 0 \text{ utile,} \end{cases}$$
$$(15.36)$$

and

$$E(U|\,D_I) = 804.75 \text{ utile} \quad \& \quad E(U|\,D_{II}) = 869.10 \text{ utile,} \qquad (15.37)$$

and

$$UB(\text{any } k|\,D_I) = 804.75 \text{ utile} \quad \& \quad UB(k = 1|\,D_{II}) = 965.67 \text{ utile.} \qquad (15.38)$$

We summarize, the first bet has a probabilistic worst-case scenario that is either 225.35, or 515.05 or 804.75 utile more desirable than the a probabilistic worst-case scenario of the second bet, depending on the risk aversion sigma level k_1, and an expected utility and a probabilistic best-case scenario that are 64.35 and 161.08 utile less desirable, respectively.

So, the observed and intuitive preference for the first bet in the second problem of choice may be explained by way of a probabilistic worst-case scenario modulation in the second uncertainty bet that is triggered by the presence of a large certain gain in the first bet.

15.3.5 Some Equilibrium Values

The certainty and uncertainty bets are in a decision theoretical equilibrium whenever their utility loci (i.e., positions of their utility probability distributions) are equal

$$\text{loc}(D_{\text{I}}|\text{ any } k) = \text{loc}^+(D_1|, k_1, k_2), \tag{15.39}$$

or, equivalently, (15.22) and (15.33),

$$q \log \frac{m + x_C}{m} = q \log \left(\frac{m + x_U}{m} \right) \frac{2p - k_1 \sqrt{p\,(1-p)} + 1}{3}. \tag{15.40}$$

If we solve (15.40) for the certain and uncertain outcomes x_C and x_U, we obtain the functions

$$x_C^{(\text{fair})} = m \left[\left(\frac{m + x_U}{m} \right)^{\frac{2p - k_1 \sqrt{p(1-p)}+1}{3}} - 1 \right] \tag{15.41}$$

and

$$x_U^{(\text{fair})} = m \left[\left(\frac{m + x_C}{m} \right)^{\frac{3}{2p - k_1 \sqrt{p(1-p)}+1}} - 1 \right]. \tag{15.42}$$

where m is the reference asset position.

If we set $p = 0.9$ and $x_U = 500{,}000{,}000$, (15.18), assume a reference asset position of $m = 32{,}000$ and modulate the probabilistic worst-case scenario with $k_1 = 1, 2, 3$, with an unmodulated probabilistic best-case scenario $k_2 = 1$, then we find corresponding fair certainty outcomes of (15.41), respectively,

$$x_C^{(\text{fair})} = \begin{cases} 99{,}973{,}333, & \text{for } k_1 = 1, \ k_2 = 1, \\ 38{,}042{,}865, & \text{for } k_1 = 2, \ k_2 = 1, \\ 14{,}464{,}181, & \text{for } k_1 = 3, \ k_2 = 1. \end{cases} \tag{15.43}$$

So, it is predicted by the Bayesian decision theory that an extreme risk averse personality will start to consider to forgo of the uncertainty bet D_{II} and its corresponding opportunity to become a halfway-billionaire with 90% in exchange for the certainty of becoming a multimillionaire with 14.5 million euro to his name.

And if we set $p = 0.9$ and $x_C = 100{,}000{,}000$, (15.19), assume a reference asset position of $m = 32{,}000$ and modulate the probabilistic worst-case scenario with $k_1 = 1, 2, 3$, with an unmodulated probabilistic best-case scenario $k_2 = 1$, then we find corresponding fair uncertainty outcomes of (15.41), respectively,

$$x_U^{(\text{fair})} = \begin{cases} 500{,}160{,}000, & \text{for } k_1 = 1, \ k_2 = 1, \\ 1{,}886{,}540{,}000, & \text{for } k_1 = 2, \ k_2 = 1, \\ 10{,}055{,}730{,}000, & \text{for } k_1 = 3, \ k_2 = 1. \end{cases} \qquad (15.44)$$

So, it is predicted by the Bayesian decision theory that an extreme risk averse personality will start to consider to forgo of the certainty bet D_I and the promise of becoming a multimillionaire with 100 million euro to his name in exchange for the very real chance to become a multi-billionaire with 10 billion euro.

Also, if we numerically solve (15.40) for p, while setting $x_U = 500{,}000{,}000$ and $x_C = 100{,}000{,}000$, (15.18) and (15.19), then we find for the reference asset position $m = 32{,}000$ a corresponding fair probability of

$$p^{(\text{fair})} = \begin{cases} 0.9000, & \text{for } k_1 = 1, \ k_2 = 1, \\ 0.9557, & \text{for } k_1 = 2, \ k_2 = 1, \\ 0.9766, & \text{for } k_1 = 3, \ k_2 = 1. \end{cases} \qquad (15.45)$$

So, it is predicted that that same personality will only start consider the uncertainty bet and the corresponding possibility of becoming a halfway-billionaire if the odds of becoming one are more than 97 to 3.

Equilibrium predictions like (15.43), (15.44), and (15.45) may be used to experimentally determine the implied sigma value. For example, for given probability p and uncertain outcome x_U, the fair certain outcome $x_C^{(\text{fair})}$ may be elicited in hypothetical betting experiments and substituted in (15.40). If we then solve (15.40) for k_1, we find the implied lower confidence bound sigma level:

$$k_1^{(\text{implied})} = \frac{1}{\sqrt{p\,(1-p)}} \left(1 + 2p - 3\, \frac{\log\left(m + x_C^{(\text{fair})}\right) - \log m}{\log\left(m + x_U\right) - \log m} \right). \qquad (15.46)$$

The same procedure may also be used to determine the used confidence sigma level k_2 for risk seeking in the negative domain, as will be discussed in the next chapter. So, where in Von Neumann and Morgenstern's expected utility one has to determine the unknown utility function u in its totality by way of the elicitation of fair values, we have here an analogous program by which to determine the confidence sigma levels k_1 and k_2 of the position measure (13.25).

15.3.6   Some Derivations

The uncertainty decision D_1 has an outcome probability distribution (16.48) of the form

$$f(x \mid D_1) = \begin{cases} p, & x = x_U, \\ 1 - p, & x = 0. \end{cases} \qquad (15.47)$$

The corresponding utility distribution is be obtained by a simple relabeling, wherein the outcomes x in (16.48) are transformed into their corresponding utilities by way Bernoulli's utility function (13.4):

$$f(u \mid D_1) = \begin{cases} p, & u = u(x_U \mid m), \\ 1 - p, & u = 0, \end{cases} \qquad (15.48)$$

where

$$u(x_U \mid m) = q \ \log \frac{m + x_U}{m}. \qquad (15.49)$$

Since the probability distribution (15.48) is a Bernoulli distribution with one zero outcome, we have that the expected value and standard deviation are given as (16.30) and (16.31). So, for $u(x_U \mid m) > 0$, the k_1-sigma uncorrected lower and the k_2-sigma uncorrected upper bound for negative uncertain outcomes x_U are given as, (13.11), (13.15), (15.49), (16.30), and (16.31),

$$lb^+(k_1) = u(x_U \mid m) \ p - k_1 \ |u(x_U \mid m)| \ \sqrt{p(1-p)}$$

$$\qquad (15.50)$$

$$= u(x_U \mid m) \left(p - k_1 \ \sqrt{p(1-p)} \right)$$

and

$$ub^+(k_2) = u(x_U \mid m) \ p + k_2 \ |u(x_U \mid m)| \ \sqrt{p(1-p)}$$

$$\qquad (15.51)$$

$$= u(x_U \mid m) \left(p + k_2 \ \sqrt{p(1-p)} \right).$$

Since we have positive $u(x_U \mid m)$, the minimax and maximax values, (13.10) and (13.14), may be read from (15.48) as

$$a^+ = 0 \qquad \text{and} \qquad b^+ = u(x_U \mid m). \qquad (15.52)$$

It follows that lower bound undershoot and upper bound overshoot will occur whenever, (15.50), (15.52), and (16.36),

$$\begin{cases} lb^+(k_1) < a^+ & p < \frac{k_1^2}{1+k_1^2}, \\ lb^+(k_1) \geq a^+ & p \geq \frac{k_1^2}{1+k_1^2}, \end{cases} \qquad (15.53)$$

and, (15.51), (15.52), and (16.35),

$$\begin{cases} ub^+(k_2) \leq b^+, & p \leq \frac{1}{1+k_2^2}, \\ ub^+(k_2) > b^+, & p > \frac{1}{1+k_2^2}. \end{cases} \qquad (15.54)$$

So the corrected lower and upper sigma bounds are given as, (13.13), (15.50), (15.52), and (15.53),

$$LB^+(k_1) = \begin{cases} 0, & p < \frac{k_1^2}{1+k_1^2}, \\ u(x_U \mid m) \left(p - k_1 \ \sqrt{p(1-p)} \right), & p \geq \frac{k_1^2}{1+k_1^2}, \end{cases} \qquad (15.55)$$

and, (13.16), (15.51), (15.52), and (15.54),

$$UB^+(k_2) = \begin{cases} u(x_U \,|\, m) \left(p + k_2 \sqrt{p\,(1-p)} \right), & p \leq \frac{1}{1+k_2^2}, \\ u(x_U \,|\, m), & p > \frac{1}{1+k_2^2}. \end{cases} \qquad (15.56)$$

By way of (15.30), (16.30), (15.55), and (15.56), we may set up the following identity for the general locus of uncertainty bets in the positive domain:

$$\mathrm{loc}^+(D_1 \,|\, k_1, k_2) = \begin{cases} u(x_U \,|\, m) \frac{1}{3} \left[3p + (k_2 - k_1) \sqrt{p\,(1-p)} \right], & \frac{k_1^2}{1+k_1^2} \leq p \leq \frac{1}{1+k_2^2}, \\ u(x_U \,|\, m) \frac{1}{3} \left(2p - k_1 \sqrt{p\,(1-p)} + 1 \right), & p \geq \max \left(\frac{k_1^2}{1+k_1^2}, \frac{1}{1+k_2^2} \right), \\ u(x_U \,|\, m) \frac{1}{3} \left(2p + k_2 \sqrt{p\,(1-p)} \right), & p \leq \min \left(\frac{k_1^2}{1+k_1^2}, \frac{1}{1+k_2^2} \right), \\ u(x_U \,|\, m) \frac{1}{3} (1+p), & \frac{1}{1+k_2^2} < p < \frac{k_1^2}{1+k_1^2}, \end{cases}$$

$$(15.57)$$

where

$$\frac{k_1^2}{1+k_1^2} \leq p \leq \frac{1}{1+k_2^2} \qquad (15.58)$$

will only hold for $k_1, k_2 \leq 1$, and where, because in imbalanced risk aversion we have, by construction, that $k_1 > k_2$,

$$\min \left(\frac{k_1^2}{1+k_1^2}, \frac{1}{1+k_2^2} \right) = \frac{1}{1+k_2^2}, \qquad (15.59)$$

and

$$\max \left(\frac{k_1^2}{1+k_1^2}, \frac{1}{1+k_2^2} \right) = \frac{k_1^2}{1+k_1^2}. \qquad (15.60)$$

It follows that we may simplify (15.57) as

$$\mathrm{loc}^+(D_1 \,|\, , k_1, k_2) = u(x_U \,|\, m)\, g^+(p, k_1, k_2) \qquad (15.61)$$

or, equivalently, (15.49),

$$\mathrm{loc}^+(D_1 \,|\, , k_1, k_2) = q \log \frac{m+x_U}{m}\, g^+(p, k_1, k_2), \qquad (15.62)$$

where

$$g^+(p, k_1, k_2) = \begin{cases} \frac{2p + k_2\sqrt{p(1-p)}}{3}, & p \leq \frac{1}{1+k_2^2}, \\ \frac{1+p}{3}, & \frac{1}{1+k_2^2} < p < \frac{k_1^2}{1+k_1^2}, \\ \frac{2p - k_1\sqrt{p(1-p)}+1}{3}, & p \geq \frac{k_1^2}{1+k_1^2}, \end{cases} \qquad (15.63)$$

for general k_1 and k_2.

15.4 Allais' Variance Preferences

In the Bayesian framework both the expected values and standard deviations, or, equivalently, variances, of the utility probability distributions are taken into account in the making of decisions, while the skewness of the utility probability distributions is indirectly taken into account by way of the concept of under- and overshoot corrections in (13.13) and (13.16), which introduce skewness-reflecting asymmetries in the consequent fair position measure of the utility probability distribution (13.23).

It turns out that an approach along these lines was also envisaged by both Allais [2, 3, 4] and Georgescu-Roegen[2] [28]. Moreover, Allais constructed his famous paradox in order to demonstrate the utility spread of the utility probability distribution around the expected utility should be the fundamental psychological element of the theory of risk [3, 5]. Stated differently, people not only try to maximize the expected value of utility, they also take into account the variances of the respective utility probability distributions.

The main reason that Allais' concept of 'variance preferences' never caught on is probably because Edwards, who was the mentor of Kahneman and Tversky, deemed the problem of utility measurement for monetary stimuli to be insoluble, even though he endorsed the suggestion of variance preferences in principle [22]:

> There are instances in which this argument seems convincing. You would probably prefer the certainty of a million dollars to a 50-50 chance of getting either four million or nothing. I do not think that this preference is due to the fact that the expected utility of the 50-50 bet is less than the utility of one million dollars to you, although this is possible. A more likely explanation is simply that the variances of the two propositions are different.

Edwards, nonetheless, rejected the suggestion on the practical grounds [22]:

> [T]he introduction of the variance and higher moments makes the problem of applying the theory [as envisaged by Allais] experimentally seem totally insoluble. It is difficult enough [, in von Neumann and Morgenstern's expected utility theory, as opposed to Bernoulli's expected utility theory, that is,] to derive reasonable methods of measuring utility alone from risky choices; when it also becomes necessary [...] to take the higher moments of the utility distribution into account, the problem seems hopeless.

Allais' suggestion to solve this problem of intractability by using psycho-physical methods was dismissed by Edwards' on the grounds that it would be impossible to assign subjective values to monetary increments [22]:

> The dollar scale of the value of money is so thoroughly taught to us that it seems almost impossible to devise a psychophysical situation

---

[2]We were unable to find Georgescu-Roegen's article, referenced in [22], on-line.

in which subjects would judge the utility, rather than the dollar value, of dollars.

We, however, are of the opinion that Edwards may have been in the wrong in regards to the absoluteness of the dollar scale. The issue is not whether or not we can discriminate between sums of money. Rather, the question is whether or not a given sum of money may induce a sense of loss, if taken from us, or a sense of gain, if given to us.

In other words, we trust that both the rich and the poor will be able to discriminate between sums of money up to the cent. However, we do not expect them to assign the same utility to the same sum of dollars; e.g., the archetypical image of the rich fat-cat who lights his cigar with a 100 dollar bill.

If we use the psycho-physical Weber-Fechner law to translate the dollar scale into an utility scale, then the only technical issue remaining is which value to assign to the Weber constant of monetary stimuli. But in practice we will see that all references to the unknown Weber constant q fall away once we have set up our decision theoretic (in)equalities. So, the actual value of the Weber constant will not be an issue in a large class of decision theoretical problems. A fact which also is mentioned by Bernoulli himself [6].

And for those instances where we do have need for an explicit value of the Weber constant, for example, if we want to graph a utility function in a certain income domain for some initial reference asset position (e.g., Figures 12.1 and 12.2) or if we want to discuss our results in utiles (i.e., decision theoretical 'decibels'), then a Weber constant may readily be assigned by way of introspection, as was done earlier in (12.14).

Had Edwards not dismissed Allais' suggestion of variance preferences, which he endorsed in principle, on the practical grounds that the Weber-Fechner law would be inapplicable to monetary stimuli, then the decision theoretical landscape might have been radically different from the way it is today. Since it were Edwards' post-docs that brought us prospect theory and the paradigm of behavioral economics[3].

---

[3]Note that Edwards himself had fundamental issues with behavioral economics' implicit characterization of 'man as a cognitive cripple'. It can be read in [82] that even though Edwards struggled for many years to make his peace with the behavioral economic research of his former pupils, he never succeeded in doing so.

Chapter 16

The Reflection Effect and the Fourfold Pattern

If we offer a choice between a certain and an uncertain option, the outcomes of these options having the same sign, and the uncertain option consisting of a zero outcome and a non-zero outcome, then the reflection effect is an observed pattern in which the preference for either the certain or uncertain option tends to reverse as the sign of the outcomes is reversed [53].

To be more specific, let risk seeking be the choosing of the uncertain option and risk aversion be the choosing for the certain option. Then for an uncertain option with an absolute non-zero outcome that both is much larger than the outcome of the certain option and has a small probability of occurring, the tendency for risk seeking (in the positive domain) will go to a tendency for risk aversion (in the negative domain), as we change the sign of the outcomes from positive to negative. The buying of lottery tickets is an example of risk seeking in the positive domain and the buying of an insurance an example of risk aversion in the negative domain.

Also, for an uncertain option with an absolute non-zero outcome that is both somewhat larger than the outcome of the certain option and has a large probability of occurring, the tendency for risk aversion (in the positive domain) will go to a tendency for risk seeking (in the negative domain), as we change the sign of the outcomes from positive to negative. The opting for certainty in the Allais paradox is an example of risk aversion in the positive domain and the trying to make good on previous losses by taking a large chance on an even greater loss an example of risk seeking in the negative domain.

It follows that the reflection effect gives rise to a fourfold pattern of observed preferences in our choosing between certain and uncertain options. Prospect theory explains this characteristic reflection pattern of attitudes to risky prospects by way of the form of its probability weighting functions[1].

Overweighting of small probabilities contributes to the popularity of both

---

[1]See Figures 14.1 and 14.2 in Chapter 14.

lotteries (i.e., risk seeking in the positive domain) and insurance (i.e., risk aversion in the negative domain). Underweighting of high probabilities contributes both to the prevalence of risk aversion in choices between a probable somewhat greater gain and a somewhat smaller but certain gain (e.g., the Allais paradox) and to the prevalence of risk seeking in choices between a highly probable and large loss and a somewhat smaller but certain loss.

This explanation of the fourfold pattern of preferences is said to be one of the core achievements of prospect theory [55]. It will be demonstrated in this chapter that the Bayesian decision theory, which can predict the probability weighting functions from first principles, as discussed in Chapter 14, can also accommodate the reflection effect and the consequent fourfold pattern.

Also, it is found that the observed tendency risk seeking in negative domain points to an imbalanced risk seeking in which the probabilistic best-case scenario is taken more forcefully into account, by way of a sigma level k_2 of the overshoot corrected upper confidence bound in Weaver's criterion of choice that is greater than the (normative) default value of one.

Stated differently, if the 'possibility effect' increases the desirability of desperate gambles in which we cling to a 'sliver of hope' as we accept a high probability of making things worse in exchange for a small chance of avoiding a large loss [55], then risk seeking in negative domain seems to be under the sway of this effect, as the uncertain option, with its small probability of no-loss, becomes ever more attractive as the certain loss increases in size relative to the asset reference position (i.e., the initial wealth.) And it is proposed that the sigma level k_2 of Weaver's criterion of choice is the parameter that modulates the extent of this imbalanced risk seeking.

16.1   Risk Seeking in the Positive Domain

We first give an example of risk seeking in the case of a small probability of winning a large prize. The outcome probability distributions for the respective bets in our risk seeking example are from [53] and are given as

$$p(x|D_1) = \begin{cases} 0.001, & x = 5000, \\ 0.999, & x = 0, \end{cases} \tag{16.1}$$

and

$$p(x|D_2) = \begin{cases} 1.0, & x = 5, \end{cases} \tag{16.2}$$

where the x are in Israeli pounds. It is found that 72% of $N = 72$ subjects prefer decision D_1 over D_2 [53], even though both bets have the same expectation value; i.e.,

$$E(x|D_1) = 5 = E(x|D_2).$$

The preference for an uncertain but much larger gain over a sure but much smaller gain constitutes risk seeking in the positive domain.

Risk seeking in the positive domain represents our tendency to maximize profits, as it moves us to invest in a long shot if the pay-out is high enough. We now interpret this finding in terms of the Bayesian decision theoretical framework.

16.1.1   The Predicted Choice

Let x_C and x_U, respectively, be the certainty and the uncertainty outcomes in Israeli pounds, where the uncertainty outcome x_U has a probability of p of being realized. In the case that x_U is not realized the outcome will be zero, and the probability corresponding with this outcome is $1 - p$. The certainty outcome x_C is certain and, hence, has a probability of one.

It was derived in the previous chapter that the uncertainty and certainty bets have, (16.1) and (16.2), have 1-sigma utility loci of, (14.3) and (14.41),

$$\text{loc}(D_1 | k = 1) = q \log \frac{m + x_U}{m} \, g(p) \,, \tag{16.3}$$

where

$$g(p) = \begin{cases} \frac{2p + \sqrt{p(1-p)}}{3}, & p < \frac{1}{2}, \\ p, & p = \frac{1}{2}, \\ \frac{2p - \sqrt{p(1-p)} + 1}{3}, & p > \frac{1}{2}, \end{cases} \tag{16.4}$$

and, for any sigma level k, (14.3) and (14.21),

$$\text{loc}(D_2 | \text{any } k) = q \log \frac{m + x_C}{m} \,, \tag{16.5}$$

where m is the reference asset position of the decision maker.

If we set $p = 0.001$, $x_U = 5000$, and $x_C = 5$, and if we assume[2] a reference asset position of freely expendable income of $m = 1000$ for Israeli graduate students, then we find the respective loci, (16.3), (16.5), and a Weber constant of $q = 100$, (12.14),

$$\text{loc}(D_1 | k = 1) = 2.0 \text{ utile} \tag{16.6}$$

and

$$\text{loc}(D_2 | \text{any } k) = 0.5 \text{ utile.} \tag{16.7}$$

So, the Bayesian decision theory picks the first bet as the most profitable one, as the first option is slightly preferred over the second option with

$$\text{loc}(D_1 | k = 1) - \text{loc}(D_2 | \text{any } k) = 1.5 \text{ utile,} \tag{16.8}$$

where one utile corresponds with a just noticeable difference in utility. The result of this decision theoretical analysis is consistent with the finding that 72% of $N = 72$ subjects prefer decision D_1 over D_2 [53].

---

[2]In [53] it is stated that the modal income in Israel is 3000 Israeli pounds.

16.2 Risk Aversion in the Negative Domain

The above analysis may also be performed for the case when we change the sign in the outcomes in (16.1) and (16.2), so that there is a small probability of loosing a large sum of money. We then will see a reversal in the preference for bet D_1 over bet D_2 to a preference for bet D_2 over bet D_1.

The outcome probability distributions for the respective bets are:

$$p(x|D_1) = \begin{cases} 0.001, & x = -5000, \\ 0.999, & x = 0, \end{cases} \tag{16.9}$$

and

$$p(x|D_2) = \begin{cases} 1.0, & x = -5, \end{cases} \tag{16.10}$$

where x is in Israeli pounds. It is found that 83% of $N = 72$ subjects preferred the bet D_2 over D_1 [53].

The preference for a sure but much smaller loss over an uncertain but much larger loss constitutes risk aversion in the negative domain. Risk aversion in the negative domain represents our tendency to hedge against large and painful losses.

16.2.1 The Predicted Choice

If we set $p = 0.001$, $x_U = -5000$ and $x_C = -5$, (16.9) and (16.10), then we find for a reference asset position of [3] $m = 6000$ respective loci of, (16.3), (16.5), and a Weber constant of $q = 100$, (12.14),

$$\mathrm{loc}(D_1|k=1) = -2.01 \text{ utile} \tag{16.11}$$

and

$$\mathrm{loc}(D_2|\text{any } k) = -0.08 \text{ utile.} \tag{16.12}$$

So, the Bayesian decision theory picks the second bet as the most profitable one, as the second option is slightly preferred over the second option with

$$\mathrm{loc}(D_2|\text{any } k) - \mathrm{loc}(D_1|k=1) = 1.92 \text{ utile,} \tag{16.13}$$

where one utile corresponds with a just noticeable difference in utility. The result of this decision theoretical analysis is consistent with the finding that 83% of $N = 72$ subjects prefer decision D_2 over D_1 [53].

16.3 Risk Seeking in the Negative Domain

We now give an example of risk seeking when people must choose between a sure loss and a substantial probability of a larger loss. The outcome probability

---

[3]We cannot lose more wealth than we have without invoking the concept of the utility of debt (12.37), which is why we now set m from thousand to six thousand pounds.

distributions for the respective bets in our risk seeking example are

$$p(x \mid D_1) = \begin{cases} 0.8, & x = -4000, \\ 0.2, & x = 0, \end{cases} \tag{16.14}$$

and

$$p(x \mid D_2) = \begin{cases} 1.0, & x = -3000, \end{cases} \tag{16.15}$$

where x is in Israeli pounds. It is found that 92% of $N = 95$ subjects preferred the bet D_1 over D_2 [53].

The preference for a somewhat larger but uncertain loss over a somewhat smaller but certain loss constitutes risk seeking in the negative domain. Risk seeking in the negative domain represents our tendency to try to evade large and catastrophic losses.

16.3.1   The Predicted Choice, Part I

If we set $p = 0.8$ and $x_U = -4000$, (16.14), then we find for a reference asset position of $m = 6000$ the respective loci, (16.3), (16.5), and a Weber constant of $q = 100$, (12.14),

$$\text{loc}(D_1 \mid k = 1) = -73.24 \text{ utile} \tag{16.16}$$

and

$$\text{loc}(D_2 \mid \text{any } k) = -69.31 \text{ utile.} \tag{16.17}$$

So, the Bayesian decision theory picks the first bet as the most profitable one, as the second option is preferred over the first with

$$\text{loc}(D_2 \mid \text{any } k) - \text{loc}(D_1 \mid k = 1) = 3.93 \text{ utile,} \tag{16.18}$$

where one utile corresponds with a just noticeable difference in utility. The result of this decision theoretical analysis is inconsistent with the finding that an overwhelming 92% of $N = 95$ subjects preferred the bet D_1 over D_2 [53].

Now, if we take a reference asset position of $m = 36,000$, or, equivalently, the modal yearly aggregated Israeli household income [53], then the solutions will tip over, as they align themselves with the observed preferences. But it does so with only the smallest of margins (i.e., 0.06 utiles), which does not justify this very pronounced preference of 92% for the uncertainty decision D_1.

16.3.2   Imbalanced Risk Seeking in the Negative Domain

Until now it sufficed in our decision theoretical modeling of the fourfold pattern to assume a balanced trade-off between probabilistic worst- and best-cases that correspond with 1-sigma plausibility bounds. But in light of the observed strong preference for the uncertainty decision D_1 we are now forced to invoke the postulate of an imbalanced trade-off between probabilistic worst- and best-cases in order to accommodate the observed risk seeking in the negative domain.

In problems of choice involving large (near) certain losses there seems to be a tendency towards a more pronounced risk-taking than we would normally expect. This is the so-called 'possibility effect' which increases the desirability of desperate gambles in which we cling to a 'sliver of hope' as we accept a high probability of making things worse in exchange for a small chance of avoiding a large loss [55].

And as we examine our intuitions regarding risk seeking in the negative domain, then it would seem that this behavior is indeed a commitment to 'a leap of faith' towards the happy outcome of some best-case scenario, which in the current problem of choice would be the no-loss outcome of the uncertainty bet (16.14). Also, it is expected that the larger the loss we wish to avoid, the more desperate the gambles that we are willing to accept. Stated differently, the larger the certain loss relative to our reference asset position, the more extreme will be the best-case scenarios that we are willing to entertain.

The probabilistic best-case scenario, in the case of no upper bound overshoot, is given as a function of probabilities, outcomes, and the used k-sigma level, (13.9), (13.12), (13.14), (13.15) and (13.16),

$$ub(k_2) = E(U) + k_2 \operatorname{std}(U).$$ (16.19)

For k-sigma levels greater than

$$k_2 \geq \frac{b - E(U)}{\operatorname{std}(U)},$$ (16.20)

the maximax criterion b becomes the accounted for best-case scenario, as opposed to the less optimistic probabilistic best-case scenario $ub(k_2)$.

It follows that the sigma level k_2 is the parameter that modulates the forcefulness in which the probabilistic best-case scenario is taken into account, as we are under the sway of the possibility effect [55], with a limit of optimism that tends to the maximax criterion (13.14):

$$UB(k_2) \to b, \qquad \text{as } k_2 \to \frac{b - E(U)}{\operatorname{std}(U)}.$$ (16.21)

So, the postulate of imbalanced risk-seeking translates to the general criterion of choice (13.25),

$$\operatorname{loc}(k_1, k_2) = \frac{LB(k_1) + E(U) + UB(k_2)}{3},$$ (16.22)

where the forcefulness of the best-case scenario and, as a consequence, the attractiveness of that choice, is modulated upward as the upper bound sigma level k_2 is set to some value greater than one, while the weighting of the worst-case scenario remains the same, as k_1 is set to its default value of one.

It may be derived, as is done in Section 16.3.4, that the locus (16.22) for the utility probability distribution (16.14) simplifies to[4]

$$\operatorname{loc}^-(D_1|, k_1, k_2) = q \log \frac{m + x_U}{m} \, g^-(p, k_1, k_2),$$ (16.23)

---

[4]In order to keep the flow of exposition going, we have moved the derivation of the imbalanced locus under risk seeking to the end of this section.

where

$$
g^-(p, k_1, k_2) = \begin{cases} \frac{2p + k_1\sqrt{p(1-p)}}{3}, & p \leq \frac{1}{1+k_1^2}, \\ \frac{1+p}{3}, & \frac{1}{1+k_1^2} < p < \frac{k_2^2}{1+k_2^2}, \\ \frac{1 + 2p - k_2\sqrt{p(1-p)}}{3}, & p \geq \frac{k_2^2}{1+k_2^2}, \end{cases} \tag{16.24}
$$

for general k_1 and k_2.

16.3.3 The Predicted Choice, Part II

If we set $p = 0.8$, $x_U = -4000$, $x_C = -3000$, $m = 6000$, $k_1 = 1$, and $k_2 = 2$, then we find the respective loci, (16.5), (16.23), and a Weber constant of $q = 100$, (12.14),

$$
\mathrm{loc}^-(D_1 | k_1 = 1, k_2 = 2) = -65.92 \text{ utile} \tag{16.25}
$$

and

$$
\mathrm{loc}(D_2 | \text{any } k) = -69.31 \text{ utile}. \tag{16.26}
$$

So, the Bayesian decision theory picks the first bet as the most profitable one, as we now find that the first option is preferred over the second option with

$$
\mathrm{loc}^-(D_1 | k_1 = 1, k_2 = 2) - \mathrm{loc}(D_2 | \text{any } k) = 3.40 \text{ utile},
$$

where one utile corresponds with a just noticeable difference in utility.

The result of this decision theoretical analysis in which an imbalanced criterion of choice is used in the direction of daring is in line with the finding that 92% of $N = 95$ subjects prefer decision D_1 over D_2 [53].

So, for risk seeking in the negative domain with large certain losses, we find that the observed betting preferences, which seem to be under the influence of the possibility effect [55], point to an imbalanced risk seeking in which the probabilistic best-case scenario is taken more forcefully into account than the probabilistic worst-case scenario.

16.3.4 Some Derivations

The uncertainty decision D_1 has an outcome probability distribution (16.14) of the form

$$
f(x | D_1) = \begin{cases} p, & x = x_U, \\ 1 - p, & x = 0. \end{cases} \tag{16.27}
$$

The corresponding utility distribution is be obtained by a simple relabeling, wherein the outcomes x in (16.14) are transformed into their corresponding utilities by way Bernoulli's utility function (13.4):

$$
f(u | D_1) = \begin{cases} p, & u = u(x_U | m), \\ 1 - p, & u = 0, \end{cases} \tag{16.28}
$$

where

$$u(x_U \,|\, m) = q \,\log \frac{m + x_U}{m}. \tag{16.29}$$

Since the probability distribution (16.28) is a Bernoulli distribution with one zero outcome, we have that the expected value and standard deviation are given as

$$E(U \,|\, D_1) = u(x_U \,|\, m) \; p \tag{16.30}$$

and

$$\mathrm{std}(U \,|\, D_1) = |u(x_U \,|\, m)| \sqrt{p\,(1-p)}. \tag{16.31}$$

For negative prospects $u(x_U \,|\, m) < 0$, the k_1-sigma uncorrected lower and the k_2-sigma uncorrected upper bound for negative uncertain outcomes x_U are given as, (13.11), (13.15), (16.29), (16.30), and (16.31),

$$lb^-(k_1) = u(x_U \,|\, m) \; p - k_1 \; |u(x_U \,|\, m)| \sqrt{p\,(1-p)}$$

$$\tag{16.32}$$

$$= u(x_U \,|\, m) \left(p + k_1 \; \sqrt{p\,(1-p)} \right)$$

and

$$ub^-(k_2) = p\, u(x_U \,|\, m) + k_2 \; \sqrt{p\,(1-p)} \; |u(x_U \,|\, m)|$$

$$\tag{16.33}$$

$$= u(x_U \,|\, m) \left(p - k_2 \; \sqrt{p\,(1-p)} \right).$$

Since we have negative $u(x_U \,|\, m)$, the minimax and maximax values, (13.10) and (13.14), may be read from (16.28) as

$$a^- = u(x_U \,|\, m) \qquad \text{and} \qquad b^- = 0. \tag{16.34}$$

We also have that

$$p - k\sqrt{p\,(1-p)} < 0, \qquad \text{for } p < \frac{k^2}{1 + k^2}. \tag{16.35}$$

and

$$p + k\sqrt{p\,(1-p)} > 1, \qquad \text{for } p > \frac{1}{1 + k^2}. \tag{16.36}$$

It follows that lower bound undershoot and upper bound overshoot will occur whenever, (16.32), (16.34), and (16.36),

$$\begin{cases} lb^-(k_1) < a^- & p > \frac{1}{1+k_1^2}, \\ lb^-(k_1) \geq a^- & p \leq \frac{1}{1+k_1^2}, \end{cases} \tag{16.37}$$

and, (16.33), (16.34), and (16.35),

$$\begin{cases} ub^-(k_2) \leq b^-, & p \geq \frac{k_2^2}{1+k_2^2}, \\ ub^-(k_2) > b^-, & p < \frac{k_2^2}{1+k_2^2}. \end{cases} \tag{16.38}$$

So the corrected lower and upper sigma bounds are given as, (13.13), (16.32), (16.34), and (16.37),

$$
LB^-(k_1) = \begin{cases} u(x_U \mid m), & p > \frac{1}{1+k_1^2}, \\ u(x_U \mid m)\left(p + k_1 \sqrt{p(1-p)}\right), & p \le \frac{1}{1+k_1^2}, \end{cases} \tag{16.39}
$$

and, (13.16), (16.33), (16.34), and (16.38),

$$
UB^-(k_2) = \begin{cases} u(x_U \mid m)\left(p - k_2 \sqrt{p(1-p)}\right), & p \ge \frac{k_2^2}{1+k_2^2}, \\ 0, & p < \frac{k_2^2}{1+k_2^2}. \end{cases} \tag{16.40}
$$

By way of (16.22), (16.30), (16.39), and (16.40), we may set up the following identity for the general locus of uncertainty bets in the negative domain:

$$
\mathrm{loc}^-(D_1 \mid k_1, k_2) = \begin{cases} u(x_U \mid m)\frac{1}{3}\left[3p + (k_1 - k_2)\sqrt{p(1-p)}\right], & \frac{k_2^2}{1+k_2^2} \le p \le \frac{1}{1+k_1^2}, \\ u(x_U \mid m)\frac{1}{3}\left(1 + 2p - k_2\sqrt{p(1-p)}\right), & p \ge \max\left(\frac{1}{1+k_1^2}, \frac{k_2^2}{1+k_2^2}\right), \\ u(x_U \mid m)\frac{1}{3}\left(2p + k_1\sqrt{p(1-p)}\right), & p \le \min\left(\frac{1}{1+k_1^2}, \frac{k_2^2}{1+k_2^2}\right), \\ u(x_U \mid m)\frac{1}{3}(1+p), & \frac{1}{1+k_1^2} < p < \frac{k_2^2}{1+k_2^2}, \end{cases}
$$
$$\tag{16.41}$$

where

$$
\frac{k_2^2}{1+k_2^2} \le p \le \frac{1}{1+k_1^2} \tag{16.42}
$$

will only hold for $k_1, k_2 \le 1$, and where, because in imbalanced risk seeking we have, by construction, that $k_2 > k_1$,

$$
\min\left(\frac{1}{1+k_1^2}, \frac{k_2^2}{1+k_2^2}\right) = \frac{1}{1+k_1^2} \tag{16.43}
$$

and

$$
\max\left(\frac{1}{1+k_1^2}, \frac{k_2^2}{1+k_2^2}\right) = \frac{k_2^2}{1+k_2^2}. \tag{16.44}
$$

It follows that we may simplify (16.41) as

$$
\mathrm{loc}^-(D_1 \mid, k_1, k_2) = u(x_U \mid m)\ g^-(p, k_1, k_2), \tag{16.45}
$$

or, equivalently, (16.29),

$$
\mathrm{loc}^-(D_1 \mid, k_1, k_2) = q \log \frac{m + x_U}{m}\ g^-(p, k_1, k_2), \tag{16.46}
$$

where

$$
g^-(p, k_1, k_2) = \begin{cases} \frac{2p + k_1\sqrt{p(1-p)}}{3}, & p \le \frac{1}{1+k_1^2}, \\ \frac{1+p}{3}, & \frac{1}{1+k_1^2} < p < \frac{k_2^2}{1+k_2^2}, \\ \frac{1+2p - k_2\sqrt{p(1-p)}}{3}, & p \ge \frac{k_2^2}{1+k_2^2}, \end{cases} \tag{16.47}
$$

for general k_1 and k_2.

16.4   Risk Aversion in the Positive Domain

The previous analysis may also be performed for the opposite case of a sure gain and a substantial probability of a larger gain. We then will see a reversal in the preference for bet D_1 over bet D_2 to a preference for bet D_2 over bet D_1.

The outcome probability distributions for this problem of choice are[5]:

$$p(x|D_1) = \begin{cases} 0.8, & x = 4000, \\ 0.2, & x = 0, \end{cases} \tag{16.48}$$

and

$$p(x|D_2) = \begin{cases} 1.0, & x = 3000, \end{cases} \tag{16.49}$$

where x is in Israeli pounds. It is found that 80% of $N = 95$ subjects preferred the bet D_2 over D_1 [53]. Risk aversion in the positive domain represents our tendency to want to secure our profits.

16.4.1   The Predicted Choice

If we set $p = 0.8$ and $x_U = 4000$, (16.14), then we find for a reference asset position of $m = 1000$ the respective loci, (16.3), (16.5), and a Weber constant of $q = 100$, (12.14),

$$\text{loc}(D_1|k = 1) = 118.03 \text{ utile} \tag{16.50}$$

and

$$\text{loc}(D_2|\text{any } k) = 138.63 \text{ utile}. \tag{16.51}$$

So, the Bayesian decision theory picks the second bet as the most profitable one, the second option is preferred over the first with a very forceful

$$\text{loc}(D_2|\text{any } k) - \text{loc}(D_1|k = 1) = 20.60 \text{ utile}, \tag{16.52}$$

where one utile corresponds with a just noticeable difference in utility.

The result of this decision theoretical analysis is in line with the finding that 80% of $N = 95$ subjects prefer decision D_2 over D_1 [53].

In closing, it is to be noted that the procedure in Section 15.3.5 can be used to determine the actual upper confidence bound sigma level k_2 that were used. For it is very possible that preference of the second option over the first was greater than the in (16.52) reported 20.60 utile, as the the experimental subjects may have used implicit sigma levels k_2 greater than one.

16.5   Discussion

It has been found that Weaver's criterion of choice (16.22) can accommodate risk seeking, risk aversion, and the reflection effect [53], or, equivalently, the

---

[5]Compare with (16.14) and (16.15).

fourfold pattern of preferences, which is said to be one of the core achievements of prospect theory [55].

It is stated in [55] that the 'certainty effect' increases the desirability of certain large gains, whereas the 'possibility effect' increases the desirability of desperate gambles in which we cling to a 'sliver of hope' as we accept a high probability of making things worse in exchange for a small chance of avoiding a large loss. And it has been found, in both this chapter and the previous one, that the sigma levels k_1 and k_2 in the general Weaver criterion of choice (16.22) are the parameters by which the strength of the certainty and possibility effects can be modulated, respectively.

For risk aversion in the positive domain with large certain gains, it is found that the observed betting preferences, which seem to be under the influence of the possibility effect, point to an imbalanced risk seeking in which the probabilistic worst-case scenario is taken more forcefully into account, by way of a sigma level k_1 of the undershoot corrected lower confidence bound in (16.22) greater than one, as discussed in the previous chapter.

And for risk seeking in the negative domain with large certain losses, it is found that the observed betting preferences, which seem to be under the influence of the possibility effect, point to an imbalanced risk seeking in which the probabilistic best-case scenario is taken more forcefully into account, by way of a sigma level k_2 of the overshoot corrected upper confidence bound in (16.22) greater than one, as discussed in this chapter.

Chapter 17

Bottomry Loans

Before premium-based insurances were well and truly introduced in the Northern Netherlands, approximately around the mid-sixteenth century, merchants and ship-owners fell back on different methods for dealing with the financial consequences of long-distance maritime trade. A well-known and often applied construction was known as bottomry (bodemerij).

With bottomry a loan was taken out, which was only to be repaid, with an agreed upon interest, if the vessel or merchandise arrived safely at the port of destination. So, bottomry loans incorporated both an insurance and a finance component. In what follows, we will take this practice of bottomry as a decision theoretical case study.

In what follows we will derive lower and upper bounds on the bottomry loan interest rate, as set by the money lender and the merchant, respectively, for the criterion of choice that is proposed by the Bayesian decision theory. It is found that these interest rate bounds are intimately connected to the concept of the odds of winning a bet; i.e., the probability of winning a bet divided by the probability of losing that bet.

In order to bring this better to the fore, we also discuss the lower and upper bounds on the bottomry loan interest rates which follow from the alternative and more traditional expected utility maximization and Value at Risk (VaR) minimization, or, equivalently, for negatively signed losses, a lower confidence bound maximization. We then compare the predicted interest rate bounds on bottomry loans under the different criteria of choice, by way of a historical data point.

Finally, we will take a look at the effect of the regular interest rates on the bounds of the bottomry interest rates. And it is again found that these updated bounds are also intimately connected to the concept of the odds of winning a bet. For the lower bound on the bottomry interest rates leads to a class of adjusted odds that take into account the cost of money over some time period.

17.1   The Position Measure

In this case study we will use a balanced position measure, in which the worst- and best-case scenarios are given the same weight in terms of their plausibility; i.e., $k_1 = k_2 = k$ in (13.25), or, equivalently, (13.23). Moreover, we will let the decision makers navigate their decision space by way of 1-sigma, i.e., realistically plausible, worst- and best-case scenarios, (13.13) and (13.16).

In the understanding that $k_1 = k_2 = 1$ in (13.25), or, equivalently, $k = 1$ in (13.23), we now will drop explicit mention of the sigma levels, as this will allow us to remove some of the notational clutter down the line.

Let the 1-sigma undershoot corrected lower confidence bound be given as

$$LB = \begin{cases} a, & lb < a, \\ E(U) - \text{std}(U), & lb \geq a, \end{cases} \tag{17.1}$$

where a is minimax value and lb is the traditional 1-sigma lower confidence bound:

$$lb = E(U) - \text{std}(U). \tag{17.2}$$

Let the 1-sigma overshoot corrected upper confidence bound be given as

$$UB = \begin{cases} E(U) + \text{std}(U), & ub \leq b, \\ b, & ub > b, \end{cases} \tag{17.3}$$

where b is maximax value and ub is the traditional 1-sigma upper confidence bound:

$$ub = E(U) + \text{std}(U). \tag{17.4}$$

Then the position measure which is to be maximized is given as, (13.23),

$$\text{loc} = \frac{LB + E(U) + UB}{3}, \tag{17.5}$$

where, because of the under- and overshoot corrections in (17.1) and (17.3),

$$\frac{LB + E(U) + UB}{3} = \begin{cases} E(U), & lb \geq a, \ ub \leq b, \\ \frac{1}{3}\left[a + 2E(U) + \text{std}(U)\right], & lb < a, \ ub \leq b, \\ \frac{1}{3}\left[2E(U) - \text{std}(U) + b\right], & lb \geq a, \ ub > b, \\ \frac{1}{3}\left[a + E(U) + b\right], & lb < a, \ ub > b. \end{cases} \tag{17.6}$$

The balanced 1-sigma criterion of choice (17.5) is the position measure that takes into account the most likely scenario, together with the plausible worst- and best-case scenarios.

17.2   An Intermediate Result

In order to allow for a smooth discussion of our bottomry case study, we now will here derive the balanced 1-sigma position (17.5) for general two-outcome

probability distributions:

$$f(u) = \begin{cases} p, & u = u_1, \\ 1 - p, & u = u_2. \end{cases} \tag{17.7}$$

where

$$u_1 < u_2. \tag{17.8}$$

The minimax and maximax values of (17.7) are, respectively, (17.8),

$$a = \min(u_1, u_2) = u_1 \quad \text{and} \quad b = \max(u_1, u_2) = u_2. \tag{17.9}$$

The expected value and standard deviation of (17.7) are, respectively, [74],

$$E(U) = p\, u_1 + (1 - p)\, u_2 \tag{17.10}$$

and

$$\text{std}(U) = \sqrt{p\,(1 - p)}\,(u_2 - u_1). \tag{17.11}$$

If we solve the lower bound undershoot condition in (17.1),

$$lb < a, \tag{17.12}$$

or, equivalently, (17.2), (17.9), (17.10), and (17.11),

$$p\, u_1 + (1 - p)\, u_2 - \sqrt{p\,(1 - p)}\,(u_2 - u_1) < u_1, \tag{17.13}$$

for the probability p, then we find that lower bound undershoot will occur whenever $p > 1/2$. This allows us to rewrite the undershoot corrected lower bound (17.1) as

$$LB = \begin{cases} a, & p > \frac{1}{2}, \\ E(U) - \text{std}(U), & p \le \frac{1}{2}. \end{cases} \tag{17.14}$$

Likewise, if we solve the upper bound overshoot condition in (17.3),

$$ub > b, \tag{17.15}$$

or, equivalently, (17.4), (17.9), (17.10), and (17.11),

$$p\, u_1 + (1 - p)\, u_2 + \sqrt{p\,(1 - p)}\,(u_2 - u_1) > u_2, \tag{17.16}$$

for the probability p, then we find that upper bound overshoot will occur whenever $p < 1/2$. This allows us to rewrite the overshoot corrected upper bound (17.3) as

$$UB = \begin{cases} E(U) + \text{std}(U), & p \ge \frac{1}{2}, \\ b, & p < \frac{1}{2}. \end{cases} \tag{17.17}$$

Moreover, (17.14) and (17.17) allow us to rewrite the 1-sigma balanced position measure (17.6) as

$$
\frac{LB + E(U) + UB}{3} = \begin{cases} \frac{1}{3}\left[2E(U) - \operatorname{std}(U) + b\right], & p < \frac{1}{2}, \\ E(U), & p = \frac{1}{2}, \\ \frac{1}{3}\left[a + 2E(U) + \operatorname{std}(U)\right], & p > \frac{1}{2}, \end{cases} \tag{17.18}
$$

or, equivalently, (17.9), (17.10), and (17.11),

$$
\frac{LB + E(U) + UB}{3} = \begin{cases} \frac{1}{3}\left(2\left[p\,u_1 + (1-p)\,u_2\right] - \sqrt{p(1-p)}\,(u_2 - u_1) + u_2\right), & p < \frac{1}{2}, \\ p\,u_1 + (1-p)\,u_2, & p = \frac{1}{2}, \\ \frac{1}{3}\left(u_1 + 2\left[p\,u_1 + (1-p)\,u_2\right] + \sqrt{p(1-p)}\,(u_2 - u_1)\right), & p > \frac{1}{2}, \end{cases} \tag{17.19}
$$

Collecting the terms in u_1 and u_2, we may rewrite (17.19) more succinctly as

$$
\frac{LB + E(U) + UB}{3} = u_1\,g_1(p) + u_2\,g_2(p), \tag{17.20}
$$

where

$$
g_1(p) = \begin{cases} \frac{2p + \sqrt{p(1-p)}}{3}, & p < \frac{1}{2}, \\ p, & p = \frac{1}{2}, \\ \frac{2p - \sqrt{p(1-p)} + 1}{3}, & p > \frac{1}{2}, \end{cases} \tag{17.21}
$$

and

$$
g_2(p) = \begin{cases} \frac{2(1-p) - \sqrt{p(1-p)} + 1}{3}, & p < \frac{1}{2}, \\ 1 - p, & p = \frac{1}{2}, \\ \frac{2(1-p) + \sqrt{p(1-p)}}{3}, & p > \frac{1}{2}, \end{cases} \tag{17.22}
$$

It is to be noted that (17.21) is the 1-sigma position (17.5) of a Bernoulli event s having probability p:

$$
f(s) = \begin{cases} 1 - p, & s = 0, \\ p, & s = 1, \end{cases} \tag{17.23}
$$

whereas, (17.22) is the 1-sigma position of the complementary Bernoulli event $\overline{s}$ having probability $1 - p$:

$$
f(\overline{s}) = \begin{cases} p, & \overline{s} = 0, \\ 1 - p, & \overline{s} = 1. \end{cases} \tag{17.24}
$$

This may be verified if one replaces (17.7) with either (17.23) or (17.24), and then proceeds from equation (17.9) to equation (17.20).

17.3 The Bottomry Loan Case Study

We have a merchant with a current wealth of m guilders. The one contingency the merchant wishes to have covered is the loss of his cargo. This loss represents a monetary damage of L guilders, which would greatly reduce the merchants wealth. But if his cargo safely reaches the harbor, then the merchant stands to generate a revenue with which he can buy his cargo C times over, with $C > 1$.

We also have a money lender with a much greater current wealth of M guilders, which vastly exceeds the range of the monetary damages L. The money lender, a retired merchant, provides bottomry loans for an interest rate c, collectable, together with the loan itself, once the cargo safely reaches the harbor.

The probability of the cargo being lost at sea is estimated by both merchant and money lender to be p.

17.3.1 The Money Lender

The money lender, having an initial wealth of M, may decide to provide the merchant with a bottomry loan of L in exchange for an interest rate c, where $c < C - 1$, to be collectable, together with the loan itself, once the cargo safely reaches the harbor. Under this decision D_1, the worst-case scenario for the money lender is that the ship and its cargo would be lost at sea (probability of p), in which case the money lender will incur a loss of

$$x = -L. \tag{17.25}$$

The best-case scenario is that the ship and its cargo safely reach the harbor (probability of $1 - p$), in which case the money lender will gain a profit of

$$x = -L + (1 + c)\, L = cL. \tag{17.26}$$

So, the under the decision D_1 to provide a bottomry loan, the money lender has the following outcome probability distribution, (17.25) and (17.26):

$$f(x|\, D_1) = \begin{cases} p, & x = -L, \\ 1 - p, & x = cL. \end{cases} \tag{17.27}$$

By way of Bernoullis utility function (13.4), we may transform this outcome probability distribution to a corresponding utility probability distribution:

$$f(u|\, D_1) = \begin{cases} p, & u = q \log \frac{M-L}{M}, \\ 1 - p, & u = q \log \frac{M+cL}{M}. \end{cases} \tag{17.28}$$

If an asset position m vastly exceeds the monetary increment x, then, by way of the series expansion

$$\log (1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \tag{17.29}$$

we will have that for the person who holds such an asset position there will be a near linear utility for money in the neighborhood of x, or, equivalently, (13.4) and (17.29),

$$u(x\,|\,m) = q \log \frac{m+x}{m} \approx q \frac{x}{m}. \tag{17.30}$$

So, if we assume that the money lender, who is a rich retired merchant, has a wealth M that vastly exceeds the worth of the insured cargo L, then we may apply the approximation (17.30) to the utilities in (17.28). This gives the following approximate utility probability distribution:

$$f(u\,|\,D_1) = \begin{cases} p, & u = -q\frac{L}{M}, \\ 1-p, & u = q\frac{cL}{M}, \end{cases} \tag{17.31}$$

which has the 1-sigma position (i.e., risk index), (17.5) and (17.20),

$$\mathrm{loc}(D_1) = q \frac{L}{M} \left[-g_1(p) + c\, g_2(p) \right]. \tag{17.32}$$

Under the decision D_2 that the money lender does not provide a bottomry loan to the merchant, the money lender will be certain to incur neither a loss nor a gain:

$$x = 0. \tag{17.33}$$

This corresponds with the outcome probability distribution:

$$f(x\,|\,D_2) = \Big\{ 1, \qquad x = 0, \tag{17.34}$$

By way of Bernoullis utility function (13.4), we may transform this outcome probability distribution to a corresponding utility probability distribution:

$$f(u\,|\,D_2) = \Big\{ 1, \qquad u = q \log \tfrac{M}{M} = 0. \tag{17.35}$$

Since the position of a certain outcome is that certain outcome, (13.64) and (14.21), we then have that

$$\mathrm{loc}(D_2) = 0. \tag{17.36}$$

Now, it is assumed that the money lender will agree to provide to provide a bottomry contract if the position of the utility probability distribution under D_1 is greater than the position of the utility probability distribution under D_2:

$$\mathrm{loc}(D_1) > \mathrm{loc}(D_2), \tag{17.37}$$

or, equivalently, (17.32) and (17.36),

$$q \frac{L}{M} \left[-g_1(p) + c\, g_2(p) \right] > 0. \tag{17.38}$$

The decision theoretical inequality (17.38) solves for a lower bound on the interest rate c, as determined by the money lender:

$$c > h(p), \tag{17.39}$$

where, (17.21) and (17.22),

$$h(p) = \frac{g_1(p)}{g_2(p)} = \begin{cases} \frac{2p + \sqrt{p(1-p)}}{2(1-p) - \sqrt{p(1-p)} + 1}, & p < \frac{1}{2}, \\ \frac{p}{1-p}, & p = \frac{1}{2}, \\ \frac{2p - \sqrt{p(1-p)} + 1}{2(1-p) + \sqrt{p(1-p)}}, & p > \frac{1}{2}, \end{cases} \qquad (17.40)$$

is the odds of the 1-sigma positions of the Bernoulli events of the sinking and the not-sinking of the ship and its cargo, (17.23) and (17.24), respectively.

Note that the interest rate (17.40) goes to the regular odds for $p = 1/2$, as the skewness of the utility probability distribution (17.31) goes to zero and the 1-sigma lower and upper confidence bounds, consequently, cancel each other out.

17.3.2 The Merchant

Let m be the amount of money the merchant initially had, before buying his cargo. If the merchant, after having bought L guilders worth of cargo decides to take out the bottomry loan D_1, then the worst-case scenario for the merchant is that the ship and its cargo would be lost at sea (probability of p), in which case there is neither a gain nor a loss for the merchant:

$$x = -L + L = 0, \qquad (17.41)$$

as the merchant in this case may keep his bottomry loan. The best-case scenario is that the ship and its cargo safely reach the harbor (probability of $1 - p$), in which case the merchant will gain a profit of

$$x = -L + L + [C - (1 + c)]\,L = (C - c - 1)\,L, \qquad (17.42)$$

as we let the repayment factor $1 + c$ be smaller than the return factor C,

$$1 + c < C, \qquad (17.43)$$

in order to ensure a minimal compensation for the merchant for all his hard work.

So, the under the decision D_1 to take out a bottomry loan, the merchant has the following outcome probability distribution, (17.42) and (17.43):

$$f(x|\,D_1) = \begin{cases} p, & x = 0, \\ 1 - p, & x = (C - c - 1)\,L. \end{cases} \qquad (17.44)$$

By way of Bernoullis utility function (13.4), we may transform this outcome probability distribution to a corresponding utility probability distribution:

$$f(u|\,D_1) = \begin{cases} p, & u = 0, \\ 1 - p, & u = q\log\frac{m + (C - c - 1)L}{m}, \end{cases} \qquad (17.45)$$

which has the 1-sigma position (i.e., risk index), (17.5) and (17.20),

$$\mathrm{loc}(D_1) = q \log \left(\frac{m + (C - c - 1)L}{m} \right) g_2(p). \qquad (17.46)$$

Alternatively, under the decision not to hedge against the possible loss of his cargo D_2 the merchant will either incur a loss on the cargo he bought if the cargo is lost at sea (probability of p)

$$x = -L, \qquad (17.47)$$

or make a profit if the cargo safely reaches its destination (probability of $1 - p$)

$$x = -L + CL = (C - 1)L, \qquad (17.48)$$

where $C > 1$. So, the under the decision D_2 to refrain from taking out a bottomry loan, the merchant has the following outcome probability distribution, (17.47) and (17.48): (17.42) and (17.43):

$$f(x|D_2) = \begin{cases} p, & x = -L, \\ 1 - p, & x = (C - 1)L. \end{cases} \qquad (17.49)$$

By way of Bernoullis utility function (13.4), we may transform this outcome probability distribution to a corresponding utility probability distribution:

$$f(u|D_2) = \begin{cases} p, & u = q \log \frac{m - L}{m}, \\ 1 - p, & u = q \log \frac{m + (C - 1)L}{m}, \end{cases} \qquad (17.50)$$

which has the 1-sigma position (i.e., risk index), (17.5) and (17.20),

$$\mathrm{loc}(D_2) = q \log \left(\frac{m - L}{m} \right) g_1(p) + q \log \left(\frac{m + (C - 1)L}{m} \right) g_2(p). \qquad (17.51)$$

It is assumed that the merchant will take out a bottomry loan only if the utility distribution under the decision to take out the loan D_1 is located more to the right than the utility probability distribution under the decision not to take out the loan D_2:

$$\mathrm{loc}(D_1) > \mathrm{loc}(D_2), \qquad (17.52)$$

or, equivalently, (17.46) and (17.51),

$$q \log \left(\frac{m + (C - c - 1)L}{m} \right) g_2(p) > q \log \left(\frac{m - L}{m} \right) g_1(p) + q \log \left(\frac{m + (C - 1)L}{m} \right) g_2(p).$$
$$(17.53)$$

Solving the decision theoretical inequality, we find the upper bound of the interest rate c, as determined by the merchant:

$$c < \left(C - 1 + \frac{m}{L} \right) \left[1 - \left(\frac{m - L}{m} \right)^{h(p)} \right], \qquad (17.54)$$

where, (17.40),

$$h(p) = \frac{g_1(p)}{g_2(p)} = \begin{cases} \frac{2p + \sqrt{p(1-p)}}{2(1-p) - \sqrt{p(1-p)} + 1}, & p < \frac{1}{2}, \\ \frac{p}{1-p}, & p = \frac{1}{2}, \\ \frac{2p - \sqrt{p(1-p)} + 1}{2(1-p) + \sqrt{p(1-p)}}, & p > \frac{1}{2}, \end{cases}$$

is the odds of the 1-sigma positions (17.21) and (17.22).

Note that in (17.54) we again encounter the adjusted odds (17.40). Moreover, it may be checked numerically that if the merchant himself has an ample fortune, such that $L/m \to 0$, or, equivalently, his utility for money becomes linear, (17.29) and (17.30), then the interest upper bound (17.54) will tend to the lower bound (17.39): i.e.,

$$\left(C - 1 + \frac{m}{L}\right)\left[1 - \left(\frac{m-L}{m}\right)^{h(p)}\right] \to h(p) \qquad \text{as} \qquad \frac{L}{m} \to 0. \qquad (17.55)$$

So, as the merchant gets richer, the maximum fraction of profit the merchant is willing to share in return for bottomry insurance converges to the minimum fraction of the profit the money lender wishes to receive for that insurance. And it follows that the margin of profit for the money lender will evaporate, as the merchant he services becomes rich enough to become his own insurer. This phenomenon, on a conceptual level, seems to be so intuitive that it could serve as a fundamental common-sense boundary condition in the modeling of any insurance problem. On a mathematical level, however, this phenomenon translates to a limit (17.55) which is highly non-intuitive, in terms of its inner workings.

In closing, also note that it follows from (17.54) that the interest rate c is linear in the return factor C. So, it is predicted by this decision theoretical analysis that a projected increase in the turnover C will make the merchant more inclined to pay a larger interest rate c on the same bottomry loan of L guilders, which guarantees him a risk-less opportunity for profit.

17.3.3   Examining the Odds

The margin of profit which is to be had by the money lender for his bottomry loan, is determined by both the interest lower bound of the money lender and the interest upper bound of the merchant. It follows from these bounds, (17.39), and (17.54), that the interest rate c that is both acceptable to the money lender and the merchant must lie in the range

$$h(p) < c < \left(C - 1 + \frac{m}{L}\right)\left[1 - \left(\frac{m-L}{m}\right)^{h(p)}\right], \qquad (17.56)$$

where, (17.40),

$$h(p) = \frac{g_1(p)}{g_2(p)} = \begin{cases} \frac{2p+\sqrt{p(1-p)}}{2(1-p)-\sqrt{p(1-p)}+1}, & p < \frac{1}{2}, \\ \frac{p}{1-p}, & p = \frac{1}{2}, \\ \frac{2p-\sqrt{p(1-p)}+1}{2(1-p)+\sqrt{p(1-p)}}, & p > \frac{1}{2}, \end{cases} \qquad (17.57)$$

is the odds of the 1-sigma positions (17.21) and (17.22).

If both our money lender and merchant are very pessimistic, in that they choose to disregard both the best-case and the most likely scenarios, as they only wish to focus on the worst-case scenario, then they may decide, in the spirit of a Value at Risk (VaR) approach [50], to choose that course of action which maximizes their k-sigma lower confidence bounds (13.13):

$$LB(D_i\,|\,k) = \begin{cases} a_i, & lb(k) < a_i, \\ E(U\,|\,D_i) - k\,\mathrm{std}(U\,|\,D_i), & lb(k) \geq a_i, \end{cases} \qquad (17.58)$$

where a_i is minimax value under decision D_i and $lb(k)$ is the traditional k-sigma lower confidence bound:

$$lb(k) = E(U\,|\,D_i) - k\,\mathrm{std}(U\,|\,D_i). \qquad (17.59)$$

By taking the 1-sigma lower bound undershoot corrected lower bound (17.58) as a position measure, while dropping the explicit mention of $k = 1$, per our notation,

$$\mathrm{loc}(D_i) = LB(D_i), \qquad (17.60)$$

we obtain, as we make the necessary adjustments in steps (17.7) through (17.54), the following bounds on the interest factor:

$$s(p) < c < \left(C - 1 + \frac{m}{L}\right)\left[1 - \left(\frac{m-L}{m}\right)^{s(p)}\right], \qquad (17.61)$$

where

$$s(p) = \begin{cases} \frac{p+\sqrt{p(1-p)}}{(1-p)-\sqrt{p(1-p)}}, & p \leq \frac{1}{2}, \\ \infty, & p > \frac{1}{2}, \end{cases} \qquad (17.62)$$

is the odds of the 1-sigma upper bound of the Bernoulli event of the cargo being lost at sea (17.23) and the 1-sigma lower bound of the cargo not being lost at sea (17.24).

So, a singular focus on the probabilistic worst-case scenario (17.60) translates to a state of informed pessimism in which we assign a high credence to the worst-case scenario of the cargo being lost at sea and a low credence to the best-case scenario of the cargo not being lost at sea. Also, as the probability of a loss at sea approaches 1/2 then, with an interest lower bound of infinity, no bottomry loan is ever to be provided by the money lender.

Alternatively, if our money lender and money lender are more traditional, in that they choose to disregard both the worst- and best-case scenarios, as they only wish to focus on the most likely scenario, then they may decide to choose that course of action which maximizes their expected utility (17.10), [6]:

$$\text{loc}(D_i) = E(U \mid D_i). \tag{17.63}$$

The criterion of choice (17.63) translates to the position measure, as we make the necessary adjustments in steps (17.7) through (17.54),

$$\omega(p) < c < \left(C - 1 + \frac{m}{L}\right)\left[1 - \left(\frac{m - L}{m}\right)^{\omega(p)}\right], \tag{17.64}$$

where

$$\omega(p) = \frac{p}{1 - p} \tag{17.65}$$

are the traditional odds that the ship and its cargo will sink.

The adjusted odds (17.57) and (17.62) take into account the 1-sigma bound information we have regarding the probabilities of the winning and the losing of the 'bet', just like the position measures from which they were derived, (17.5) and (17.60), respectively. The traditional odds does not take this information account, just like the position measure from which it was derived, (17.63).

17.3.4 Some Tentative Historical Data Points

It may be distilled from [8] and [29] that for the 16th century Dutch Levant trade, which was both risky and profitable, return factors ranging from $C = 2$ to $C = 4$ and a ship-loss frequency of $p = 1/20$ would seem to be reasonable estimates.

So, if both the money lender and the merchant use the 1-sigma balanced position measure (17.5) and if the latter has a cargo which represents a $L = 200$ guilders in worth, then the minimum interest rate which the money lender will demand on his loan in order to cover his risk exposure is $c = 0.12$, or, equivalently, 12%. Whereas, the merchant, who has an initial wealth of $m = 300$ guilders, of which $L = 200$ guilders is invested in cargo bound for the Levant with a promised return factor of $C = 3$ upon delivery, is willing to repay the loan with an interest rate of $c = 0.43$, or, equivalently, 43%, seeing that without this loan he stands a small but still very real chance to lose two thirds of his fortune in case of a shipwreck.

It follows that by agreeing to the bottomry contract the money lender stands to make a hefty margin of profit of 31% in interest rate, in terms of the actual risk incurred, if he manages to correctly gauge the merchants interest upper bound.

Furthermore, the upper bound of the interest rate c that he merchant is willing to pay is linear in the return factor C, (17.54). So, for $p = 1/20$,

$m = 300$ guilders, and $L = 200$ guilders, we may obtain the following relation for this interest rate:

$$c\left(C\right) = 0.061 + 0.122\,C. \qquad (17.66)$$

And we see in that for every unit increase in the return factor C will increases the maximum interest rate that the merchant is willing to pay on his bottomry loan with about 12.2%.

In Table 17.1, we give for return factors of $C = 2, 3, 4$, the bottomry interest rate upper bounds for the 1-sigma VaR criterion of choice (17.60), the expected utility criterion of choice (17.63), and the balanced 1-sigma criterion of choice (17.5), i.e., right-hand terms of (17.61), (17.64), and (17.56), respectively.

	$LB(D_i)$	$E(U\,|\,D_i)$	$\frac{1}{3}\left[LB + E(U\,|\,D_i) + LB\right]$
$C = 2$	83%	14%	31%
$C = 3$	116%	20%	43%
$C = 4$	149%	25%	55%

Table 17.1: Interest rate upper bounds for bottomry loans, as determined by the merchant, for different return factors C, for a probability of a shipwreck of $p = 1/20$, an initial wealth of $m = 300$ guilders of which $L = 200$ are invested in merchandise that is bound for the Levant.

And it can be noted that only for the balanced 1-sigma criterion of choice (17.5) these interest factors are within the historical interest bounds of 30% to 70%, as reported in [29].

Also, regarding the 1-sigma VaR criterion of choice (17.60), we may offer up additional observations which are in contradiction with the predicted interest rate lower bound, i.e., the left-hand term in (17.61), which tells us that for probabilities of a loss at sea greater than $p = 1/2$, no bottomry loan is ever to be provided by the money lender.

The Far East trade at the beginning of the 16th century was both extremely dangerous, i.e., probability p of a loss at sea greater than $1/2$, and spectacularly profitable, i.e., large return factors C. When Vasco Da Gama returned from his first voyage to the Indies in 1499, he had enough pepper, cinnamon, and cloves in his cargo hold to pay his expedition's cost sixty times over. One hundred years later, the Indies trade was still as profitable as it had been in Da Gama's time. When the Golden Hind returned to Plymouth in 1580 laden with the riches of the Far East, its contents repaid Francis Drake's backers fifty pounds for every one invested.

Also, following the discovery of the Portuguese spice sea route in 1597 by the Dutch, there was a rush on fine spices by Amsterdam merchants [8]. Within the year fourteen expeditions by six different trading companies, sixty-five ships in total, with an aggregated ship loss rate of about 33%, were send around the Cape of Good Hope. This influx of traders threatened to squeeze the profits right out of the spice trade. So, in order to remedy the situation the Dutch government established in 1602 a single combined monopoly organization to handle all commerce to the Indies. Investors provided this newly established

V.O.C. with 6.5 million guilders in initial funding to hire men, purchase ships, and acquire silver and trade goods to exchange for spices [8].

So, in the Far East trade we have multiple historical examples of a high-risk, i.e., probability p of failure of a commercial venture in excess of a half, be it on an individual ship level or on a company level[1]) and high-yield commercial venture where investment money was easy enough to come by, i.e., interest factors c smaller than infinity, which is in contradiction with (17.61).

17.4 The Cost of Money

We now add an additional layer of complexity in our decision theoretical analysis, as we also take into account the cost of money, by way of the interest rate for regular, non-bottomry loans.

In 16th century Holland reputable borrowers paid 4% (i.e., $r = 0.04$) on their loans, with the Dutch government getting its credit at the lowest rate of all. For comparison, in England reputable borrowers paid 10%, and the crown, who was not considered to be reputable, as it could, and often did, repudiate its loans, had to pay higher rates than good commercial borrowers [8].

17.4.1 The Money Lender

The money lender, rather than providing a risky bottomry loan to the merchant, can also choose to lend his L guilders out to reputable borrowers in exchange for a regular interest on that loan. Under the updated decision $D_{2'}$ that the money lender does not provide a bottomry loan to the merchant, the money lender will obtain with certainty a gain of

$$x = \left((1+r)^t - 1 \right) L, \tag{17.67}$$

where r is the going interest rate (in decimals) per time period t on regular, non-bottomry loans. So, the outcome probability distribution (17.34) is updated to the outcome probability distribution

$$f(x \mid D_{2'}) = \left\{ 1, \qquad x = \left((1+r)^t - 1 \right) L, \right. \tag{17.68}$$

By way of the approximation (17.30), we may transform this outcome probability distribution to a corresponding utility probability distribution:

$$f(u \mid D_{2'}) = \left\{ 1, \qquad u = q^{\frac{\left((1+r)^t - 1 \right) L}{M}}. \right. \tag{17.69}$$

Since the position of a certain outcome is that certain outcome, (13.64) and (14.21), we then have the following position of the the utility probability distribution under the updated decision $D_{2'}$:

$$\mathrm{loc}(D_{2'}) = q \frac{\left((1+r)^t - 1 \right) L}{M}. \tag{17.70}$$

---

[1]The Dutch West India Company (W.I.C.), a sister company to the V.O.C. that was chartered twenty years later, eventually went bankrupt.

It is assumed again that the money lender will agree to provide to provide a bottomry contract if the position of the utility probability distribution under D_1 is greater than the position of the utility probability distribution under $D_{2'}$:

$$\text{loc}(D_1) > \text{loc}(D_{2'}) , \tag{17.71}$$

or, equivalently, (17.32) and (17.70),

$$q\frac{L}{M}\left[-g_1(p) + c\, g_2(p)\right] > q\frac{L}{M}\left((1+r)^t - 1\right). \tag{17.72}$$

The decision theoretical inequality (17.72) solves for a lower bound on the bottomry interest rate c, as determined by the money lender and the going regular interest rate r:

$$c > h(p, r, t) , \tag{17.73}$$

where

$$h(p, r, t) = \frac{\left((1+r)^t - 1\right) + g_1(p)}{g_2(p)} \tag{17.74}$$

and, (17.21) and (17.22),

$$\frac{\left((1+r)^t - 1\right) + g_1(p)}{g_2(p)} = \begin{cases} \frac{3\left((1+r)^t-1\right)+2p+\sqrt{p(1-p)}}{2(1-p)-\sqrt{p(1-p)}+1}, & p < \frac{1}{2}, \\ \frac{3\left((1+r)^t-1\right)+p}{1-p}, & p = \frac{1}{2}, \\ \frac{3\left((1+r)^t-1\right)+2p-\sqrt{p(1-p)}+1}{2(1-p)+\sqrt{p(1-p)}}. & p > \frac{1}{2}. \end{cases} \tag{17.75}$$

And it is to be noted that in the case of a risk-less bottomry loan (i.e., $p = 0$) the lower bound on the bottomry interest rate (17.73) collapses to the regular interest rate that is demanded of any reputable borrower for a loan that is made available over a time period t:

$$c > \left((1+r)^t - 1\right), \tag{17.76}$$

which is intuitive enough.

The bottomry interest rate lower bound for the 1-sigma VaR criterion of choice (17.60) is given as

$$c > s(p, r, t) , \tag{17.77}$$

where

$$s(p, r, t) = \begin{cases} \frac{\left((1+r)^t-1\right)+p+\sqrt{p(1-p)}}{(1-p)-\sqrt{p(1-p)}}, & p \le \frac{1}{2}, \\ \infty, & p > \frac{1}{2}, \end{cases} \tag{17.78}$$

For the expected utility criterion of choice (17.63) this lower bound is given as

$$c > \omega(p, r, t) , \tag{17.79}$$

where

$$\omega(p, r, t) = \frac{\left((1 + r)^t - 1\right) + p}{1 - p}.$$

(17.80)

The adjusted odds (17.74), (17.78), and (17.80) not only take the same sigma bound information into account as the position measures from which they were derived, (17.5), (17.60), and (17.63), respectively. But they also take into account the cost of money, by way of the regular interest rate, r, and the duration of the bet, t.

17.4.2 The Merchant

The merchant, once having obtained a bottomry loan of L guilders, can choose to lend out this loan to reputable borrowers $D_{1'}$, in order to cash in on this loan via a regular interest rate, while waiting for his profits to return home. The worst-case scenario for the merchant then will be that the ship and its cargo would be lost at sea (probability of p), in which case there is a gain of only

$$x = -L + L + \left((1 + r)^t - 1\right) L = \left((1 + r)^t - 1\right) L.$$

(17.81)

The best-case scenario is that the ship and its cargo safely reach the harbor (probability of $1 - p$), in which case the merchant will gain a much larger profit of

$$x = -L + L + \left((1 + r)^t - 1\right) L + [C - (1 + c)] L$$

(17.82)

$$= \left[\left((1 + r)^t - 1\right) + (C - c - 1)\right] L.$$

So, the under the updated decision $D_{1'}$ to take out a bottomry loan, the merchant has the following outcome probability distribution, (17.81) and (17.82):

$$f(x|D_{1'}) = \begin{cases} p, & x = \left((1 + r)^t - 1\right) L, \\ 1 - p, & x = \left[\left((1 + r)^t - 1\right) + (C - c - 1)\right] L. \end{cases}$$

(17.83)

By way of Bernoullis utility function (13.4), we may transform this outcome probability distribution to a corresponding utility probability distribution:

$$f(u|D_{1'}) = \begin{cases} p, & u = q \log \frac{m + \left((1+r)^t - 1\right) L}{m}, \\ 1 - p, & u = q \log \frac{m + \left[\left((1+r)^t - 1\right) + (C - c - 1)\right] L}{m}, \end{cases}$$

(17.84)

which has the 1-sigma position (i.e., risk index), (17.5) and (17.20),

$$\text{loc}(D_{1'}) = q \log \frac{m + \left((1+r)^t - 1 \right) L}{m} \, g_1(p) \tag{17.85}$$

$$+ \, q \log \left(\frac{m + \left[\left((1+r)^t - 1 \right) + (C - c - 1) \right] L}{m} \right) g_2(p) \, .$$

It is again assumed that the merchant will take out a bottomry loan only if the utility distribution under the decision to take out the loan $D_{1'}$ is located more to the right than the utility probability distribution under the decision not to take out the loan D_2:

$$\text{loc}(D_{1'}) > \text{loc}(D_2) \, . \tag{17.86}$$

Solving the decision theoretical inequality, we find the upper bound of the bottomry interest rate c, as determined by the merchant and the going regular interest rate r, (17.51) and (17.85),

$$c < \left(C - 1 + \frac{m}{L} \right) \left[1 - \left(\frac{m - L}{m + \left((1+r)^t - 1 \right) L} \right)^{h(p)} \right] + \left((1+r)^t - 1 \right), \tag{17.87}$$

where, (17.40),

$$h(p) = \frac{g_1(p)}{g_2(p)} = \begin{cases} \frac{2p + \sqrt{p(1-p)}}{2(1-p) - \sqrt{p(1-p)} + 1}, & p < \frac{1}{2}, \\ \frac{p}{1-p}, & p = \frac{1}{2}, \\ \frac{2p - \sqrt{p(1-p)} + 1}{2(1-p) + \sqrt{p(1-p)}}, & p > \frac{1}{2}. \end{cases}$$

And it is to be noted that in the case of a risk-less bottomry loan (i.e., $p = 0$) the upper bound on the bottomry interest rate (17.87) also collapses to the regular interest rate that is demanded of any reputable borrower for a loan that is made available over a time period t:

$$c < \left((1+r)^t - 1 \right), \tag{17.88}$$

which is intuitive enough.

Moreover, we find again, as in (17.55), that the margin of profit for the money lender will evaporate, as the merchant he services becomes rich enough to become his own insurer. This phenomenon, on a conceptual level, seems to be so intuitive that it could serve as a fundamental common-sense boundary condition in the modeling of any insurance problem. But on a mathematical level this phenomenon translates to a limit which again is highly non-intuitive, in terms of its inner workings.

For it may numerically be checked that the interest upper bound (17.88) will tend to the lower bound (17.76) as the merchant gets richer; i.e.,

$$\left(C-1+\frac{m}{L}\right)\left[1-\left(\frac{m-L}{m+\left((1+r)^t-1\right)L}\right)^{h(p)}\right]+\left((1+r)^t-1\right)\to h\left(p,r,t\right),$$
(17.89)

or, equivalently, (17.40) and (17.76),

$$\left(C-1+\frac{m}{L}\right)\left[1-\left(\frac{m-L}{m+\left((1+r)^t-1\right)L}\right)^{\frac{g_1(p)}{g_2(p)}}\right]+\left((1+r)^t-1\right)$$

$$\to\frac{\left((1+r)^t-1\right)+g_1(p)}{g_2(p)}$$
(17.90)

as $L/m\to 0$.

The bottomry interest rate upper bounds for the 1-sigma VaR criterion of choice (17.60) and the expected utility criterion of choice (17.63) are given as

$$c<\left(C-1+\frac{m}{L}\right)\left[1-\left(\frac{m-L}{m+\left((1+r)^t-1\right)L}\right)^{s(p)}\right]+\left((1+r)^t-1\right),$$
(17.91)

where, (17.62),

$$s(p)=\begin{cases}\frac{p+\sqrt{p(1-p)}}{(1-p)-\sqrt{p(1-p)}}, & p\le\frac{1}{2},\\ \infty, & p>\frac{1}{2},\end{cases}$$

and

$$c<\left(C-1+\frac{m}{L}\right)\left[1-\left(\frac{m-L}{m+\left((1+r)^t-1\right)L}\right)^{\omega(p)}\right]+\left((1+r)^t-1\right),$$
(17.92)

where, (17.65),

$$\omega(p)=\frac{p}{1-p}.$$

17.4.3 A Historical Data Point

In Table 17.2, we give for return factors of $C=2,3,4$, the bottomry interest rate upper bounds for the 1-sigma VaR criterion of choice (17.60), the expected utility criterion of choice (17.63), and the balanced 1-sigma criterion of choice (17.5), i.e., (17.92), (17.91), and (17.87), respectively. And it is found that only for the balanced 1-sigma criterion of choice (17.5) the for the cost of money corrected bottomry interest rates are within the reported historical bounds of 30% to 70%, [29].

	$LB(D_i)$	$E(U\,|\,D_i)$	$\frac{1}{3}\left[LB + E(U\,|\,D_i) + LB\right]$
$C = 2$	88%	18%	35%
$C = 3$	122%	24%	48%
$C = 4$	156%	30%	60%

Table 17.2: Interest rate upper bounds for bottomry loans, as determined by the merchant and the going regular interest rate $r = 0.04$, for different return factors C, for a probability of a shipwreck of $p = 1/20$, an initial wealth of $m = 300$ guilders of which $L = 200$ are invested in merchandise that is bound for the Levant and word of which is expected to reach Holland within $t = 1$.

17.5   Discussion

In this chapter we have applied the Bayesian decision theory to a simple case study that is inspired on the sixteenth century practice of bottomry.

This case study has led us to a class of odds adjustments that may be found by way of the use of different maximization criteria and by taking into account the cost of money, by way of the regular interest rate, r, and the duration of the bet, t. Also, there have been presented some historical data points that may seem to point at the optimality of the balanced 1-sigma position measure that takes into account the most likely scenario, together with the plausible worst- and best-case scenarios.

Chapter 18

Premium Based Insurance

We now turn to the rationale of the individual to take out insurance and the rationale of the insurance company to provide insurance contracts. The example given here is a generalization of an example given by Jaynes [47].

18.1 The Position Measure

In this case study we will use a balanced position measure, in which the worst- and best-case scenarios are given the same weight in terms of their plausibility; i.e., $k_1 = k_2 = k$ in (13.25), or, equivalently, (13.23). Moreover, we will let the decision makers navigate their decision space by way of 1-sigma, i.e., realistically plausible, worst- and best-case scenarios, (13.13) and (13.16).

In the understanding that $k_1 = k_2 = 1$ in (13.25), or, equivalently, $k = 1$ in (13.23), we now will drop explicit mention of the sigma levels, as this will allow us to remove some of the notational clutter down the line.

Let the 1-sigma undershoot corrected lower confidence bound be given as

$$LB = \begin{cases} a, & lb < a, \\ E(U) - \text{std}(U), & lb \geq a, \end{cases} \tag{18.1}$$

where a is minimax value

$$a = \min(U) \tag{18.2}$$

and lb is the traditional 1-sigma lower confidence bound:

$$lb = E(U) - \text{std}(U). \tag{18.3}$$

Let the 1-sigma overshoot corrected upper confidence bound be given as

$$UB = \begin{cases} E(U) + \text{std}(U), & ub \leq b, \\ b, & ub > b, \end{cases} \tag{18.4}$$

where b is maximax value

$$b = \max(U) \tag{18.5}$$

and ub is the traditional 1-sigma upper confidence bound:

$$ub = E(U) + \text{std}(U). \tag{18.6}$$

Then the position measure which is to be maximized is given as, (13.23),

$$\text{loc} = \frac{LB + E(U) + UB}{3}, \tag{18.7}$$

where, because of the under- and overshoot corrections in (18.1) and (18.4),

$$\frac{LB + E(U) + UB}{3} = \begin{cases} E(U), & lb \geq a, \ ub \leq b, \\ \frac{1}{3}\left[a + 2E(U) + \text{std}(U)\right], & lb < a, \ ub \leq b, \\ \frac{1}{3}\left[2E(U) - \text{std}(U) + b\right], & lb \geq a, \ ub > b, \\ \frac{1}{3}\left[a + E(U) + b\right], & lb < a, \ ub > b. \end{cases} \tag{18.8}$$

The balanced 1-sigma criterion of choice (18.7) is the position measure that takes into account the most likely scenario, together with the plausible worst- and best-case scenarios.

18.2 The Insurance Case

Let P be the premium for some proposed insurance contract between one individual customer and an insurance company.

The insurance contract covers n contingencies, where the ith contingency C_i has a probability of p_i of occurring and a cost associated with it of L_i. We assume that the insurance company and their customers make the same probability and costs assessments for the n contingencies.

It is also assumed that the probabilities of the contingencies covered are independent; i.e., the knowledge of a burglary occurring does not modify the probability of a car accident occurring, and vice versa. Moreover, for notational simplicity and computational tractability, we let the probabilities for the contingencies as well as their associated costs be equal:

$$p = p_1 = \cdots = p_n \quad \& \quad L = L_1 = \cdots = L_n.$$

18.3 The Insurance Company

We will first take a look at the minimal premium P an insurance company must set for each of its N contracts, in order to be rational.

The insurance company has two decisions it can choose from, either it offers N insurance contracts, each of which covers n contingencies, or it does not:

$$D_1 = \text{offer } N \text{ insurance contracts},$$
$$D_2 = \text{offer no insurance contracts}.$$

If the insurance company has N outstanding contracts, each contract covering n contingencies with probability p and a payout of L for each contingency, then a total of Nn separate contingencies are covered each having probability p. So, the probability of s contingencies occurring in conjunction for over all N customers follow the binomial distribution:

$$p(s|\, D_j) = \binom{Nn}{s} p^s \, (1-p)^{Nn-s}\,, \tag{18.9}$$

for $s = 0, 1, \ldots, Nn$ and for $k = 1, 2$, as the decisions D_j will not modulate the probabilities of the number of contingencies occurring in conjunction.

18.3.1   The Utility Probability Distributions

Let the insurance company have an initial wealth of M. If one customer pays an insurance premium P, then n contingencies can occur in conjunction. And if N customers pay an insurance premium P, then Nn contingencies can occur in conjunction.

Under the decision to provide N insurance contracts D_1, the increment of the wealth of the insurance company X_s for a given outcome s is given as

$$X_s = NP - sL, \tag{18.10}$$

for $s = 0, 1, \ldots, Nn$. By way of (13.36) and (18.10), we may construct the following conditional utility distribution, used for mapping outcomes to utilities:

$$p(u|\, s, E_1) = \begin{cases} 1, & u = q\log \frac{M+NP-sL}{M} \\ 0, & u \neq q\log \frac{M+NP-sL}{M} \end{cases} \tag{18.11}$$

or, equivalently, (13.37),

$$p(u|\, s, D_1) = \delta\left(u - q\log \frac{M + NP - sL}{M}\right), \tag{18.12}$$

where δ is the delta-function (13.38).

By way of the product rule (4.1), we may combine (18.9) and (18.12) in order to obtain the bivariate distribution of the utility u and the outcome s:

$$p(u, s|\, D_1) = p(u|\, s, D_1)\, p(s|\, D_1)$$

$$\tag{18.13}$$

$$= \delta\left(u - q\log \frac{M + NP - sL}{M}\right) \binom{Nn}{s} p^s \, (1-p)^{Nn-s}.$$

Marginalizing over the outcomes s by way of the generalized sum rule (4.4), we find the utility probability distribution

$$p(u|\, D_1) = \sum_{s=0}^{Nn} \delta\left(u - \log \frac{M + NP - sL}{M}\right) \binom{Nn}{s} p^s \, (1-p)^{Nn-s}. \tag{18.14}$$

In order to get a more intuitive feel for (18.14), we observe that (18.11) is an one-on-one mapping. So, we may make a change of variable by interchanging the label s by its corresponding utility value. This then allows us to write (18.14) in the alternative form:

$$P\left(u = q\log \frac{M + NP - sL}{M} \,\middle|\, D_1 \right) = \binom{Nn}{s} p^s \,(1-p)^{Nn-s}. \qquad (18.15)$$

It may be found, however, that after one gets used to the delta-function notation, it will be notation (18.15) which becomes awkward to the eye, rather than (18.14).

If the insurance company does not provide any insurance contracts D_2, then for each outcome s the initial wealth M of the insurance company remains as it was:

$$p(\,u|\,s, D_2) = \delta\left(u - q\log \frac{M}{M} \right) = \delta(u)\,, \qquad (18.16)$$

By way of the product and generalized sum rule, we then find[1]:

$$p(\,u|\,D_2) = \delta(u) \sum_{s=0}^{Nn} \binom{n}{s} p^s \,(1-p)^{Nn-s} = \delta(u)\,, \qquad (18.17)$$

or, equivalently, a probability of one of neither a loss nor a gain:

$$P(\,u = 0|\,D_2) = 1. \qquad (18.18)$$

18.3.2   The Loci

The utility probability distributions for decisions D_1 and D_2 are (18.14) and (18.17), respectively. We now proceed to compute the balanced 1-sigma loci of these utility probability distributions.

The log-function allows the following expansion

$$\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots. \qquad (18.19)$$

Furthermore, the utilities for decision D_1 in (18.14) may be rewritten as

$$u = q\log\frac{M + NP - sL}{M} = q\log\left(1 + \frac{NP - sL}{M} \right). \qquad (18.20)$$

Since insurance companies, typically, have asset positions M in the hundreds of millions, far exceeding the collective premiums and the (reasonably probable[2]) contingency costs over N insurance contracts, we make the simplifying assumption that

$$\left(\frac{NP - sL}{M} \right)^c \to 0, \qquad (18.21)$$

---

[1]Compare with (18.13) and (18.14).
[2]For example, the probability of all Nn contingencies occurring all at once is p^{Nn}, (18.17), which is highly improbable (i.e., tends to zero) for large N and small p.

for $c \geq 2$. By way of (18.19) and (18.21), we find that

$$q \log \left(1 + \frac{P - sL}{M} \right) \approx q \, \frac{NP - sL}{M}, \tag{18.22}$$

and we say that the insurance company has such a large initial amounts of money M, that relative to any (reasonably probable) gain and loss, its utility for money is linear. This approximation simplifies the probability distribution of the utility (18.14) to

$$p(\, u|\, D_1) = \sum_{s=0}^{Nn} \delta \left(u - q \, \frac{NP - sL}{M} \right) \binom{Nn}{s} p^s \, (1-p)^{Nn-s}. \tag{18.23}$$

The integral of a sum is equivalent to a sum of integrals. So, using the delta-function property (13.38), the expected value of u evaluates as

$$E(\, u|\, D_1) = \int u \, p(\, u|\, D_1) \, du$$

$$= \sum_{s=0}^{Nn} \left[\int u \, \delta \left(u - q \, \frac{NP - sL}{M} \right) du \right] \binom{Nn}{s} p^s \, (1-p)^{Nn-s} \tag{18.24}$$

$$= \sum_{s=0}^{Nn} q \, \frac{NP - sL}{M} \binom{Nn}{s} p^s \, (1-p)^{Nn-s}$$

$$= E \left(q \, \frac{NP - sL}{M} \right).$$

In analogy with (18.24), we also may find that

$$\mathrm{var}(\, u|\, D_1) = \mathrm{var} \left(q \, \frac{NP - sL}{M} \right). \tag{18.25}$$

For any variable X and constants a, b, we have that, [74]:

$$E(a + bX) = a + b \, E(X),$$

$$\tag{18.26}$$

$$\mathrm{var}(a + bX) = b^2 \, \mathrm{var}(X).$$

So, using the fact that s is a variable with a binomial distribution, having an expected value of

$$E(s) = (Nn)\, p, \tag{18.27}$$

we may evaluate (18.24), by way of (18.26), directly as

$$E(\, u|\, D_1) = q \, \frac{NP - E(s)\, L}{M} = q \, \frac{NP - (Nn)\, pL}{M}. \tag{18.28}$$

Likewise, using the fact that s is a variable having variance

$$\text{var}(s) = (Nn)\,p\,(1-p),\tag{18.29}$$

we may evaluate (18.25), by way of (18.26), directly as

$$\text{var}(u|\,D_1) = \left(q\,\frac{L}{M}\right)^2 \text{var}(i) = q^2\,\frac{(Nn)\,p\,(1-p)\,L^2}{M^2},\tag{18.30}$$

or, equivalently,

$$\text{std}(u|\,D_1) = q\,\frac{\sqrt{(Nn)\,p\,(1-p)}L}{M},\tag{18.31}$$

as the costs of a given contingency L and the asset position of the insurance company M are both positive.

In order to compute the balanced 1-sigma locus (18.7) of the utility distribution (18.14), we first need to determine whether or not lower bound undershoot and whether or not upper bound overshoot occurs. That is, we need to check whether the probabilistic worst-case scenario (18.1) corresponds with a minimax utility (18.2) or a traditional, uncorrected 1-sigma lower confidence bound (18.3) and whether the probabilistic best-case scenario (18.4) corresponds with a maximax utility (18.5) or a traditional, uncorrected 1-sigma upper confidence bound (18.6).

The traditional 1-sigma lower confidence bound is given as (18.3)

$$lb = E(u|\,D_1) - \text{std}(u|\,D_1).\tag{18.32}$$

Substituting (18.28) and (18.31) into (18.3), we can compute the 1-sigma utility lower bound under decision D_1 to be

$$E(u|\,D_1) - \text{std}(u|\,D_1) = q\,\frac{NP - (Nn)\,pL}{M} - q\,\frac{\sqrt{np\,(1-p)}L}{M}.\tag{18.33}$$

And from (18.20) and (18.22) it is found that the minimax utility (18.2) equals

$$a = q\,\frac{NP - (Nn)\,L}{M},\tag{18.34}$$

as the insurance company stands to have to pay the most to the insurance holders should all n contingencies occur in a given year for all N contracts.

If we solve the inequality, (18.33) and (18.34),

$$E(u|\,D_1) - \text{std}(u|\,D_1) < a,\tag{18.35}$$

or, equivalently, as the Weber constant q cancels out,

$$\frac{NP - (Nn)\,pL}{M} - \frac{\sqrt{(Nn)\,p\,(1-p)}L}{M} < \frac{NP - (Nn)\,L}{M},\tag{18.36}$$

for the probability p, then we find that lower bound undershoot, for a given number of contingencies Nn and a given 1-sigma lower bound, will only occur for contingency probabilities

$$p > \frac{Nn}{1 + Nn}. \tag{18.37}$$

So, it follows that, (18.1),

$$LB = \begin{cases} a, & p > \frac{Nn}{1+Nn}, \\ E(u|\, D_1) - \text{std}(u|\, D_1), & p \leq \frac{Nn}{1+Nn}. \end{cases} \tag{18.38}$$

Since probabilities of insurable contingencies are quite small, it can be assumed that lower bound undershoot will not occur, not even if there is only one covered contingency.

The traditional 1-sigma upper confidence bound is given as (18.6)

$$ub = E(u|\, D_1) + \text{std}(u|\, D_1). \tag{18.39}$$

Substituting (18.28) and (18.31) into (18.6), we can compute the 1-sigma utility upper bound under decision D_1 to be

$$E(u|\, D_1) + \text{std}(u|\, D_1) = q\,\frac{NP - (Nn)\,pL}{M} + q\,\frac{\sqrt{(Nn)\,p\,(1-p)}L}{M}. \tag{18.40}$$

It follows from (18.20) and (18.22) that the maximax utility (18.5) is

$$b = q\,\frac{NP}{M}, \tag{18.41}$$

as the insurance company only stands to gain a total collective premium of NP should no contingencies occur in a given year for any of the N contracts. If we solve the inequality, (18.40) and (18.41),

$$E(u|\, D_1) + \text{std}(u|\, D_1) > b, \tag{18.42}$$

or, equivalently, as the Weber constant q cancels out,

$$\frac{NP - (Nn)\,pL}{M} + \frac{\sqrt{(Nn)\,p\,(1-p)}L}{M} > \frac{NP}{M}, \tag{18.43}$$

for the probability p, then we find that upper bound overshoot, for a given number of contingencies Nn and a given k_2-sigma upper bound, occurs for contingency probabilities

$$p < \frac{1}{1 + Nn}. \tag{18.44}$$

So, it follows that, (18.4),

$$UB = \begin{cases} E(u|\, D_1) + \text{std}(u|\, D_1), & p \geq \frac{1}{1+Nn}, \\ b, & p < \frac{1}{1+Nn}. \end{cases} \tag{18.45}$$

For a large number of insurance contracts N, it can be assumed that upper bound overshoot will not occur, even if the probabilities of the insurable contingencies are quite small.

It follows from (18.38), (18.45), and the fact that for large Nn the probability of a contingency will generally adhere to the inequality

$$\frac{1}{1 + Nn} < p < \frac{Nn}{1 + Nn},\tag{18.46}$$

that the balanced 1-sigma locus (18.7) of the insurance company reduces to a traditional expected utility value, (18.28) and (18.33):

$$\mathrm{loc}(D_1) = \frac{LB + E(U \mid D_1) + UB}{3}$$

$$= E(U \mid D_1)\tag{18.47}$$

$$= q\,\frac{NP - (Nn)\,pL}{M}.$$

And it is to be noted this is the first time in this thesis that we have an actual example of a case in which Weaver's criterion of choice (18.8) collapses to a traditional expected value criterion of choice.

Let D_2 be the decision to not provide any insurances, then

$$\mathrm{loc}(D_2) = 0.\tag{18.48}$$

seeing that under decision D_2 the whole utility distribution is concentrated at $u = 0$, (18.17).

18.3.3   A Premium Lower Bound

It is assumed in this thesis that rational agents will base their decisions on a comparison of the position of the utility probability distributions. So, it is expected that the insurance company will be certain to offer the insurance contracts if the utility probability distribution decision D_1 is located more to the right than the utility probability distribution under decision D_2, that is, when it is expected to be more profitable to sell insurances rather than not.

If we solve the the decision inequality

$$\mathrm{loc}(D_1) > \mathrm{loc}(D_2),\tag{18.49}$$

or, equivalently, (18.47) and (18.48),

$$q\,\frac{NP - (Nn)\,pL}{M} > 0,\tag{18.50}$$

then we find the lower bound of N premiums

$$NP > (Nn)\,pL,\tag{18.51}$$

which is the expected value of the monetary damages $E(sL)$ in N insurance contracts, (18.26) and (18.27).

Stated differently, the position of the utility probability distribution under decision D_1 (i.e., provide N insurance contracts) exceeds the position of the utility probability distribution under decision D_2 (i.e., provide no insurance contracts) whenever the collective premium NP exceeds the expected value of the total monetary damages outcome probability distribution. A finding which is intuitive enough.

If both sides of (18.59) are divided by the number of insurance contracts N, then we obtain the individual premium lower bound

$$P > npL, \tag{18.52}$$

or, equivalently,

$$P > E(iL). \tag{18.53}$$

where

$$E(iL) = npL \tag{18.54}$$

and

$$\mathrm{std}(iL) = \sqrt{np\,(1-p)}\,L \tag{18.55}$$

are the expected value and a standard deviation of the monetary damages outcome probability distribution for one single customer:

$$p(iL) = \binom{n}{i} p^i\,(1-p)^{n-i}. \tag{18.56}$$

In other words, the premium lower bound on a single contract, as determined by the insurance company, is the expected value of the monetary damages for one single customer.

It is to be noted that in this analysis the operating costs of the insurance company have been neglected, as well as the cost of money.

18.3.4 Spreading the Risks

The insurance company has N contracts over which it can spread its risks. This allows the insurance company to set its small premium lower bound of (18.53).

In order to better illustrate the profound effect of this decision theoretical law of large numbers, we now will give the premium lower bound for the special case wherein the insurance company, like the money lender from the previous chapter, just holds the one insurance contract.

It may be checked that for an insurance company that only has the one insurance contract (i.e., $N = 1$) and where the probability p on the occurrence of a contingency adheres to the inequality, (18.46),

$$p < \frac{1}{1+n} < \frac{n}{1+n}, \tag{18.57}$$

that the balanced 1-sigma locus (18.7) of the insurance company goes to, (18.28), (18.33), and (18.41):

$$\text{loc}(D_1) = \frac{LB + E(U \mid D_1) + UB}{3}$$

$$= \frac{2\,E(U \mid D_1) - \text{std}(U \mid D_1) + b}{3} \tag{18.58}$$

$$= q\left[\frac{P}{M} - \frac{1}{3}\left(\frac{2\,npL}{M} + \frac{\sqrt{np\,(1-p)}\,L}{M}\right)\right],$$

as opposed to the traditional expected value of (18.47).

Substituting (18.121) into (18.49), we then obtain the premium lower bound for an insurance company that only has the one customer:

$$P > \frac{2\,npL + \sqrt{np\,(1-p)}L}{3}, \tag{18.59}$$

or, equivalently, (18.54) and (18.55),

$$P > \frac{2\,E(iL) + \text{std}(iL)}{3}. \tag{18.60}$$

If we compare (18.53) with (18.60), then we see that having N identical and independent insurance contracts over which to spread the risks allows the insurance company to bring down its premium lower bound, or, equivalently, the minimal premium needed to cover the incurred risks, with an amount of

$$\frac{2\,E(iL) + \text{std}(iL)}{3} - E(iL) = \frac{\text{std}(iL) - E(iL)}{3}. \tag{18.61}$$

So, the decision theoretical law of large numbers tells us that there is strength in numbers. For it allows the members of a corporation, with a large enough membership of N, to divide up the risks between themselves, thus, reducing the risks they incur by an amount of (18.61).

18.4   The Customer

We now take a look at the maximum premium P the customer is willing to pay for his insurance contract.

There are two decisions the customer can choose from:

$$D_1 = \text{buy insurance},$$
$$D_2 = \text{do not buy insurance}.$$

The probability of i of the n contingencies occurring in conjunction for a single customer is

$$p(i|D_j) = \binom{n}{i} p^i (1-p)^{n-i}, \tag{18.62}$$

for $i = 0, 1, \ldots, n$ and $j = 1, 2$, as the decisions D_j will not modulate the probabilities of the number of contingencies occurring.

18.4.1   The Utility Probability Distributions

Let the customer have an initial amount of money m. If the customer buys the insurance, D_1, then for any number of contingencies i the monetary outcome will always be the same. The customer pays the premium P and now has an updated amount of wealth $m - P$, and whatever the number of contingencies, his damages are always refunded to the level $m - P$.

Stated differently, if the customer does buy the insurance, then

$$p(u|i, D_1) = \delta\left(u - q\log\frac{m-P}{m}\right). \tag{18.63}$$

By way of the product rule (4.1), we have that, (18.62) and (18.63),

$$p(u, i|D_1) = p(i|D_j)\, p(u|i, D_1) = \delta\left(u - q\log\frac{m-P}{m}\right)\binom{n}{i} p^i (1-p)^{n-i}. \tag{18.64}$$

It then follows from (18.64) and the generalized sum rule (4.4) that

$$p(u|D_1) = \sum_{i=0}^{n} p(u, i|D_1) = \delta\left(u - q\log\frac{m-P}{m}\right), \tag{18.65}$$

which is equivalent to the statement that

$$P\left(u = \log\frac{m-P}{m}\,\middle|\,D_1\right) = 1.$$

Now, if the customer decides not to buy insurance, D_2, then for a given number of contingencies i the monetary damage is iL, which corresponds with a decrement in wealth of

$$x_i = -iL. \tag{18.66}$$

By way of (13.36) and (18.66), we may construct the following conditional utility distribution, used for mapping outcomes to utilities:

$$p(u|i, D_2) = \begin{cases} 1, & u = q\log\frac{m-iL}{m} \\ 0, & u \neq q\log\frac{m-iL}{m} \end{cases} \tag{18.67}$$

or, equivalently, (13.37),

$$p(u|i, D_2) = \delta\left(u - q\log\frac{m-iL}{m}\right). \tag{18.68}$$

By way of (18.62), (18.68), and the product rule (4.1), we have

$$p(u, i|\, D_2) = \delta\left(u - q\log\frac{m - iL}{m}\right)\binom{n}{i} p^i\, (1 - p)^{n-i}\,. \qquad (18.69)$$

By way (18.69) and the generalized sum rule (4.4), we then have

$$p(u|\, D_2) = \sum_{i=0}^{n} \delta\left(u - q\log\frac{m - iL}{m}\right)\binom{n}{i} p^i\, (1 - p)^{n-i}\,, \qquad (18.70)$$

which is equivalent to the statement that

$$P\left(u = q\log\frac{m - iL}{m}\,\middle|\, D_2\right) = \binom{n}{i} p^i\, (1 - p)^{n-i}\,.$$

18.4.2   The Loci

The utility probability distributions for decisions D_1 and D_2 are (18.65) and (18.70), respectively. We now proceed to compute the balanced 1-sigma loci of these utility probability distributions.

We define the taking out of an insurance to be a hedge to mitigate some worst-case scenario. The mitigating action is to take out an insurance D_1, which has a balanced locus of

$$\operatorname{loc}(D_1) = q\log\frac{m - P}{m}\,, \qquad (18.71)$$

seeing that under decision D_1 the whole utility distribution is concentrated at the certain utility $u = q\log\left[(m - P)/m\right]$, (18.65).

The scenarios we wish to mitigate against under the decision not take out an insurance D_2 is the occurrence of one or more contingencies in a given year. The first two moments and the standard deviation of (18.70) are given as

$$E(u|\, D_2) = \int u\, p(u|\, D_2)\, du$$

$$\qquad\qquad\qquad (18.72)$$

$$= \sum_{i=0}^{n} q\log\frac{m - iL}{m}\binom{n}{i} p^i\, (1 - p)^{n-i}\,,$$

and

$$E\left(u^2\,\middle|\, D_2\right) = \int u^2 p(u|\, D_2)\, du$$

$$\qquad\qquad\qquad (18.73)$$

$$= \sum_{i=0}^{n} \left(q\log\frac{m - iL}{m}\right)^2\binom{n}{i} p^i\, (1 - p)^{n-i}\,,$$

and

$$\text{std}(U|D_2) = \sqrt{E(u^2|D_2) - [E(u|D_2)]^2}. \tag{18.74}$$

In order to compute the balanced 1-sigma locus (18.7) of the utility distribution (18.70), we first need to determine whether or not lower bound undershoot and whether or not upper bound overshoot occurs. That is, we need to check whether the probabilistic worst-case scenario (18.1) corresponds with a minimax utility (18.2) or a traditional 1-sigma lower confidence bound (18.3) and whether the probabilistic best-case scenario (18.4) corresponds with a maximax utility (18.5) or a traditional 1-sigma upper confidence bound (18.6).

For a customer wealth m that tends to infinity, we have that (18.19), (18.20), and (18.21),

$$q \log\left(1 - \frac{iL}{m}\right) \to -q\,\frac{iL}{m}. \tag{18.75}$$

So as $m \to \infty$, we have that (18.26), (18.27), (18.72), and (18.75),

$$E(u|D_2) \to -q\,\frac{npL}{m} \tag{18.76}$$

and (18.26), (18.29), (18.31), (18.74), and (18.75),

$$\text{std}(u|D_2) \to q\,\frac{\sqrt{np(1-p)}L}{m}. \tag{18.77}$$

By way of (18.32), (18.76) and (18.77), we find that the 1-sigma utility lower bound under decision D_2 tends to

$$E(u|D_2) - \text{std}(u|D_2) \to -q\,\frac{npL}{m} - q\,\frac{\sqrt{np(1-p)}L}{m}, \tag{18.78}$$

as the customer wealth m tends to infinity. It follows from (18.75) that the minimax utility tends to

$$a = q\,\log\frac{m-nL}{m} \to -q\,\frac{nL}{m}, \tag{18.79}$$

as the customer stands to have to have the largest damages should all n contingencies occur in a given year. If we solve the limit of the inequality

$$E(u|D_2) - \text{std}(u|D_2) < a, \tag{18.80}$$

or, equivalently, (18.78) and (18.79), as the Weber constant q cancels out,

$$-\frac{npL}{m} - \frac{\sqrt{np(1-p)}L}{m} < -\frac{nL}{m}, \tag{18.81}$$

for the probability p, then we find that lower bound undershoot, for a given number of contingencies n and a 1-sigma lower bound, will only occur for contingency probabilities

$$p > \frac{n}{1+n}, \tag{18.82}$$

as m tends to infinity. Moreover, as we let the initial wealth tend, in the reverse direction, to the maximum possible damage $m \to nL$, then inequality (18.82) will morph into

$$p > \frac{c\,n}{1+n}. \tag{18.83}$$

So, it follows that, (18.1),

$$LB = \begin{cases} a, & p > \frac{c\,n}{1+n}, \\ E(\,u|\,D_2) - \mathrm{std}(\,u|\,D_2), & p \le \frac{c\,n}{1+n}, \end{cases} \tag{18.84}$$

where c is some constant greater than one, as the utility probability distribution $p(\,u|\,D_2)$ becomes ever more left-skewed, as the monetary damages translate to ever greater disutilities. So, as the probabilities of insurable contingencies are quite small, it is predicted that lower bound undershoot will not occur, not even if we have only one covered contingency $n = 1$.

Also, by way of (18.39), (18.76) and (18.77), we find that the 1-sigma utility upper bound under event D_2 tends to

$$E(\,u|\,D_2) + \mathrm{std}(\,u|\,D_2) \to -q\,\frac{npL}{m} + q\,\frac{\sqrt{np\,(1-p)}L}{m}, \tag{18.85}$$

as the customer wealth m tends to infinity. The maximax utility is equal to, (18.68),

$$b = q \log \frac{m}{m} = 0, \tag{18.86}$$

as the customer need not make good on any monetary damages should no contingencies occur in a given year. If we solve the limit of the inequality

$$E(\,u|\,D_2) + \mathrm{std}(\,u|\,D_2) > b, \tag{18.87}$$

or, equivalently, (18.85) and (18.86), where the Weber constant q cancels out,

$$-\frac{npL}{m} + \frac{\sqrt{np\,(1-p)}L}{m} > 0, \tag{18.88}$$

for the probability p, then we find that upper bound overshoot, for a given number of contingencies n and a 1-sigma upper bound, occurs for contingency probabilities

$$p < \frac{1}{1+n}, \tag{18.89}$$

as m tends to infinity. Moreover, as we let the initial wealth tend, in the reverse direction, to the maximum possible damage $m \to nL$, then inequality (18.89) will morph into

$$p < \frac{c}{1+n}. \tag{18.90}$$

So, it follows that, (18.4),

$$UB = \begin{cases} E(\,u|\,D_2) + \mathrm{std}(\,u|\,D_2), & p \ge \frac{c}{1+n}, \\ b, & p < \frac{c}{1+n}, \end{cases} \tag{18.91}$$

where c is some constant greater than one, as the utility probability distribution $p(u|D_2)$ becomes ever more left-skewed, as the as the monetary damages translate to ever greater disutilities. So, as the probabilities of insurable contingencies are quite small, it is predicted that upper bound overshoot is almost certain to occur, even for an 1-sigma upper bound level.

From (18.83), (18.90), and the fact that the probabilities p of the n insurable contingencies are quite small, or equivalently,

$$p < \frac{c}{1+n} \leq \frac{c\,n}{1+n}, \qquad \text{for } c > 1, \tag{18.92}$$

it follows that the locus (18.7) of the utility probability distribution (18.70) reduces to, (18.84), (18.86), (18.91), and (18.92),

$$\text{loc}(D_2) = \frac{LB + E(U|D_2) + UB}{3}$$

$$= \frac{2\,E(U|D_2) - \text{std}\,(U|D_2)}{3}. \tag{18.93}$$

18.4.3 A Premium Upper Bound

It is assumed in this thesis that rational agents will base their decisions on a comparison of the position of the utility probability distributions. So, it is expected that the customer will decline to take the insurance contract if the utility probability distribution under decision D_1 is located more to the right than the utility probability distribution under decision D_2, that is, when it is expected to be more profitable to decline an insurance rather than not.

If the locus (18.71) is greater than the locus (18.93),

$$\text{loc}(D_1) > \text{loc}(D_2), \tag{18.94}$$

or, equivalently,

$$q\log\frac{m-P}{m} > \frac{2E(u|D_2) - \text{std}(u|D_2)}{3}, \tag{18.95}$$

then we can say that the outcome under D_1 will tend to be better then the outcome under D_2, that is, D_1 will tend to mitigate the outcome under D_2. Taking the exponent of both sides and doing some simple algebra then gives us the upper bound for the premium P, as dictated by the customer:

$$P < m\left\{1 - \exp\left[\frac{2E(u|D_2) - \text{std}(u|D_2)}{3\,q}\right]\right\}. \tag{18.96}$$

For a given insurance problem, the inequality (18.96) can simply be evaluated numerically by computing the mean and standard deviation of (18.70) and substituting the corresponding values into (18.96).

18.5 The Margin of Profit on a Single Insurance Contract

The premium P of an insurance contract gets its lower bound from the provider of the contract (18.53) and its upper bound from the customer of that contract (18.96):

$$E(iL) < P < m \left\{ 1 - \exp \left[\frac{2E(u|\,D_2) - \mathrm{std}(\,u|\,D_2)}{3\,q} \right] \right\}, \qquad (18.97)$$

where it is understood that both the insurance company and the customer share the same 1-sigma level for their respective probabilistic worst- and best-case scenarios.

The Margin Of Profit (MOP) for the insurance company is the premium P minus the lower bound (18.53) that the insurance company needs to set in order to cover its risks, (18.54),

$$\mathrm{MOP} = P - E(iL). \qquad (18.98)$$

It follows that the MOP will lie in the range, (18.97) and (18.98),

$$0 < \mathrm{MOP} < m \left\{ 1 - \exp \left[\frac{2E(u|\,D_2) - \mathrm{std}(\,u|\,D_2)}{3\,q} \right] \right\} - E(iL). \qquad (18.99)$$

We now will look at how reference asset position m of the customer influences the potential margin of profit for the insurance company.

18.5.1 Extremely Wealthy Customers

If the wealth of the customer tends to infinity, then the inequality (18.94) will tend to, (18.19), (18.20), (18.21), (18.71), and (18.76),

$$-q\,\frac{P}{m} > \frac{1}{3} \left[-2\,q\,\frac{npL}{m} - q\,\frac{\sqrt{np\,(1-p)}L}{m} \right], \qquad (18.100)$$

which solves for the premium upper bound

$$P < \frac{2\,npL + \sqrt{np\,(1-p)}L}{3}, \qquad (18.101)$$

or, equivalently, (18.54) and (18.55),

$$P < \frac{2\,E(iL) + \mathrm{std}(iL)}{3}. \qquad (18.102)$$

So, for the very wealthy customer the premium range (18.97) will tend to, (18.102),

$$E(iL) < P < \frac{2\,E(iL) + \mathrm{std}(iL)}{3}. \qquad (18.103)$$

And it follows that the MOP for the very rich customer will lie in the range, (18.98),

$$0 < \text{MOP} < \frac{\text{std}(iL) - E(iL)}{3}. \tag{18.104}$$

For the insurance contract that cover $n = 4$ contingencies, each contingency having a probability of $p = 10^{-4}$ of occurring, and a maximum payout for each contingency of $L = 20{,}000$ euros, we have,(18.54) and (18.55),

$$E(iL) = npL = 8.00 \text{ euro}. \tag{18.105}$$

and

$$\text{std}(iL) = \sqrt{np\,(1-p)}L = 399.98 \text{ euro}, \tag{18.106}$$

from which there follows a margin of profit upper bound of, (18.104),

$$\text{MOP} < 130.66 \text{ euro}. \tag{18.107}$$

And it follows that the insurance company can still make a good profit on its very rich customers.

Now, the reason that the insurance company can make profits off of its rich customers is because the insurance company has N contracts over which it can spread its risks. This help from the law of large numbers allows the insurance company to set its small premium lower bound of (18.53), as opposed to the much higher premium lower bound of (18.60).

In closing, if the insurance company only has one customer, like the money lender in the previous chapter, and if the wealth of that customer tends to infinity, then the premium range (18.97) will tend to, (18.54), (18.55), (18.60) and (18.102),

$$\frac{2\,E(iL) + \text{std}(iL)}{3} < P < \frac{2\,E(iL) + \text{std}(iL)}{3}, \tag{18.108}$$

in which case no profit is to be made by the insurance company, as upper bound of the margin of profit becomes zero, (18.98):

$$\text{MOP} < 0. \tag{18.109}$$

18.5.2 Moderately Wealthy Customers

We now assume customers with an initial amount of money $m = 100{,}000$ euro, as we again assume insurance contracts that cover $n = 4$ contingencies, each contingency of those having a probability of $p = 10^{-4}$ of occurring and a maximum payout for each contingency of $L = 20{,}000$ euro.

We have that

$$E(u|\,D_2) = \sum_{i=0}^{n} q \log \frac{m - iL}{m} \binom{n}{i} p^i\,(1-p)^{n-i}$$

$$\tag{18.110}$$

$$= -q\left(8.9261 \times 10^{-5}\right) \text{ utile}$$

and

$$E\left(u^2\middle|\, D_2\right) = \sum_{i=0}^{n} \left(q \log \frac{m - iL}{m}\right)^2 \binom{n}{i} p^i \left(1 - p\right)^{n-i}$$

(18.111)

$$= q^2 \left(1.9927 \times 10^{-5}\right) \text{ utile-square,}$$

from which it follows

$$\operatorname{std}\left(u\middle|\, D_2\right) = \sqrt{E\left(u^2\middle|\, D_2\right) - \left[E\left(u\middle|\, D_2\right)\right]^2}$$

(18.112)

$$= q \left(4.4631 \times 10^{-3}\right) \text{ utile.}$$

So, the upper bound of the premium, as determined by the customer who has a wealth of $m = 100{,}000$ euro, is set to, (18.96), (18.110), and (18.112),

$$m \left\{1 - \exp\left[\frac{2E\left(u\middle|\, D_2\right) - \operatorname{std}\left(u\middle|\, D_2\right)}{3\,q}\right]\right\} = 154.60 \text{ euro.} \qquad (18.113)$$

The upper bound of the margin of profit for the insurance company then becomes, (18.99), (18.105), and (18.113),

$$\text{MOP} < 154.60 - 8.00 = 146.60 \text{ euro.} \qquad (18.114)$$

18.5.3   Regular Customers

The insurance company may exact a considerably larger margin of profit from those that are not moderately rich, as for those less wealthy there will be a very real chance of suffering financial ruin, were all the n contingencies occur at once.

Stated differently, for the less wealthy customers there will be a higher willingness to buy insurance, as expressed by a higher upper bound for the premium (18.96) and, consequent, higher margins of profit for the insurance companies.

In order demonstrate the effect of the potential threat of financial ruin, we now assume an initial amount of money of only $m = 40{,}000$ for the same insurance contract as before; i.e., the insurance contract covers $n = 4$ contingencies, each having a probability of $p = 10^{-4}$ and a maximum payout for each contingency of $L = 20.000$ euro. If the very worst-case were to happen and n separate contingencies were to occur at once, then the monetary damage nL would exceed the initial amount of wealth m of the customer.

In order to compute the locus (i.e., position) of the utility probability distribution of the customer under the decision to decline the insurance contract, we need to introduce a minimum significant amount of stimulus of, (12.12),

$$x_0 = 1000 \text{ euro.}$$

So, if the customer only has 1000 euro left of his initial wealth m, then he has, for all intents and purposes, hit rock bottom. Using this x_0, we then define the damages function x, where

$$x(i, x_0) = \begin{cases} iL, & m - iL > x_0, \\ m - x_0, & m - iL \leq x_0. \end{cases} \tag{18.115}$$

Then, in order to guard against negative log-values, we substitute the conditional utility distribution (18.68) in (18.69) and (18.70) with

$$p(u| i, D_2) = \delta \left[u - q \log \frac{m - x(i, 1000)}{m} \right]. \tag{18.116}$$

This results in a new utility probability distribution under decision D_2, of which we then compute the new expected value, standard deviation, and general 1-sigma locus (18.93):

$$E(u| D_2) = \sum_{i=0}^{n} q \log \frac{m - x(i, 1000)}{m} \binom{n}{i} p^i (1 - p)^{n-i} \tag{18.117}$$

$$= -q \left(2.7740 \times 10^{-4} \right) \text{ utile}$$

and

$$E(u^2| D_2) = \sum_{i=0}^{n} \left(q \log \frac{m - x(i, 1000)}{m} \right)^2 \binom{n}{i} p^i (1 - p)^{n-i} \tag{18.118}$$

$$= q^2 \left(1.9294 \times 10^{-4} \right) \text{ utile-square},$$

from which it follows

$$\text{std}(u| D_2) = \sqrt{E\left(u^2| D_2\right) - \left[E\left(u| D_2\right)\right]^2} \tag{18.119}$$

$$= q \left(1.389 \times 10^{-2} \right) \text{ utile}.$$

So, the upper bound of the premium, as determined by the customer who has a wealth of $m = 40{,}000$ euro, is set to, (18.96), (18.117), and (18.119),

$$m \left\{ 1 - \exp \left[\frac{2E(u| D_2) - \text{std}(u| D_2)}{3} \right] \right\} = 192.10. \tag{18.120}$$

The upper bound of the margin of profit for the insurance company then becomes, (18.99), (18.105), and (18.120),

$$\text{MOP} < 192.10 - 8.00 = 184.10. \tag{18.121}$$

Comparing (18.55) with (18.121), we see that the initial amount of wealth m has a large impact on the amount of money a customer is willing to spend on an insurance contract. The less money we have, the more willing we will be to spend money on insurance.

18.6   Discussion

It may be read in [103] that the "overweighting of small probabilities contributes to the popularity of [...] insurance." So, prospect theory seems to propose that those who buy insurance tend to overestimate the expected monetary damages, as they over-weight the probability of a contingency, whereas those who provide insurance are assumed to make more realistic lower estimates, as they do not over-weight the probability of a contingency, which allows the latter to make money by way of arbitrage.

From the perspective of the Bayesian decision theory, however, providers of insurance (i.e., insurance companies) are allowed to make the same estimates of possible monetary damages as their customers and still have a very good margin of profit, as their customers, by themselves, cannot spread their risks like the insurance company can and tend to have a markedly non-linear utility function in the region of the potential losses.

Chapter 19

An Order of Magnitude

We give here a comparison of the expected outcome theory, the expected utility theory, and the Bayesian decision theory, by way of a simple toy problem in which we look at the investment willingness to avert a high impact low probability event.

It will be demonstrated here that Weaver's criterion of choice, in which the mean of the sum of the undershoot corrected lower confidence bound, expected value, and overshoot corrected upper confidence bound of either outcome or utility probability distributions are maximized, though mathematically trivial, has non-trivial practical implications for the modeled investment willingness for high impact low probability events. For it is found that under the alternative criterion of choice of the Bayesian decision theory the investment willingness for such events may increase up to an order of magnitude.

19.1   A Simple Scenario

We now apply our Bayesian framework to a scenario in which a decision maker must decide on how it is willing to invest in a further improvement of flood defenses. The two decisions under consideration in our simple scenario are

$$D_1 = \text{keep status quo,}$$
$$D_2 = \text{improve flood defenses.}$$

The possible outcomes in our risk scenario remain the same under either decision, and as such are not dependent upon the particular decision taken. These outcomes are

$$O_1 = \text{flooding,}$$
$$O_2 = \text{no flooding.}$$

The hypothetical costs associated with a flooding are

$$C = \text{costs of flooding} \tag{19.1}$$

and the investment costs associated with the additional flood defenses are expressed by the parameter

$$I = \text{investment costs.} \tag{19.2}$$

Note that if we were to do an actual cost-benefit analysis, rather than a demonstration of the here proposed decision theoretical framework, then the (opportunity) cost of money itself should also be taken into account.

The decision whether to improve the flood defenses or not is of influence on the probabilities of the respective outcomes. Under the decision to make no additional investments in flood defenses and keep the status quo, D_1, the probabilities of the outcomes will be, say,

$$
\begin{aligned}
P(O_1|\,D_1) &= \theta, \\
P(O_2|\,D_1) &= 1 - \theta.
\end{aligned}
\tag{19.3}
$$

Under the decision to improve the flood defenses, D_2, the probabilities of the flood outcomes will be decreased, leaving us with hypothetical outcome probabilities, say,

$$
\begin{aligned}
P(O_1|\,D_2) &= \phi, \\
P(O_2|\,D_2) &= 1 - \phi,
\end{aligned}
\tag{19.4}
$$

where $\phi < \theta$; that is, the proposed flood defenses will reduce the chances of a flooding by a factor $c = \theta/\phi$, where $c > 1$.

In what follows we will give the solution of this problem of choice by way of the expected outcome theory, expected utility theory, and the Bayesian decision theory.

19.2   The Expected Outcome Solution

In the expected outcome theory the expected values of the outcome probability distributions are maximized.

By way of (19.1), (19.2), (19.3), and (19.4), we may construct the outcome probability distributions under the decisions D_1 and D_2:

$$
p(x|\,D_1) = \begin{cases} \theta, & x = -C, \\ 1 - \theta, & x = 0, \end{cases}
\tag{19.5}
$$

and

$$
p(x|\,I, D_2) = \begin{cases} \phi, & x = -C - I, \\ 1 - \phi, & x = -I, \end{cases}
\tag{19.6}
$$

where we explicitly conditionalize on the investment parameter I, which is to be to estimated.

The expected outcomes of these probability distributions are, respectively [74],

$$E(X \mid D_1) = -\theta\, C \tag{19.7}$$

and

$$E(X \mid I, D_2) = -\phi\, C - I. \tag{19.8}$$

The decision theoretical equality

$$E(X \mid D_1) = E(X \mid I, D_2) \tag{19.9}$$

represents the equilibrium situation, where we will be undecided between the decision to keep the status quo D_1 and the decision to invest in additional flood defenses. Now, if we solve for the unknown I in (19.9), by way of (19.7) and (19.8):

$$I = (\theta - \phi)\, C, \tag{19.10}$$

then we find that investment where we will be undecided between both decisions.

Stated differently, any investment smaller than (19.10) will turn (19.9) into an inequality, where D_2 becomes more attractive than D_1. If we assume that we are only motivated by monetary costs, then the equilibrium investment (19.10) is the maximal investment we will be willing to make to improve our flood defenses.

The utility of a given outcome is the perceived worth of that outcome. If we take the utilities that monetary outcomes hold for us to be an incentive for our decisions, then we may perceive money to be a stimulus.

For the rich man hundred one hundred euros is an insignificant amount of money. So, the prospect of gaining or losing hundred euros will fail to move the rich man; that is, an increment of hundred euros for him has a utility which tends to zero. For the poor man one hundred euros will be a significant amount of money. So, the prospect of gaining or losing one hundred euros will most likely move the poor man to action. It follows that an increment of one hundred euros for him has a utility significantly greater than zero.

Bernoulli in 1738 derived his utility function for the subjective value of objective monies by way of a variance argument, in which he considered the subjective effect of a given fixed monetary increment c for two persons holding different initial wealths. Based on this variance argument he derived the utility function of going from an initial asset position m to the asset position $m + x$: An alternative consistency argument for the derivation of Bernoulli's utility function may be found in [106].

19.3 Bernoulli's Expected Utility Solution

In Bernoulli's expected utility theory the expected values of utility probability distributions are maximized.

Assuming that the decision maker has an initial asset position of

$$m = \text{initial asset position in euros}, \tag{19.11}$$

then, by way of Bernoulli's utility function (13.4)

$$u(x \mid m) = q \log \frac{m+x}{m} \tag{19.12}$$

where q is some scaling constant greater than zero [6], we may construct from (19.5) and (19.6) the utility probability distributions under the decisions D_1 and D_2 as

$$p(u \mid D_1) = \begin{cases} \theta, & u = q \log \frac{m-C}{m}, \\ 1-\theta, & u = q \log \frac{m}{m}, \end{cases} \tag{19.13}$$

and

$$p(u \mid I, D_2) = \begin{cases} \phi, & u = q \log \frac{m-C-I}{m}, \\ 1-\phi, & u = q \log \frac{m-I}{m}. \end{cases} \tag{19.14}$$

The expected outcomes of the utility probability distributions are, respectively [74],

$$E(U \mid D_1) = q\,\theta\,\log \frac{m-C}{m} \tag{19.15}$$

and

$$E(U \mid I, D_2) = q\,\phi\,\log \frac{m-C-I}{m-I} + q \log \frac{m-I}{m}. \tag{19.16}$$

The decision theoretical equality

$$E(U \mid D_1) = E(U \mid I, D_2) \tag{19.17}$$

represents the equilibrium situation, where we will be undecided between decisions D_1, keep the status quo, and D_2, invest in additional flood defenses. Now, if we substitute (19.15) and (19.16) into (19.17), then we obtain the closed expression for that investment value where we will be undecided between both decisions:

$$\log \frac{m-I}{m} = \theta\,\log \frac{m-C}{m} - \phi\,\log \frac{m-C-I}{m-I}. \tag{19.18}$$

Any investment smaller than the numerical solution of I in (19.18) will turn (19.18) into an inequality, where D_2 becomes more attractive than D_1. It follows that the investment equilibrium solution of (19.18) is also the maximal investment we will be willing to make to improve our flood defenses.

19.4   The Bayesian Decision Theory Solution

In this scenario we have a decision maker who must decide on how much he is willing to invest in a further improvement of flood defenses. It is expected of this decision maker that he provides the best possible flood protection; and while money is an issue, safety will be even more so.

This is why it is perceived desirable by the public at large that the decision maker, within reason, be on the side of caution and, consequently, show some moderate to strong risk aversion. Stated differently, the decision maker needs

to consider his investment willingness under probabilistic worst-case scenarios which are increasingly severe in that they have sigma levels $k_1 = 1, 2$, and 3, while refraining from risk-seeking behavior, as he keeps the weight of the probabilistic best-case scenario at a sigma level of $k_2 = 1$.

In the Bayesian decision theory the mean of the lower confidence bound, expected value, and upper confidence bound is taken as the position measure (i.e., locus) of the underlying utility probability distribution which is to be maximized (13.25):

$$\text{loc}(D_i \,|\, k_1, k_2) = \frac{LB(k_1) + E(U \,|\, D_i) + UB(k_2)}{3}, \qquad (19.19)$$

where the k_1-sigma lower confidence bound is corrected for undershoot of the worst possible outcome, (19.13) and (19.14),

$$a_i = \begin{cases} \min \left[q \log \frac{m-C}{m}, \; q \log \frac{m}{m} \right], & i = 1, \\ \min \left[q \log \frac{m-C-I}{m}, \; q \log \frac{m-I}{m} \right], & i = 2, \end{cases} \qquad (19.20)$$

giving

$$LB(k_1) = \begin{cases} a_i, & lb(k_1) < a_i, \\ lb(k_1), & lb(k_1) \geq a_i, \end{cases} \qquad (19.21)$$

where

$$lb(k_1) = E(U \,|\, D_i) - k_1 \, \text{std}(U \,|\, D_i), \qquad (19.22)$$

and the k_2-sigma upper confidence bound is corrected for overshoot of the best possible outcome, (19.13) and (19.14),

$$b_i = \begin{cases} \max \left[q \log \frac{m-C}{m}, \; q \log \frac{m}{m} \right], & i = 1, \\ \max \left[q \log \frac{m-C-I}{m}, \; q \log \frac{m-I}{m} \right], & i = 2, \end{cases} \qquad (19.23)$$

giving

$$UB(k_2) = \begin{cases} ub(k_2), & ub(k_2) \leq b_i, \\ b_i, & ub(k_2) > b_i, \end{cases} \qquad (19.24)$$

where

$$ub(k_2) = E(U \,|\, D_i) + k_2 \, \text{std}(U \,|\, D_i), \qquad (19.25)$$

Substituting (19.21) and (19.24) into (19.19), we obtain the general locus (i.e., position measure) of the utility probability distributions which allows for either a premium on cautiousness, by way of the sigma level k_1 of the lower bound, or a premium on daring, by way of the sigma level k_2 of the upper bound, or both, (19.20) and (19.25):

$$\text{loc}(D_i \,|\, k_1, k_2) = \begin{cases} E(U \,|\, D_i) + \frac{1}{3}(k_2 - k_1) \, \text{std}(U \,|\, D_i), & lb(k_1) \geq a_i, \; ub(k_2) \leq b_i, \\ \frac{1}{3}\left[a_i + 2E(U \,|\, D_i) + k_2 \, \text{std}(U \,|\, D_i) \right], & lb(k_1) < a_i, \; ub(k_2) \leq b_i, \\ \frac{1}{3}\left[2E(U \,|\, D_i) - k_1 \, \text{std}(U \,|\, D_i) + b_i \right], & lb(k_1) \geq a_i, \; ub(k_2) > b_i, \\ \frac{1}{3}\left[a_i + E(U \,|\, D_i) + b_i \right], & lb(k_1) < a_i, \; ub(k_2) > b_i, \end{cases}$$

$$(19.26)$$

where it is to be noted that for balanced sigma bounds $k_1 = k_2$ the first row will collapse to Bernoulli's expected utility proposal.

In the toy problem under consideration we have a high impact low probability scenario; that is, both large monetary costs and small probabilities for the high-impact event, or, equivalently, $x \gg 0$ and $\theta, \phi \ll 0.5$.

Stated differently, the outcome probability distributions (19.5) and (19.6) under consideration will both be highly skewed to the left and, as a consequence, will lead to the third condition in (19.26), upper confidence bound overshoot and no lower confidence bound undershoot. It follows that the operating criterion of choice will be, (19.26),

$$\text{loc}(\,D_i\,|\,k_1, k_2) = \frac{2E(U\,|\,D_i) - k_1\,\text{std}(U\,|\,D_i) + b_i}{3}. \tag{19.27}$$

The best possible outcome under decision D_1 is, (19.23),

$$b_1 = \max\left(q\log\frac{m-C}{m}, q\log\frac{m}{m} \right) = q\log\frac{m}{m} = 0, \tag{19.28}$$

and the standard deviation of (19.13) is [74]

$$\text{std}(U\,|\,D_1) = \left| q\log\frac{m-C}{m} \right| \sqrt{\theta\,(1-\theta)} = -q\,\sqrt{\theta\,(1-\theta)}\,\log\frac{m-C}{m}. \tag{19.29}$$

So from (19.15), (19.27), (19.28), and (19.29), the locus, or, equivalently, risk index under the decision to keep the status quo is

$$\text{loc}(\,D_1\,|\,k_1, k_2) = \frac{q\log\frac{m-C}{m}\left[2\theta + k_1\sqrt{\theta\,(1-\theta)}\right]}{3}. \tag{19.30}$$

The best possible outcome under decision D_2 is (19.23)

$$b_2 = \max\left(q\log\frac{m-C-I}{m}, q\log\frac{m-I}{m} \right) = q\log\frac{m-I}{m}, \tag{19.31}$$

and the standard deviation of (19.14) is [74]:

$$\text{std}(U\,|\,I, D_2) = \left| q\log\frac{m-C-I}{m-I} \right| \sqrt{\phi\,(1-\phi)} = -q\,\sqrt{\phi\,(1-\phi)}\,\log\frac{m-C-I}{m-I}. \tag{19.32}$$

So from (19.16), (19.27), (19.31), and (19.32), the locus, or, equivalently, risk index under the decision invest in additional flood defenses is

$$\text{loc}(\,D_1\,|\,I, k_1, k_2) = \frac{q\log\frac{m-C-I}{m-I}\left[2\phi + k_1\sqrt{\phi\,(1-\phi)}\right]}{3} + q\log\frac{m-I}{m}. \tag{19.33}$$

The decision theoretical equality

$$\text{loc}(\,D_1\,|\,k_1, k_2) = \text{loc}(\,D_1\,|\,I, k_1, k_2) \tag{19.34}$$

represents the equilibrium situation, where we will be undecided between decisions D_1, keep the status quo, and D_2, invest in additional flood defenses. Now, if we substitute (19.32) and (19.33) into (19.34), then we obtain the closed expression for that investment value where we will be undecided between both decisions:

$$\log \frac{m-I}{m} = \frac{1}{3}\left[\left(2\theta + k_1\sqrt{\theta(1-\theta)}\right)\log\frac{m-C}{m} - \left(2\phi + k_1\sqrt{\phi(1-\phi)}\right)\log\frac{m-C-I}{m-I}\right].$$
(19.35)

Any investment smaller than the numerical solution of I in (19.35) will turn (19.34) into an inequality, where D_2 becomes more attractive than D_1. It follows that the equilibrium investment (19.35) is also the maximal investment we will be willing to make to improve our flood defenses.

Note that the Weber constant q has fallen away in both the decision theoretical equalities (19.18) and (19.35).

This will hold in general, as both the expected values and standard deviations of the utility probability distributions (19.13) and (19.14) are linear in the unknown constant q. It follows that we may always set, without any loss of generality, q to one.

19.5 Some Numerical Results

In our simple toy problem we have a decision maker who must decide on how much he is willing to invest in a further improvement of his flood defenses.

After the great Dutch flooding the 'Oosterschelde Waterkering' was built. This was a movable dike that allowed for an improved safety from $\theta = 1/100$ to $\phi = 1/4000$, while keeping the Oosterschelde connected to the North Sea. This open connection to the North Sea was decided upon in order to keep the salt-sea ecological system of the Oosterschelde lake intact.

The total costs of the Oosterschelde Waterkering were about 2.5 billion euros. The bulk of these costs were due to the movable character of this dike. Had the Dutch government decided to build an immovable dike, then the costs would only have been about 175 million euros.

The total value of the assets at risk were about 1/20th of the GDP at the time, so that in (19.1),

$$C = 3.75 \times 10^9 \text{ euros.}$$
(19.36)

The wealth of the decision maker, that is, the Dutch government, was about 40% of the Dutch GDP at the time. Aggregated over a period of five years to account for the building time of the movable Oosterschelde dike, the relevant wealth was

$$m = 1.5 \times 10^{11} \text{ euros.}$$
(19.37)

Right after the great flood the probability in (19.3) of a catastrophic flood had been estimated to be

$$\theta = \frac{1}{100},$$
(19.38)

whereas the probability in (19.4) of a catastrophic flood under the improved flood defenses had been estimated as

$$\phi = \frac{1}{4000}. \tag{19.39}$$

Substituting the values (19.36) through (19.39) into (19.10), (19.18), and (19.35), we obtain the solutions for the maximal investments I:

- Expected outcome theory:

    - $I = 36.6 \times 10^6$ euros (any sigma level),

- Bernoulli's expected utility theory:

    - $I = 37.0 \times 10^6$ euros (any sigma level),

- Bayesian decision theory:

    - $I = 130.6 \times 10^6$ euros ($k_1 = 1$ and $k_2 = 1$),
    - $I = 236.3 \times 10^6$ euros ($k_1 = 2$ and $k_2 = 1$),
    - $I = 342.0 \times 10^6$ euros ($k_1 = 3$ and $k_2 = 1$).

We note here that after the great Dutch flood the discussion was not whether to build additional flood defenses, but, rather, whether to choose for the expensive solution which would keep the Oosterschelde salt-sea ecosystem intact over the 'cheap' solution which would not. Under the expected utility theory solution the cheap solution of an immovable dike would have been too expensive by a factor of three, whereas under the Bayesian decision theory solution with utility transformation the cheap solution was well within the cautionary 2-sigma lower bound.

We also note that the actual project was justified under neither one of the solutions. This is because the (in)tangible costs of losing the Oosterschelde salt-sea ecosystem and the (in)tangible benefits of human safety were not factored explicitly into this particular decision analysis. But the very fact that the Dutch chose to invest 2.5 billion euros in a movable Oosterschelde Waterkering, rather than opt for the cheap immovable dike solution of 175 million euros, is an important data point which shows that these additional (in)tangibles must have played an important role in the actual decision making process.

Now, should we want to do such an extended decision analysis, then we have to differentiate decision D_2 into the decisions $D_{2'}$, "improve flood defenses by way of an immovable dike," and $D_{3'}$, "improve flood defenses by way of a movable dike." After which we have to construct an outcome probability distribution for the different costs that the loss of the Oosterschelde salt-sea ecosystem might have entailed.

Chapter 20

Discussion

The work of behavioural economists is unified by a substantial project of revision of economic theory, by replacing the *homo oeconomicus* with a psychological model that better fits the empirical data of (hypothetical) betting experiments [33]. The neo-Bernoullian decision theory, as proposed in this project, in its turn endeavours to replace the psychological model of the behavioural economists with a slightly adjusted *homo oeconomicus*. And it is our hope that we have succeeded in providing some new insights in a debate that has been going on for the past 70 years.

It may be read in Jaynes [47], that to the best of his knowledge, there are as yet no formal principles at all for assigning numerical values to utility functions, not even when the criterion is purely economic, because the utility of money remains ill-defined. In the absence of these formal principles, Jaynes' final verdict was that decision theory could not be fundamental.

The Bernoulli utility function, initially derived by Bernoulli by way of common sense first principles [6], has now been derived by way of a consistency argument. This consistency argument explains why it is that Bernoulli's utility function, both in its original Fechner-Weber law and in its alternative Stevens' power law form, has proven to be so ubiquitous and successful in the field of sensory perception research. The reason is simply that human sense perception, like human plausibility perception [47], adheres to the desideratum of consistency.

The first two algorithmic steps of the Bayesian decision theory, respectively the construction of outcome probability distributions, by way of the Bayesian probability theory, and the construction of utility probability distributions, by way of the Bernoulli utility function, allow us no freedom. To construct our outcome and utility probability distributions otherwise, would be to invite inconsistency. But there is one degree of freedom remaining in the Bayesian decision theory as a whole. This remaining degree of freedom lies in the choice of the position measure for a given probability distribution.

In any problem of choice we will endeavor to choose that decision which has a corresponding utility probability distribution that is lying farthest to the right on the utility axis; that is, we will choose to maximize the position of our

utility probability distributions. In this there is little freedom. We are free, in principle, to choose the measures of the positions of our utility probability distributions any way we see fit. Nonetheless, we believe that it is always a good policy to take into account all the pertinent information we have.

If we maximize only the expected values of the utility probability distributions, then we will, by definition, neglect the information that the standard deviations of the utility probability distributions provide regarding our problem of choice, by way of the symmetry breaking in the case of an overshoot of one of the confidence bounds. Likewise, we are free to maximize only one of the confidence bounds of our utility probability distributions, while neglecting the other. But in doing so, we will be performing probabilistic minimax or maximax analyses, and, consequently, neglecting the possibilities of both potentially astronomical gains in the upper bound and potentially catastrophic losses in the lower bound. If we only maximize the sum of the lower and upper bound, or a scalar multiple thereof, then we will make a trade-off between the probabilistic worst- and best-case scenarios, but in the process, we will, for unimodal distributions, be neglecting the location of the bulk of our probability distributions.

This is why, in our minds, the mean of the sum of the lower confidence bound, expected value, and upper bound bound, currently is the best all-round position measure for a given probability distribution, as it reflects the position of the probabilistic worst- and best-case scenarios, as well as the position of the expected outcome.

Having removed the degree of freedom of the utility function by way of a consistency derivation, we now should endeavor to find a similar consistency derivation of the measure of the position of a given probability distribution. Until the time we do so, we will have to make do with the kind of *ad hoc* common-sense reasoning that led us from the expected value as the traditional position measure to the alternative position measure of Weaver's criterion of choice, as we let actual empirical performance be the arbiter of which of these position measures is optimal [106, 105].

In regards to empirical performance of Weaver's criterion of choice, it has been found in Chapter 14 that a balanced Weaver position measure together with Bernoulli's utility function odels suffice to explain the postulated phenomenon of probability weigting, or, equivalently, the observed inverse S-shape of the ratio of the certain and uncertain outcomes as a function of the probability of the uncertain outcome [103]. Also, in Chapters 15 and 16 we find that the reflection effect and its consequent fourfold pattern may be modeled by way of Weaver's position measure together with Bernoulli's utility function.

And in the face of large certain gains both introspection and psychological experimentation would indicate that imbalanced risk aversion will occur, as the sigma bound of the probabilistic worst-case scenario in Weaver's criterion of choice is set to some value greater than one. Likewise, for and large certain losses an imbalanced risk seeking is found, as the sigma bound of the best-case scenario in Weaver's criterion of choice is set to some value greater than one.

In Chapter 17 we find that Weaver's balanced position measure gives us a new class of adjusted odds ratios. And in Chapter 18 it is demonstrated

that insurance companies may assign not only the same probabilities and consequences to the contingencies that are to be insured as their customers, but also may have the same utility functions as their customers, and still be able to make a healthy profit, as the insurance companies have a decision theoretical law of large numbers at their side that allows them to spread their risk over multiple contracts.

Furthermore, in Chapter 19 we have an example where imbalanced risk aversion will be expected on the part of the decision maker by those who stand to be affected by the decisions. For if decisions are to be made about flood defenses, then public opinion will demand that it is better to be safe than sorry. And it is found that Weaver's criterion of choice together with Bernoulli's utility function can justify an investment willingness into safety that is an order of magnitude greater than would be expected under the traditional expected value criterion of choice and a linear utility function.

Part III

Bayesian Information Theory

Chapter 21

Introduction

In Bayesian probability theory we assign probabilities to propositions and then operate on these probabilities by way of the probability theoretic product and sum rules. Bayesian probability theory has now been supplemented with an extended information theory, or equivalently, an inquiry calculus [58, 59, 61, 62, 63, 65, 104]. This new information theory, which is Bayesian in its outlook, constitutes an expansion of the canvas of rationality [94], and, consequently, of the range of psychological phenomena which are amenable to mathematical analysis.

In the inquiry calculus relevancies are assigned to sources of information. By way of the information theoretic product and sum rules, relevancies may then be operated upon in order to determine the relevancy of a source of information in regards to some issue of interest. So, if the probability measure of probability theory assigns numerical values to the plausibilities of our propositions, then the relevance measure of information theory assigns, for some issue of interest, numerical values to the potential pertinence of the variables we are considering to include in our inference. And the fact that the extended information theory admits a product rule and a sum rule puts this theory on the same footing as the Bayesian probability theory, which is why we use the qualifier 'Bayesian' in connection with this information theory.

For example, if there is the possibility of some danger, then the Bayesian information theory allows us to assign relevancies to statements made by officials in regards to that danger. It is then found that the relevance of that official source is directly related to its unbiasedness and competence. A high probability of unbiasedness and competence imply a corresponding high relevance, and a low probability of unbiasedness and competence imply a low relevance. This mathematical derived result is in close correspondence with social scientific findings [95].

In this part of the thesis we first give the derivation of both the Bayesian probability and information theories, by way of consistency on the lattices of, respectively, statements and questions. Having established the axiomatic underpinnings of the Bayesian information theory, we proceed to give the reader

some feeling for the relevance measure. We then give a risk communication case study. We conclude this part of the thesis with a short historical overview of the development of information theory. This development goes from the very specific, i.e., Shannon's information entropy, to the more general, i.e., the cross-entropy, and the final generality only comes with relevancies and their sum and product rules.

Chapter 22

The Probability and Inquiry Calculi

In this chapter we derive both the probability and the inquiry caluculi, or, equivalently, the Bayesian probablity and information theories, by defining measures on lattices.

22.1 Lattices

Lattices follow from ordering. In lattice theory elements are ordered by way of joins $\vee$ and meets $\wedge$, where the join of two elements, as a convention, is taken to be the upper bound of these elements and, whereas the meet is taken as the lower bound. So, in a lattice we have that the join of all the elements is located at the top of the lattice, whereas the meet is located at the bottom. In Figure 22.1 we give a general lattice of elements x, y, and z. It can be seen in this figure that the element $x \vee y$ is located above the element x, whereas the element $x \wedge y$ is located below. By way of the binary ordering relations $\leq$ and $\geq$, we may denote these relations as $x \leq x \vee y$ and $x \geq x \wedge y$, respectively. If $x \leq y$, then we say that element x is contained by element y, and if $x \geq y$, then we say that element x contains element y.

22.2 Unconditial Valuations

Valuations v may be assigned to the lattices that take the lattice elements x, $x \wedge y$, and $x \vee y$ to the numbers $v(x)$, $v(x \wedge y)$, and $v(x \vee y)$. The structure of the lattice constrains the valuations v and these constraints are enforced by way of constraint equations. In order for the valuations to be consistent with all the binary ordering relations within the lattice, the valuations of the contained elements in the lattice must be either smaller-equal or greater-equal than the

valuations of their containing elements:

$$x \leq y \qquad \text{implies} \qquad v\left(x\right) \leq v\left(y\right) \quad \text{or} \quad v\left(y\right) \leq v\left(x\right). \qquad (22.1)$$

This is the first and most primitive of our constraint equations; a constraint equation that is so fundamental that we shall call it the fundamental constraint.

The second constraint equation on valuations, which does have a name, is the so-called general sum rule[1] [58, 59, 60, 61, 62, 63]:

$$v\left(x \vee y\right) = v\left(x\right) + v\left(y\right) - v\left(x \wedge y\right). \qquad (22.2)$$

The sum rule ensures that the binary ordering relations of the valuations are consistent with those of the lattice itself, just like the fundamental constraints (22.1). But the sum rule provides much more structure, as it relates the valuations of the binary ordered elements as an identity, rather than an inequality.

It is to be noted that the sum rule allows for a strictly monotonic decreasing or increasing one-to-one regrade

$$\Theta\left[\nu\left(x\right)\right] = v\left(x\right) \qquad (22.3)$$

such that

$$\nu\left(x\right) = \Theta^{-1}\left[v\left(x\right)\right]. \qquad (22.4)$$

Substituting (22.3) into (22.2) we obtain

$$\Theta\left[\nu\left(x \vee y\right)\right] = \Theta\left[\nu\left(x\right)\right] + \Theta\left[\nu\left(y\right)\right] - \Theta\left[\nu\left(x \wedge y\right)\right], \qquad (22.5)$$

from which it follows that the regraded valuation (22.4) also admits a sum rule:

$$\nu\left(x \vee y\right) = \Theta^{-1}\left(\Theta\left[\nu\left(x\right)\right] + \Theta\left[\nu\left(y\right)\right] - \Theta\left[\nu\left(x \wedge y\right)\right]\right). \qquad (22.6)$$

This observation is useful in that it is found that a linear rescaling and division of valuations, i.e,

$$\Theta\left(x\right) = Cx \qquad \text{and} \qquad \Theta\left(x\right) = 1/x, \qquad (22.7)$$

do not destroy the ordering which is enforced by the sum rule.

22.3   Bi-Valuations

If we want to quantify the *degree* of ordering of element x relative to some context element c, then we need to go from our initial univariate valuation $v\left(x\right)$ to the conditional valuation $m\left(x|c\right)$. In order for the bivaluations to be consistent with all the binary ordering relations within the lattice, as with the valuations, (22.1), we must have

$$x \leq y \qquad \text{implies} \qquad m\left(x|c\right) \leq m\left(y|c\right) \quad \text{or} \quad m\left(y|c\right) \leq m\left(x|c\right). \qquad (22.8)$$

[1]See also Chapter 8.

$$x \vee y \vee z$$

$$x \vee y \qquad y \vee z$$

$$x \qquad y \qquad z$$

$$x \wedge y \qquad y \wedge z$$

$$x \wedge y \wedge z$$
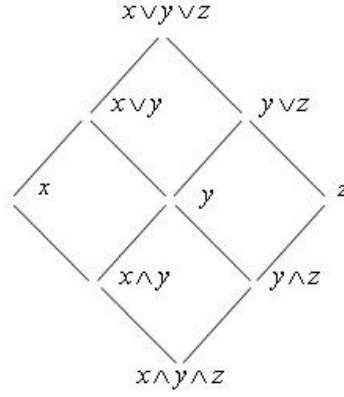
Figure 22.1: Lattice of elements x, y, and z.

Also, the introduction of the degree of ordering does not do away of our need to maintain the order of the unquantified lattice in our valuations. It follows that for conditional valuations the general sum rule must also hold as a second constraint equation, (22.2):

$$m \left(x \vee y \mid c \right) = m \left(x \mid c \right) + m \left(y \mid c \right) - m \left(x \wedge y \mid c \right). \tag{22.9}$$

Moreover, by introducing the concept of a context there is introduced a commensurate concept of a change of context. From changes of context there then follows the chain rule[2] as a third constraint equation on conditional valuations [64, 65]:

$$m \left(x \mid c \right) = m \left(x \mid y \right) m \left(y \mid c \right), \tag{22.10}$$

for chained lattice elements $x \leq y \leq c$ and $c \leq y \leq x$. The chain rule ensures that the binary ordering relations of the conditional valuations are consistent with those of the lattice itself as we go from one context to the other.

We may apply (22.10) to the chained elements $x \leq x \leq c$ and $c \leq x \leq x$. This gives for both chains the identity

$$m \left(x \mid c \right) = m \left(x \mid x \right) m \left(x \mid c \right),$$

which is only consistent for

$$m \left(x \mid x \right) = 1. \tag{22.11}$$

It follows that the degree of ordering of any element in the lattice relative to itself is constrained by the chain rule to be 1.

In the sum and chain rules, (22.9) and (22.10), we have the general bi-valuation calculus for the universal Platonic lattice in Figure 22.1, which has

---

[2]See also Chapter 8

lattice elements whose meaning is not yet specified. If we specify the nature of the elements and the join and meet operators in our lattice, then we go from the one universal lattice, the elements of which carry no specific meaning, to all the specific lattices we might conceive of.

We will now apply to the general bi-valuation calculus to the specific lattices of propositions and questions. It will be found that the specifics of these lattices introduce additional constraints that are unique to these lattices.

22.4   Valuations on the Lattice of Propositions

If define the elements in the Platonic lattice of Figure 22.1 to be propositions, the join $\vee$ to be the OR-operator of Boolean logic, and the meet $\wedge$ be the AND-operator, then we have the specific (Boolean) lattice of propositions.

The ordering relation of the lattice of propositions naturally encodes logical implication, such that a given proposition implies all the propositions above it. Logical deduction is straightforward in this framework since every proposition in the lattice implies (i.e., is included by) all the proposition above it with certainty. For example, x implies $x \vee y$, $x \vee y \vee z$, etc. The lattice of propositions is in this sense an algebra of deduction.

Logical induction, however, works backwards. In induction we quantify the degree to which one's current state of knowledge implies a proposition of lower certainty below it. So, in order to go from deduction to induction, we need to generalize the algebra of deduction to a calculus of induction, by way of a bi-valuation on the lattice of propositions. In what follows we derive the constraints on a bi-valuation measure, called probability, that quantifies the degree to which one proposition implies another.

Since we let the ordering relation be degree of implication, we may interpret the constraint (22.11) to signify that the proposition x implies itself absolutely. Moreover, by way of of the lattice of propositions' natural encoding of logical implication, we have that any proposition above x is implied with absolute certainty:

$$x \leq c \quad \text{implies} \quad m\left(c|\,x\right) = 1. \tag{22.12}$$

By way of this additional constraint, which is introduced by the specific meaning we have assigned to the lattice elements and lattice join and meet, together with the sum rule, we may further constrain the chain rule into a product rule that is specific for upper contexts.

If for the small diamond in Figure 22.1 which is defined by x, $x \vee y$, y, and $x \wedge y$ we consider the context to be x, then the sum rule for this diamond may be written down as, (22.9),

$$m\left(x \vee y|\,x\right) + m\left(x \wedge y|\,x\right) = m\left(x|\,x\right) + m\left(y|\,x\right). \tag{22.13}$$

Since $x \leq x$ and $x \leq x \vee y$, we have that the statement x implies both statements x and $x \vee y$ with absolute certainty, that is, (22.12),

$$m\left(x|\,x\right) = m\left(x \vee y|\,x\right) = 1. \tag{22.14}$$

Substituting (22.14) into sum rule (22.13), we obtain the further constraint:

$$m\left(x \wedge y\,|\,x\right) = m\left(y\,|\,x\right), \tag{22.15}$$

which holds for arbitrary elements x and y in the lattice of propositions that are closed under the join and meet, and which is expressed by the equivalence of the arrows in Figure 22.2.
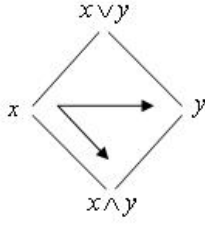
Figure 22.2: The diamond x, $x \vee y$, y, and $x \wedge y$

Consider the chain where the bi-valuation $m\left(x \wedge y \wedge z\,|\,x\right)$ with context x is decomposed into two parts, by introducing the intermediate context $x \wedge y$. The chain rule (22.10) gives

$$m\left(x \wedge y \wedge z\,|\,x\right) = m\left(x \wedge y \wedge z\,|\,x \wedge y\right) m\left(x \wedge y\,|\,x\right) \tag{22.16}$$

and the constraint (22.15) gives

$$
\begin{aligned}
m\left(x \wedge y \wedge z\,|\,x\right) &= m\left(y \wedge z\,|\,x\right), \\
m\left(x \wedge y \wedge z\,|\,x \wedge y\right) &= m\left(z\,|\,x \wedge y\right), \\
m\left(x \wedge y\,|\,x\right) &= m\left(y\,|\,x\right).
\end{aligned}
\tag{22.17}
$$

If we substitute the simplifications (22.17) into (22.16), we obtain the specific product rule for upper contexts:

$$m\left(y \wedge z\,|\,x\right) = m\left(z\,|\,x \wedge y\right) m\left(y\,|\,x\right). \tag{22.18}$$

For the lattice of propositions the meet $\wedge$ is the Boolean AND-operator. Since this operator is commutative (i.e., $y \wedge z = z \wedge y$), the constraint (22.18) relaxes to

$$m\left(y\,|\,x \wedge z\right) m\left(z\,|\,x\right) = m\left(y \wedge z\,|\,x\right) = m\left(z\,|\,x \wedge y\right) m\left(y\,|\,x\right). \tag{22.19}$$

So, by assigning meaning to the lattice we obtain the additional constraints (22.12). This then translates to the constraint (22.15), which then refines the chain rule of the general bi-valuation calculus into the product rule for upper contexts.

The product rule for upper contexts allows us determine the valuation for impossibility. If proposition y implies to impossibility under context x, then so must the logical conjunction of the propositions y and z (i.e., $y \wedge z$). Substituting these valuations into (22.18),

$$m\left(\text{impossibility}|\, x\right) = m\left(\,z|\, x \wedge y\right) m\left(\text{impossibility}|\, x\right),$$

it is found that impossibility must either translate to a valuation of 0 or ∞, in order for this identity to hold. We then have, from (22.12) and the sum rule (22.9), for a context $x \vee y$ where the propositions x and y are mutually exclusive and, as a consequence, $x \wedge y$ is impossible, the following identity:

$$1 = m\left(\,x|\, x \vee y\right) + m\left(\,y|\, x \vee y\right) - m\left(\text{impossibility}|\, x \vee y\right).$$

From which it follows that the valuation of the impossibility $x \wedge y$ necessarily equals zero:

$$m\left(\text{impossibility}|\, c\right) = 0. \tag{22.20}$$

By way of (22.12) and (22.20), it then follows that for the sum rule (22.9) the first option of the fundamental constraint (22.8) must hold, rather than the second. The removal of this degree of freedom allows us to determine the unknown bivariate function F that quantifies the degree of ordering of element x relative to some higher context element y in the lattice with an upper context.

Because of (22.15), we have that

$$m\left(\,x|\, y\right) = m\left(\,x \wedge y|\, y\right).$$

So, we are looking for the bivariate bi-valuation function

$$F\left[v\left(x \wedge y\right), v\left(y\right)\right] = m\left(\,x|\, y\right). \tag{22.21}$$

Since the sum rule allows for linear rescaling of the valuations, (22.7), we want our bi-valuation to be invariant for such a rescaling. This then puts the following constraint on the unknown function F:

$$F\left[v\left(x \wedge y\right), v\left(y\right)\right] = F\left[C\, v\left(x \wedge y\right), C\, v\left(y\right)\right]. \tag{22.22}$$

The general solution of this homogenous function of degree zero is [1]

$$F\left[v\left(x \wedge y\right), v\left(y\right)\right] = f\left[\frac{v\left(x \wedge y\right)}{v\left(y\right)}\right], \tag{22.23}$$

where f is some unknown function. Because of (22.8) and (22.12), it then follows that f is the identity function:

$$f\left(x\right) = x. \tag{22.24}$$

Substituting (22.23) and (22.24) into (22.21), we obtain the bi-valuation for the lattices with an upper context:

$$m\left(\,x|\, y\right) = \frac{v\left(x \wedge y\right)}{v\left(y\right)}. \tag{22.25}$$

By relabeling the measure m to p, the context symbol c to I, and the operator symbols $\vee$ and $\wedge$ to the corresponding Boolean OR- and AND-operators, '+' and (optional) '·', we may recognize in the constraints (22.9) and (22.19) the sum and product rules of probability theory:

$$p\left(x+y\middle|\,I\right) = p\left(x\middle|\,I\right) + p\left(y\middle|\,I\right) - p\left(x\cdot y\middle|\,I\right) \tag{22.26}$$

and

$$p\left(y\middle|\,I\right)p\left(x\middle|\,y\right) = p\left(x\cdot y\middle|\,I\right) = p\left(x\middle|\,I\right)p\left(y\middle|\,x\right),$$

where use has been made of the fact that for $x, y \leq I$

$$x \cdot I = x \qquad \text{and} \qquad y \cdot I = y.$$

The probability measure has a range, (22.12) and (22.20),

$$0 \leq p\left(x\middle|\,I\right) \leq 1 \tag{22.27}$$

and is defined as a ratio of valuations, (22.25):

$$p\left(x\middle|\,y\right) = \frac{v\left(x \wedge y\right)}{v\left(y\right)}, \tag{22.28}$$

where, the first option in (22.8),

$$x \wedge y \leq x \quad \text{implies} \quad v\left(x \wedge y\middle|\,I\right) \leq v\left(y\middle|\,I\right). \tag{22.29}$$

With respect to probability theory, Knuth has laid a new foundation that encompasses and generalizes the Cox formulations [17]. By introducing probability as a bi-valuation defined on a lattice of propositions one can quantify the degree to which one proposition implies another.

This generalization from logical implication to degrees of implication not only mirrors Cox's notion of plausibility as a degree of belief, but includes it. The main difference is that in Knuth's formulation universal symmetries of lattices in general form the basis of the theory. The meaning of the derived measure is then inherited from the ordering relation of the specific lattice under consideration, which in the case of lattice of propositions is implication [64].

Cox's formulation, however, is based on a set of desiderata which are formulated with a particular notion of plausibility in mind. This makes the Cox formulation way less universal than the formulation that followed it. Case in point, the latter Knuth formulation provides both a foundation and a road map towards an information theory that has sum and product rules analogous to the inferential calculus, as we will demonstrate shortly.

22.4.1   An Admissible Regrade of the Sum Rule

As a curiosity, it is to be noted that the sum rule for bi-valuations, like (22.2), also allows one-to-one regrades:

$$\Theta\left[\mu\left(x \vee y\middle|\,c\right)\right] = \Theta\left[\mu\left(x\middle|\,c\right)\right] + \Theta\left[\mu\left(y\middle|\,c\right)\right] - \Theta\left[\mu\left(x \wedge y\middle|\,c\right)\right], \tag{22.30}$$

where (22.4)

$$\mu\left(x|\,c\right) = \Theta^{-1}\left[m\left(x|\,c\right)\right]. \tag{22.31}$$

Would we take advantage of this freedom and in (22.9) assign the regrade

$$\Theta\left(x\right) = \Theta^{-1}\left(x\right) = \frac{1}{x}, \tag{22.32}$$

such that

$$\mu\left(x|\,c\right) = \frac{1}{m\left(x|\,c\right)}, \tag{22.33}$$

then we would obtain the equivalent probability theory, mentioned in [47], in which impossibility is signified by ∞. This equivalent probability theory has a different sum rule,

$$\frac{1}{\pi\left(x+y|\,I\right)} = \frac{1}{\pi\left(x|\,I\right)} + \frac{1}{\pi\left(y|\,I\right)} - \frac{1}{\pi\left(x\cdot y|\,I\right)}, \tag{22.34}$$

but the same product rule,

$$\pi\left(y|\,I\right)\pi\left(x|\,y\right) = \pi\left(x\cdot y|\,I\right) = \pi\left(x|\,I\right)\pi\left(y|\,x\right),$$

and a probability measure π that has a range

$$1 \leq \pi\left(x|\,I\right) \leq \infty, \tag{22.35}$$

where ∞ signifies impossibility.

The alternative probability measure π maps one-to-one to the traditional probability measure:

$$\pi\left(x|\,y\right) = \frac{v\left(y\right)}{v\left(x\wedge y\right)} = \frac{1}{p\left(x|\,y\right)}, \tag{22.36}$$

If we define

$$\nu\left(x\right) = 1/v\left(x\right), \tag{22.37}$$

then we obtain from (22.36)

$$\pi\left(x|\,y\right) = \frac{\nu\left(x\cdot y\right)}{\nu\left(y\right)}, \tag{22.38}$$

where, because of (22.34) and (22.37),

$$x\cdot y \leq y \quad \text{implies} \quad \nu\left(y|\,I\right) \leq \nu\left(x\cdot y|\,I\right). \tag{22.39}$$

It is to be noted that in [47] the possibility of assigning infinity to signify impossibility, or, equivalently, is derived after the product rule and before the sum rule have been established, which is commensurate with the fact that the product rule remains the same after (i.e., is not influenced by) an inverse regrade, (22.35).

Also, comparing (22.34) and (22.39), we find that an inverse regrade of the sum rule gives a commensurate switching in the fundamental constraint (22.8) from the first to the second option. So, if we first set the fundamental constraint axiomatically to its first option and then derive the sum rule, as is done in [65], then we will find that the (monotonically decreasing) inverse regrade is inadmissible.

22.5 What is a Question?

Before we can proceed to construct a specific lattice of question we first need to have a clear notion of the nature of questions. In his last scientific publication Cox explored the logic of inquiry [19]. In this paper he defined a question as a set of all possible statements that answer it. So, given a hypothesis space of possible answers, that is, statements, one can construct questions, and compare their equivalence by comparing their sets of answers.

For example, both questions 'is it raining?' and 'is it not raining?' are equivalent since they are both answered by the same set of statements [59]. Say, we have three possible weather states

$$a \equiv \text{raining}, \quad b \equiv \text{snowing}, \quad c \equiv \text{sunny}.$$

For compactness of notation, let

$$A \equiv \{a\}, \quad B \equiv \{b\}, \quad C \equiv \{c\},$$

$$AB \equiv \{a, b, a+b\}, \quad AC \equiv \{a, c, a+c\}, \quad BC \equiv \{b, c, b+c\}$$

$$ABC \equiv \{a, b, c, a+b, a+c, b+c, a+b+c\},$$

where '+' is the symbol for a disjunction of propositions that make up an atomic element of a question and, for example, AB is the statement 'it is either raining or snowing'. Then the set of all real questions is [59]:

$$\{ABC,$$

$$AB \cup AC \cup BC,$$

$$AB \cup AC, \ AB \cup BC, \ AC \cup BC,$$

$$AB \cup C, \ AC \cup B, \ BC \cup A,$$

$$A \cup B \cup C\},$$

where the symbol '$\cup$' stands for a disjunction of the atomic elements that make a question; i.e., we use the symbol '+' for the construction of the atomic elements of which the questions are constituted, whereas the symbol '$\cup$' is used to combine these atomic into questions, or, equivalently, sets of answers.

Note that the real questions have been ordered by set inclusion, henceforth denoted as '$\leq$', with the least concise question, that is, the question with the largest set of possible answers (i.e., ABC) at the top, and the most concise question, that is, the question with the smallest set of possible answers (i.e., $A \cup B \cup C$) at the bottom.

The set of possible answers to the concrete question 'is it sunny or not?' constitutes that selfsame question 'is it sunny or not?':

$$AB \cup C = \{a, b, c, a + b\}. \tag{22.40}$$

The question (22.40) is concrete in the sense that for each weather state there can only be one correct answer; i.e., if either of the weather states a or b holds then the correct answer to the question whether it is sunny or not is the statement that it is not sunny $a + b$, whereas under weather state c the only correct answer is that it is sunny c.

The ambiguous question 'is it raining or snowing, raining or sunny, or snowing or sunny?' is answered by all the elements in the set $AB \cup AC \cup BC$, where

$$AB \cup AC \cup BC = \{a, b, c, a + b, a + c, b + c\}. \tag{22.41}$$

The question (22.41) is ambiguous in the sense that for a given weather state two answers may be given; e.g., for weather state a one may answer with either $a + b$ or $a + c$.

So, a question is the set of statements that can be given as an answer to that question. As each question represents a set of answers, related questions may be ordered by set inclusion. This ordering relation of set inclusion implements the concept of answering [59]. So, if question Q_1 is a subset of question Q_2, then $Q_1 \leq Q_2$, and by answering question Q_1 we will have necessarily answered question Q_2.

For example, the question 'is it raining, snowing or sunny?', that is,

$$A \cup B \cup C = \{a, b, c\}, \tag{22.42}$$

also answers the question 'is it raining or not?'

$$A \cup BC = \{a, b, c, b + c\}. \tag{22.43}$$

Comparing the sets corresponding with questions (22.42) and (22.43), we may easily check that the former is indeed included by the latter

$$\{a, b, c\} = A \cup B \cup C \leq AB \cup C = \{a, b, c, a + b\}. \tag{22.44}$$

Since questions are just sets of all the possible statements that answer that question, we have that the logical meet $\cap$ and join $\cup$ of set theory may be applied to questions. The meet of the questions $AB \cup C$, 'is it sunny or not?', and $A \cup BC$, 'is it raining or not?', gives the question 'is it raining, snowing, or sunny?':

$$(AB \cup C) \cap (A \cup BC) = \{a, b, c, a + b\} \cap \{a, b, c, b + c\}$$

$$= \{a, b, c\} \tag{22.45}$$

$$= A \cup B \cup C.$$

This may be seen as follows. If we first ask if it is sunny, then we will either know that it is sunny or not. If it is not sunny, then we may inquire further, and ask whether it is raining or not, after which we will know exactly what kind of weather it is. We would have gotten the same result had we asked directly whether it was raining, snowing, or sunny. So, the meet of two questions tends to give us a question that is more informative, when answered, than either question alone.

The join of the questions $AB \cup C$, 'is it sunny or not?', and $A \cup BC$, 'is it raining or not?' gives the question 'is it not raining or is not sunny?':

$$(AB \cup C) \cup (A \cup BC) = \{a, b, c, a+b\} \cup \{a, b, c, b+c\}$$

$$= \{a, b, c, a+b, b+c\} \tag{22.46}$$

$$= AB \cup BC.$$

We can see that a join of two given questions tends to give us a less informative question than either of the questions alone, or, for that matter, the meet of those same questions.

This observation will prove to be crucial in the derivation of the product rule of the lattice of questions.

22.6 Valuations on the Lattice of Questions

If define the elements in the Platonic lattice of Figure 22.1 to be questions, the join $\vee$ to be the union-operator of set theory, and the meet $\wedge$ to be the intersection-operator, then we have the specific lattice of questions.

The ordering relation of the lattice of questions naturally encodes relevance, such that a given question answers all the questions above it. Relevance assignment is straightforward in this framework since a question in the lattice is absolutely relevant for (i.e., is included by) every question above it. For example, the questions $x \wedge y$, $x \wedge y \wedge z$, etc., are all absolutely relevant for the answering of question x. The lattice of questions is in this sense an algebra of relevancy, just like the lattice of statements is an algebra of deduction.

Now, if want to quantify the degree to which a given question is relevant for some other question which is not located directly above it in the lattice of questions, then this will require a generalization of the algebra of questions to a calculus of questions. In what follows we derive the constraints on a bi-valuation measure, called relevancy, that quantifies the degree to which the answering of one question will contribute to the answering of another question.

Since we let the ordering relation be degree of relevancy, we may interpret the constraint (22.11) to signify that the answering of question x is absolutely relevant to the answering of itself. Moreover, by way of of the lattice of questions' natural encoding of relevancy, we have that any question below x will be

absolutely relevant for its answering:

$$c \le x \quad \text{implies} \quad m\left(c|\,x\right) = 1. \tag{22.47}$$

By way of this additional constraint, which is introduced by the specific meaning we have assigned to the lattice elements and lattice join and meet, together with the sum rule, we may now constrain the chain rule into a product rule that is specific for lower contexts.

If for the small diamond in Figure 22.1 which is defined by x, $x \vee y$, y, and $x \wedge y$ we consider the context to be x, then the sum rule for this diamond may be written down as, (22.9),

$$m\left(x \vee y|\,x\right) + m\left(x \wedge y|\,x\right) = m\left(x|\,x\right) + m\left(y|\,x\right). \tag{22.48}$$

Since $x \ge x$ and $x \ge x \wedge y$, we have that the questions x and $x \wedge y$ are absolutely relevant for question x, that is, (22.47),

$$m\left(x|\,x\right) = m\left(x \wedge y|\,x\right) = 1. \tag{22.49}$$

Substituting (22.49) into sum rule (22.48), we obtain the further constraint:

$$m\left(x \vee y|\,x\right) = m\left(y|\,x\right), \tag{22.50}$$

which holds for arbitrary elements x and y in the lattice of propositions that are closed under the join and meet, and which is expressed by the equivalence of the arrows in Figure 22.3.
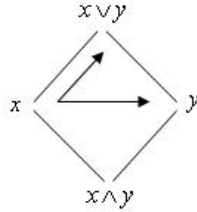
Figure 22.3: The diamond x, $x \vee y$, y, and $x \wedge y$

Consider the chain where the bi-valuation $m\left(x \vee y \vee z|\,x\right)$ with context x is decomposed into two parts, by introducing the intermediate context $x \vee y$. The chain rule (22.10) gives

$$m\left(x \vee y \vee z|\,x\right) = m\left(x \vee y \vee z|\,x \vee y\right) m\left(x \vee y|\,x\right) \tag{22.51}$$

and the constraint (22.50) gives

$$\begin{aligned} m\left(x \vee y \vee z|\,x\right) &= m\left(y \vee z|\,x\right), \\ m\left(x \vee y \vee z|\,x \vee y\right) &= m\left(z|\,x \vee y\right), \\ m\left(x \vee y|\,x\right) &= m\left(y|\,x\right). \end{aligned} \tag{22.52}$$

If we substitute the simplifications (22.52) into (22.51), we obtain the specific product rule for lower contexts:

$$m\left(y \vee z \mid x\right) = m\left(z \mid x \vee y\right) m\left(y \mid x\right). \tag{22.53}$$

For the lattice of question the union $\vee$ is the set-theoretical union-operator. Since this operator is commutative (i.e., $y \vee z = z \vee y$), the constraint (22.53) relaxes to

$$m\left(y \mid x \vee z\right) m\left(z \mid x\right) = m\left(y \vee z \mid x\right) = m\left(z \mid x \vee y\right) m\left(y \mid x\right). \tag{22.54}$$

So, by assigning meaning to the lattice we obtain the additional constraints (22.47). This then translates to the constraint (22.50), which then refines the chain rule of the general bi-valuation calculus into the product rule for lower contexts [104].

The product rule for lower contexts allows us determine the valuation for absolute irrelevancy. If question y is absolutely irrelevant under context x, then so must be the set union of the questions y and z (i.e., $y \vee z$). Substituting these valuations into (22.53),

$$m\left(\text{irrelvancy} \mid x\right) = m\left(z \mid x \vee y\right) m\left(\text{irrelvancy} \mid x\right),$$

it is found that absolute irrelevancy must either translate to a valuation of 0 or ∞, in order for this identity to hold. We then have, from (22.47) and the sum rule (22.9), for a context $x \wedge y$ where the joint question $x \vee y$ is irrelevant, the following identity:

$$1 = m\left(x \mid x \wedge y\right) + m\left(y \mid x \wedge y\right) - m\left(\text{irrelevant} \mid x \wedge y\right).$$

From which it follows that the valuation of the irrelevancy $x \vee y$ necessarily equals zero:

$$m\left(\text{irrelvancy} \mid c\right) = 0. \tag{22.55}$$

By way of (22.12) and (22.55), it then follows that for the sum rule (22.9) the second option of the fundamental constraint (22.8) must hold, rather than the second. The removal of this degree of freedom allows us to determine the unknown bivariate function F that quantifies the degree of ordering of element x relative to some higher context element y in the lattice with an upper context as, (22.25) and (22.50),

$$m\left(x \mid y\right) = \frac{v\left(x \vee y\right)}{v\left(y\right)}. \tag{22.56}$$

By relabeling the measure m to d, the context symbol c to I, and the operator symbols $\vee$ and $\wedge$ to the corresponding set theoretical union- and intersection-operators, here denoted, for notational ease, '+' and (optional) '·', we now have in the constraints (22.9) and (22.54) the sum and product rules of the inquiry calculus:

$$d\left(x + y \mid I\right) = d\left(x \mid I\right) + d\left(y \mid I\right) - d\left(x \cdot y \mid I\right) \tag{22.57}$$

and

$$d\left(y\mid I\right)d\left(x\mid y\right) = d\left(x+y\mid I\right) = d\left(x\mid I\right)d\left(y\mid x\right),$$

where use has been made of the fact that for $I \leq x, y$

$$x + I = x \qquad \text{and} \qquad y + I = y.$$

The relevancy measure has a range, (22.47) and (22.55),

$$0 \leq d\left(x\mid I\right) \leq 1 \tag{22.58}$$

and is defined as a ratio of valuations, (22.56):

$$d\left(x\mid y\right) = \frac{v\left(x+y\right)}{v\left(y\right)}, \tag{22.59}$$

where, the second option in (22.8),

$$y \leq x + y \quad \text{implies} \quad v\left(x+y\mid I\right) \leq v\left(y\mid I\right). \tag{22.60}$$

By introducing relevance as a bi-valuation defined on a lattice of questions we can quantify the degree to which one question is relevant to another. The symmetries of lattices in general form the basis of the theory and the meaning of the derived measure is inherited from the ordering relation, which in the case of questions is relevance. Because of the concept of context, we have that relevance is necessarily conditional, and a Bayes' theorem for inquiry calculus follows as a direct result of the chain rule in terms of a change in context.

22.7   Assigning Measures

We will now discuss how to construct valuations for questions. In order to best do this, we now will discuss the three spaces: the space of states, the state of statements (i.e., propositions), and the state of questions.

22.7.1   The State Space

The state space is an enumeration of all the possible states that our system may be in. For example, if we measure our system with respect to variables A and B, and if each variable can take only two values, then the state space is given as

$$AB = \{ab_{11}, ab_{12}, ab_{21}, ab_{22}\}. \tag{22.61}$$

Since the system must be in one of these states, but in no more than one, we have that the states ab_{ij} are exhaustive and mutually exhaustive.

22.7.2 The Statement Space

A given individual may not know precisely which state the system is in, but may have some information that rules out some states, but not others. So, the set of potential states defines what one can say about the state of the system. For this reason, we call a set of potential states a statement. If we let the elements in the state space AB be propositions, (22.61),

$$AB = \{ab_{11}, ab_{12}, ab_{21}, ab_{22}\},$$

then we may denote a set of potential states by way of the OR-operator $+$ of Boolean logic. For example, if our system can be in the set of states $\{ab_{11}, ab_{12}\}$, then we may denote this as

$$ab_{11} + ab_{12} = a_1 (b_1 + b_2) = a_1. \tag{22.62}$$

Alternatively, if our system can be in states in the set of states $\{ab_{11}, ab_{12}, ab_{22}\}$, then we may denote this as

$$ab_{11} + ab_{12} + ab_{22} = (ab_{11} + ab_{12}) + (ab_{12} + ab_{22}) = a_1 + b_2. \tag{22.63}$$

A statement describes a state of knowledge about the state of the system. The set of all possible statements is called the *hypothesis space*. If we let the join $\vee$ of the lattice be the OR-operator of Boolean logic and the meet $\wedge$ be the AND-operator, then we may may construct a lattice of statements. That is, the hypothesis space may be represented by way of a lattice of statements, or, equivalently, propositions.

The lattice of statements is generated by taking the power set, which is the set of all possible unions of the elements of the set of states AB, and ordering them according to inclusion. For a system of n mutually exclusive possible states, there are

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n \tag{22.64}$$

statements, including the null-meet. So, the state space AB with $n = 4$ elements has 16 possible statements which can be ordered as, (22.61) and (22.64):

$$\{ab_{11} + ab_{12} + ab_{21} + ab_{22},$$

$$ab_{11} + ab_{12} + ab_{21}, \quad ab_{11} + ab_{12} + ab_{22}, \quad ab_{11} + ab_{21} + ab_{22}, \quad ab_{12} + ab_{21} + ab_{22},$$

$$ab_{11} + ab_{12}, \quad ab_{11} + ab_{21}, \quad ab_{11} + ab_{22}, \quad ab_{12} + ab_{21}, \quad ab_{12} + ab_{22}, \quad ab_{21} + ab_{22},$$

$$ab_{11}, \quad ab_{12}, \quad ab_{21}, \quad ab_{22},$$

$$\emptyset\},$$

where the statement at the top is the truism which represents the state of knowledge where one only knows (for this particular instance) that the system can be in one of four possible states and where the statement $\emptyset$ at the bottom is the null-meet which represents logical impossibility.

Degrees of implication may be inferred from the hypothesis space by assigning bi-valuations (i.e., probabilities) to the hypotheses of the state space (22.61).

All the elements in the hypothesis space are closed under both the join $+$ and the meet $\cdot$, seeing that join of the null-meet $\emptyset$ with any element x in the hypothesis space maps to that element (i.e., $\emptyset + x = x$) and the meet of any two elements from the set AB is taken to the null-meet $\emptyset$. So it follows that the sum rule may be applied to any two elements in the hypothesis space. Also, the sum rule (22.26) may be generalized to

$$p\left(\sum_i x_i \middle| I\right) = \sum_i p\left(x_i | I\right) - \sum_{j>i} p\left(x_i \cdot x_j | I\right) + \sum_{k>j>i} p\left(x_i \cdot x_j \cdot x_k | I\right) - \cdots.$$

(22.65)

So, if the propositions x_i are mutually exclusive, then the meet of these propositions will map to the null-meet which signifies impossibility. Since we must assign a valuation of zero to logical impossibility, it follows that for mutually exclusive propositions (22.65) will simplify to

$$p\left(\sum_i x_i \middle| I\right) = \sum_i p\left(x_i | I\right).$$

(22.66)

And if we assign probabilities to the exhaustive and mutually exclusive elements of the state space, then we may generate consistent probabilities for all the non-atomic compound statements in the hypothesis space by substituting these probabilities in the right-hand of (22.66); e.g.,

$$p\left(ab_{11} + ab_{12} + ab_{22} | I\right) = p\left(ab_{11} | I\right) + p\left(ab_{12} | I\right) + p\left(ab_{22} | I\right).$$

Now, we are free to relabel the statements of the hypothesis space in a meaningful manner, (22.62) and (22.63):

$$\{I,$$

$$a_1 + b_1, \quad a_1 + b_2, \quad a_2 + b_1, \quad a_2 + b_2,$$

$$a_1, \quad b_1, \quad ab_{11} + ab_{22}, \quad ab_{12} + ab_{21}, \quad b_2, \quad a_2,$$

$$ab_{11}, \quad ab_{12}, \quad ab_{21}, \quad ab_{22},$$

$$\emptyset\},$$

where we denoted the top truism as I. And it is to be noted that the hypotheses $ab_{11} + ab_{22}$ and $ab_{12} + ab_{21}$, which are legitimate elements of the lattice of propositions, will not be entertained that often. Since in most probability analyses concerning product spaces like AB, there only will be interest for the probabilities

$$p(A|I), \quad p(B|I), \quad p(AB|I), \quad p(A+B|I), \quad p(A|B), \quad p(B|A), \tag{22.67}$$

all of which do not pertain to the hypotheses $ab_{11} + ab_{22}$ and $ab_{12} + ab_{21}$.

This simple relabeling together with the ratio-structure of the probability measure also makes insightful the mechanism by which the product rule of probability theory maintains the identity

$$p(a_1|I)(b_1|a_1) = p(a_1|I)\frac{p(ab_{11}|I)}{p(a_1|I)} = p(ab_{11}|I).$$

For we have that, (22.27) and (22.28),

$$p(a_1|I)p(b_1|a_1) = \frac{v(a_1 \cdot I)}{v(I)}\frac{v(a_1 \cdot b_1)}{v(a_1)}$$

$$= \frac{v(a_1)}{v(I)}\frac{v(ab_{11})}{v(a_1)}$$

$$= \frac{v(ab_{11})}{v(I)}$$

$$= \frac{v(ab_{11} \cdot I)}{v(I)}$$

$$= p(ab_{11}|I),$$

where use has been made of the fact that for $x, y \leq I$

$$x \cdot I = x \quad \text{and} \quad y \cdot I = y.$$

22.7.3 The Inquiry Space

The state space is an enumeration of all the possible states that the system under consideration may be in. Now, if we measure our system only with respect to the variables A and B, then the most obvious relevancies in an inquiry-analysis will be

$$d(A|I), \quad d(B|I), \quad d(AB|I), \quad d(A+B|I), \quad d(A|B), \quad d(B|A), \tag{22.68}$$

much like the probabilities (22.67) are the most obvious in a data-analysis. These relevancies pertain to the four questions A, B, AB, and $A+B$ in different contexts, just like the probabilities in (22.67) pertain to the statements a_i, b_j, ab_{ij}, and $a_i + b_j$ in different contexts.

So, we wish to determine the relevancies of the questions A, B, AB, and $A + B$ relative to the central issue $I = AB$ with which these questions form a chain, as well as the relevancies of A and B relative to the issues of interest B and A, respectively, with which these questions form an anti-chain. It will be found that the first four relevancies correspond with normalized entropies and the latter two with normalized informations.

The relevancy measure is defined as the ratio of valuations, (22.59):

$$d\left(x|\,y\right) = \frac{v\left(x+y\right)}{v\left(y\right)},$$

where x is the question for which we wish to determine the relevance in relation to the issue of interest y. When the issue of

$$d\left(x|\,y\right) = \frac{H\left(x\right)}{H\left(y\right)},$$

where $H\left(x\right)$ is the Shannon entropy of the probability distribution of all the statements that answer question x:

$$H\left(y|\,x\right) = \sum_{x} p\left(x|\,I\right) \log \frac{1}{p\left(x|\,I\right)}, \tag{22.69}$$

But when the issue of interest y forms an anti-chain with the inquiry x, then the relevance is the scaled mutual information

$$d\left(x|\,y\right) = \frac{I\left(x,y\right)}{H\left(y\right)},$$

where the mutual information is the following function of Shannon entropies

$$I\left(x,y\right) = H\left(x\right) + H\left(y\right) - H\left(x \cdot y\right). \tag{22.70}$$

The mutual information is the amount of entropy (i.e., uncertainty) that remains in y after we have subtracted the conditional entropy that is 'explained' by x:

$$I\left(x,y\right) = H\left(y\right) - H\left(y|\,x\right). \tag{22.71}$$

where

$$H\left(y|\,x\right) = \sum_{x} p\left(x|\,I\right) \sum_{y} p\left(y|\,x\right) \log \frac{1}{p\left(y|\,x\right)} = \sum_{x,y} p\left(x \cdot y|\,I\right) \log \frac{1}{p\left(y|\,x\right)}.$$
$$\tag{22.72}$$

Normalized Entropies

The question AB, when asked, immediately reveals the state of the system and has as its elements

$$ab_{11}, \quad ab_{12}, \quad ab_{21}, \quad ab_{22}.$$

If we ask question AB, then the remaining uncertainty regarding the state of our system can be quantified by way of the mean conditional surprise, (22.72),

$$H\left(AB|\,AB\right) = \sum_{ij} p\left(ab_{ij}|\,I\right) \log \frac{1}{p\left(ab_{ij}|\,ab_{ij}\right)} = \sum_{ij} p\left(ab_{ij}|\,I\right) \log \frac{p\left(ab_{ij}|\,I\right)}{p\left(ab_{ij}|\,I\right)},$$

or, equivalently,

$$H\left(AB|\,AB\right) = H\left(AB\right) - H\left(AB\right), \tag{22.73}$$

where $H\left(AB\right)$ is the Shannon information entropy, (22.69):

$$H\left(AB\right) = \sum_{ij} p\left(ab_{ij}|\,I\right) \log \frac{1}{p\left(ab_{ij}|\,I\right)}.$$

The question A, when asked, tells us if our system is in state a_1 or in state a_2, and is the down set which has as its top elements

$$a_1 = ab_{11} + ab_{12} \quad \text{and} \quad a_2 = ab_{21} + ab_{22}.$$

If we ask question A, then the remaining uncertainty regarding the state of our system can be quantified by way of the mean conditional surprise, (22.72),

$$H\left(AB|\,A\right) = \sum_{ij} p\left(ab_{ij}|\,I\right) \log \frac{1}{p\left(ab_{ij}|\,a_i\right)} = \sum_{ij} p\left(ab_{ij}|\,I\right) \log \frac{p\left(a_i|\,I\right)}{p\left(ab_{ij}|\,I\right)},$$

or, equivalently, (22.69),

$$H\left(AB|\,A\right) = H\left(AB\right) - H\left(A\right). \tag{22.74}$$

The question $A + B$ when asked, tells us that our system is not in one of either states, ab_{11}, ab_{12}, ab_{21}, or ab_{12}, and is the down set which has as its top elements

$$a_1 + b_1, \quad a_1 + b_2, \quad a_2 + b_1, \quad a_2 + b_2.$$

Questions A, B, and AB are concrete in that, for a given system state, they only admit one correct answer. The question $A + B$, however, is ambiguous in that, for a given system state, multiple correct answers are admitted. For example, for a given system state of ab_{11}, the answers $a_1 + b_1$, $a_1 + b_2$, and $a_2 + b_1$ are all legitimate answers to the question $A + B$. So, if we ask question $A + B$, then

the mean conditional surprise is given as

$$
\begin{aligned}
H\left(AB|\,A+B\right) = \ &p\left(ab_{11}|\,I\right)\frac{1}{3}\left[\log\frac{1}{p\left(a_{11}|\,a_1+b_1\right)}+\log\frac{1}{p\left(a_{11}|\,a_1+b_2\right)}+\log\frac{1}{p\left(a_{11}|\,a_2+b_1\right)}\right]\\
&+p\left(ab_{12}|\,I\right)\frac{1}{3}\left[\log\frac{1}{p\left(a_{12}|\,a_1+b_1\right)}+\log\frac{1}{p\left(a_{12}|\,a_1+b_2\right)}+\log\frac{1}{p\left(a_{12}|\,a_2+b_2\right)}\right]\\
&+p\left(ab_{21}|\,I\right)\frac{1}{3}\left[\log\frac{1}{p\left(a_{21}|\,a_1+b_1\right)}+\log\frac{1}{p\left(a_{21}|\,a_2+b_1\right)}+\log\frac{1}{p\left(a_{21}|\,a_2+b_2\right)}\right]\\
&+p\left(ab_{22}|\,I\right)\frac{1}{3}\left[\log\frac{1}{p\left(a_{22}|\,a_1+b_2\right)}+\log\frac{1}{p\left(a_{22}|\,a_2+b_1\right)}+\log\frac{1}{p\left(a_{22}|\,a_2+b_2\right)}\right],
\end{aligned}
$$

or, equivalently,

$$
\begin{aligned}
H\left(AB|\,A+B\right) = \ &p\left(ab_{11}|\,I\right)\frac{1}{3}\left[\log\frac{p\left(a_1+b_1|\,I\right)}{p\left(a_{11}|\,I\right)}+\log\frac{p\left(a_1+b_2|\,I\right)}{p\left(a_{11}|\,I\right)}+\log\frac{p\left(a_2+b_1|\,I\right)}{p\left(a_{11}|\,I\right)}\right]\\
&+p\left(ab_{12}|\,I\right)\frac{1}{3}\left[\log\frac{p\left(a_1+b_1|\,I\right)}{p\left(a_{12}|\,I\right)}+\log\frac{p\left(a_1+b_2|\,I\right)}{p\left(a_{12}|\,I\right)}+\log\frac{p\left(a_2+b_2|\,I\right)}{p\left(a_{12}|\,I\right)}\right]\\
&+p\left(ab_{21}|\,I\right)\frac{1}{3}\left[\log\frac{p\left(a_1+b_1|\,I\right)}{p\left(a_{21}|\,I\right)}+\log\frac{p\left(a_2+b_1|\,I\right)}{p\left(a_{21}|\,I\right)}+\log\frac{p\left(a_2+b_2|\,I\right)}{p\left(a_{21}|\,I\right)}\right]\\
&+p\left(ab_{22}|\,I\right)\frac{1}{3}\left[\log\frac{p\left(a_1+b_2|\,I\right)}{p\left(a_{22}|\,I\right)}+\log\frac{p\left(a_2+b_1|\,I\right)}{p\left(a_{22}|\,I\right)}+\log\frac{p\left(a_2+b_2|\,I\right)}{p\left(a_{22}|\,I\right)}\right],
\end{aligned}
$$

or, more succinctly, as we gather all the terms,

$$
H\left(AB|\,A+B\right)=\sum_{ij}p\left(ab_{ij}|\,I\right)\log\frac{1}{p\left(ab_{ij}|\,I\right)}-\frac{1}{3}\sum_{ij}p\left(a_i+b_j|\,I\right)\log\frac{1}{p\left(a_i+b_j|\,I\right)}.
$$

$$(22.75)$$

We now define for probabilities $p\left(A+B|\,I\right)$ where the variables A and B can take on n and m values, respectively, the Shannon entropy to be

$$
H\left(A+B\right)=\frac{1}{n+m-1}\sum_{ij}p\left(a_i+b_j|\,I\right)\log\frac{1}{p\left(a_i+b_j|\,I\right)},\qquad(22.76)
$$

where $n+m-1$ is number of legitimate answers allowed for each system state as well as the sum

$$
\sum_{ij}p\left(a_i+b_j|\,I\right)=\sum_{ij}p\left(a_i|\,I\right)+p\left(b_j|\,I\right)-p\left(ab_{ij}|\,I\right)=n+m-1.
$$

Then the amount of uncertainty which will remain after asking question $A+B$, can be written down, (22.69), (22.75) and (22.76),

$$
H\left(AB|\,A+B\right)=H\left(AB\right)-H\left(A+B\right).\qquad(22.77)
$$

Now, seeing that our total uncertainty in regards to the state of our system is quantified by the Shannon entropy $H\left(AB\right)$, we may take as an intermediate

relevance valuation ϕ for a given question the total uncertainty in the system minus the uncertainty that remains after having asked that question, (22.73), (22.74), and (22.77):

$$\phi\left(AB\right) = H\left(AB\right) - H\left(\left.AB\right|AB\right) = H\left(AB\right),$$
$$\phi\left(A\right) = H\left(AB\right) - H\left(\left.AB\right|A\right) = H\left(A\right), \qquad (22.78)$$
$$\phi\left(A+B\right) = H\left(AB\right) - H\left(\left.AB\right|A+B\right) = H\left(A+B\right).$$

And if we compare (22.78) with (22.71), then it can be observed that the Shannon entropy of a question is equivalent to a mutual information between that question and the central issue which answers all questions. From which it follows that all relevancies, in a sense, are normalized informations.

Since the questions include the central issue $I = AB$,

$$AB, A, (A+B) \geq I,$$

we may use the fact that

$$x \geq I \qquad \text{implies} \qquad x = x + I,$$

to come to the needed valuations v:

$$v\left(AB+I\right) = \phi\left(AB\right),$$
$$v\left(A+I\right) = \phi\left(A\right), \qquad (22.79)$$
$$v\left[\left(A+B\right)+I\right] = \phi\left(A+B\right),$$

where we let

$$v\left(I\right) = \phi\left(AB\right). \qquad (22.80)$$

Substituting (22.78), (22.79) and (22.80) into (22.59)

$$d\left(\left.x\right|I\right) = \frac{v\left(x+I\right)}{v\left(I\right)},$$

we obtain the relevancies to the questions AB, A, and $A+B$, relative to the central issue $I = AB$:

$$d\left(\left.AB\right|I\right) = \frac{H\left(AB\right)}{H\left(AB\right)}, \quad d\left(\left.A\right|I\right) = \frac{H\left(A\right)}{H\left(AB\right)}, \quad d\left(\left.A+B\right|I\right) = \frac{H\left(A+B\right)}{H\left(AB\right)},$$
$$(22.81)$$

where the Shannon entropy H for ambiguous (i.e., disjunctive) questions like $A+B$ is defined as (22.76):

$$H\left(A+B\right) = \frac{1}{n+m-1}\sum_{ij} p\left(\left.a_i+b_j\right|I\right)\log\frac{1}{p\left(\left.a_i+b_j\right|I\right)}, \qquad (22.82)$$

where $n+m-1$ is the sum

$$\sum_{ij} p\left(\left.a_i+b_j\right|I\right) = \sum_{ij} p\left(\left.a_i\right|I\right) + p\left(\left.b_j\right|I\right) - p\left(\left.ab_{ij}\right|I\right) = n+m-1.$$

In closing, here some further examples of ambiguous questions. If we have the product space ABC, with measurements on the variables A, B, and C, where the variables can take on n_A, n_B, and n_C different values, then for the ambiguous question $A + B + C$ we have that

$$H\left(A + B + C\right) = \frac{1}{C} \sum p\left(a_i + b_j + c_k \mid I\right) \log \frac{1}{p\left(a_i + b_j + c_k \mid I\right)},$$

where, (22.65),

$$\begin{aligned} p\left(a_i + b_j + c_k \mid I\right) &= p\left(a_i \mid I\right) + p\left(b_j \mid I\right) + p\left(c_k \mid I\right) \\ &\quad - p\left(a_i \cdot b_j \mid I\right) - p\left(a_i \cdot c_k \mid I\right) - p\left(b_j \cdot c_k \mid I\right) \\ &\quad + p\left(a_i \cdot b_j \cdot c_k \mid I\right). \end{aligned}$$

and

$$C = \sum_{i,j,k} p\left(a_i + b_j + c_k \mid I\right) = n_B n_C + n_A n_C + n_A n_B - n_C - n_B - n_A + 1,$$

which, for a given system state abc_{ijk}, is the number of permitted answers to the ambiguous question $A + B + C$. Alternatively, for the ambiguous question $AB + C$ we have

$$H\left(AB + C\right) = \frac{1}{C} \sum p\left(ab_{ij} + c_k \mid I\right) \log \frac{1}{p\left(ab_{ij} + c_k \mid I\right)},$$

where, (22.65),

$$p\left(ab_{ij} + c_k \mid I\right) = p\left(ab_{ij} \mid I\right) + p\left(c_k \mid I\right) - p\left(ab_{ij} \cdot c_k \mid I\right)$$

and

$$C = \sum_{i,j,k} p\left(ab_{ij} + c_k \mid I\right) = n_C + n_A n_B - 1,$$

which, for a given system state abc_{ijk}, is the number of permitted answers to the ambiguous question $AB + C$. Finally, as a last example, for the ambiguous question $AB + AC + BC$ we have

$$H\left(AB + AC + BC\right) = \frac{1}{C} \sum p\left(ab_{ij} + ac_{ik} + bc_{jk} \mid I\right) \log \frac{1}{p\left(ab_{ij} + ac_{ik} + bc_{jk} \mid I\right)},$$

where, (22.65),

$$p\left(ab_{ij} + ac_{ik} + bc_{jk} \mid I\right) = p\left(ab_{ij} \mid I\right) + p\left(ac_{ik} \mid I\right) + p\left(bc_{jk} \mid I\right) - 2\, p\left(abc_{ijk} \mid I\right)$$

and

$$C = \sum_{i,j,k} p\left(ab_{ij} + ac_{ik} + bc_{jk} \mid I\right) = n_C + n_B + n_A - 2,$$

which, for a given system state abc_{ijk}, is the number of permitted answers to the ambiguous question $AB + AC + BC$.

Normalized Mutual Informations

If now take as our system B, rather than AB, and ask question A, then the remaining uncertainty regarding the state of our system B can be quantified by way of the mean conditional surprise, (22.72),

$$H\left(B|A\right) = \sum_{ij} p\left(a_i|I\right) p\left(b_j|a_i\right) \log \frac{1}{p\left(b_j|a_i\right)} = \sum_{ij} p\left(ab_{ij}|I\right) \log \frac{p\left(a_i|I\right)}{p\left(ab_{ij}|I\right)},$$

or, equivalently, (22.69),

$$H\left(B|A\right) = H\left(AB\right) - H\left(A\right), \tag{22.83}$$

where it is to be noted that, (22.74) and (22.83),

$$H\left(B|A\right) = H\left(AB|A\right),$$

which is not that surprising, seeing that, by way of the product rule of probability theory (22.27),

$$p\left(b_j|a_i\right) = p\left(ab_{ij}|a_i\right).$$

Again, seeing that our total uncertainty in regards to the state of our system is quantified by the Shannon entropy $H\left(B\right)$, we may take as our relevance valuation v for a given question the total uncertainty in the system minus the uncertainty that remains after having asked that question, (22.78),

$$v\left(A+B\right) = H\left(B\right) - H\left(B|A\right), \tag{22.84}$$

or, equivalently, (22.70) and (22.83),

$$v\left(A+B\right) = H\left(A\right) + H\left(B\right) - H\left(AB\right) = I\left(A,B\right), \tag{22.85}$$

where it is to be noted that $v\left(A+B\right)$ corresponds with a mutual information, and $v\left[\left(A+B\right)+I\right]$ with a Shannon entropy, (22.79). Substituting (22.85) and, (22.79),

$$v\left(B+I\right) = v\left(B\right) = H\left(B\right)$$

into (22.59), we obtain the relevance we are looking for:

$$d\left(A|B\right) = \frac{H\left(A\right) + H\left(B\right) - H\left(AB\right)}{H\left(B\right)} = \frac{I\left(A,B\right)}{H\left(B\right)}. \tag{22.86}$$

So, the relevancies of the questions A, AB, and $A+B$ relative to the chained central issue $I = AB$ correspond with normalized entropies, whereas the relevancy of A relative to the anti-chained issue of interest B corresponds with normalized mutual information, (22.81) and (22.86).

In closing, the mutual information,

$$I\left(x,y\right) = H\left(x\right) + H\left(y\right) - H\left(x\cdot y\right),$$

is used to find the relevance of element x relative to the anti-chained y, or, vise versa, the relevance of y for x, for arbitrary anti-chained elements x and y. So, it follows that only the mutual information is needed to assign relevancies of the type (22.86). Stated differently, higher order informations, with their potentially negative values are not needed to assign relevancies.

22.8   Discussion

By introducing relevance as a bi-valuation defined on a lattice of questions we can quantify the degree to which one question is relevant to another. The symmetries of lattices in general form the basis of the theory and the meaning of the derived measure is inherited from the ordering relation, which in the case of questions is relevance. Because of the concept of context, we have that relevance is necessarily conditional, and a Bayes' Theorem for inquiry theory follows as a direct result of the chain rule in terms of a change in context.

By way of a quantification on the lattice of statements, it is derived in [64] that the product rule of probability calculus gives the degree of implication of the logical meet of two statements relative to some upper context. Now, in the lattice of statements the join of statements $x \vee y$ is absolutely implied by the lower contexts x and y. Whereas, because of the definition of a question, in the lattice of questions it is the meet of questions $x \wedge y$ that is absolutely relevant for the upper contexts x and y.

So, if for the alternative lattice of questions we follow the derivations for the product rule in [64] with this simple observation in mind, then there effortlessly flows forth a dual product rule of inquiry calculus which gives the degree of relevance of the logical join of two questions relative to some lower context. It follows that the derivation of the product rule of inquiry calculus is more an uncovering of what is already there, hiding in plain sight, beneath the theoretical scaffolding laid down in [64]. And it is in this specific sense that the uncovering of the information theoretical product rule in [104] is nothing more than "a technical and conceptual refinement" of Knuth's inquiry calculus [68].

Further conceptual refinements have now been presented in this thesis in that we have extended the Shannon entropy from concrete to ambiguous questions, and the realization that the mutual information is to only information needed; i.e., since all relevancy assignments for anti-chained contexts can be decomposed in a question x and a context y, say, where both x and y can be join or a meet of multiple lattice elements, there is no need to introduce an third lattice element z into the relevancy equation. For comparison, in [62, 16] there is proposed to use higher order informations to assign relevancies to ambiguous questions. But, as noted before, this proposal results in potentially negative relevancies, as higher order informations may become negative, which is prohibitively problematic.

Chapter 23

Measures of Association

It will now be shown that properly scaled mutual information and transfer entropies correspond with specific relevancies from the inquiry calculus. Also, the mutual information and transfer entropies, when reformulated in terms of relevancies, are the information theoretic equivalents of, respectively, correlation coefficients and partial correlation coefficients.

23.1 Relevancies as Measures of Association

We will now demonstrate how for known probability distributions the relevance measure may be used as a measure of association, or, equivalently, a measure of predictability.

Let $\mathbf{A}$ and $\mathbf{B}$ be the sets of propositions

$$\mathbf{A} = \{A_1, \dots, A_n\}$$

and

$$\mathbf{B} = \{B_1, \dots, B_m\},$$

and let

$$p(A_i, B_j) = p_{ij} \tag{23.1}$$

be a bivariate distribution, for $i = 1, \dots, n$ and $j = 1, \dots, m$, having marginal distributions

$$p(A_i) = \sum_{j=1}^{m} p(A_i, B_j) = \sum_{j=1}^{m} p_{ij} = p_{i+}, \tag{23.2}$$

and

$$p(B_j) = \sum_{i=1}^{n} p(A_i, B_j) = \sum_{i=1}^{n} p_{ij} = p_{+j}. \tag{23.3}$$

The product rule of relevancies is given as [104]:

$$d(\mathbf{A} \,|\, \mathbf{B}) = \frac{d(\mathbf{A} + \mathbf{B} \,|\, \mathbf{I})}{d(\mathbf{B} \,|\, \mathbf{I})}, \tag{23.4}$$

and the sum rule of relevancies is given as [59]:

$$d(\mathbf{A} + \mathbf{B} | \mathbf{I}) = d(\mathbf{A} | \mathbf{I}) + d(\mathbf{B} | \mathbf{I}) - d(\mathbf{A}\mathbf{B} | \mathbf{I}), \qquad (23.5)$$

where the central issue $\mathbf{I}$ is the joint question

$$\mathbf{A}\mathbf{B} = \{AB_{11}, AB_{12}, \ldots, AB_{nn}\}.$$

Substituting (23.5) into (23.4), we obtain the relevance of knowing the true proposition A_i in the set $\mathbf{A}$ in order to predict the true proposition B_j in the set $\mathbf{B}$:

$$d(\mathbf{A} | \mathbf{B}) = \frac{d(\mathbf{A} | \mathbf{I}) + d(\mathbf{B} | \mathbf{I}) - d(\mathbf{A}\mathbf{B} | \mathbf{I})}{d(\mathbf{B} | \mathbf{I})}, \qquad (23.6)$$

where [63]

$$d(\mathbf{A} | \mathbf{I}) = \frac{H(p_{i+})}{H(p_{ij})}, \qquad d(\mathbf{B} | \mathbf{I}) = \frac{H(p_{+j})}{H(p_{ij})}, \qquad d(\mathbf{A}\mathbf{B} | \mathbf{I}) = \frac{H(p_{ij})}{H(p_{ij})}, \quad (23.7)$$

and where

$$H(p_{ij}) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log \frac{1}{p_{ij}},$$

$$H(p_{i+}) = \sum_{i=1}^{n} p_{i+} \log \frac{1}{p_{i+}}, \qquad (23.8)$$

$$H(p_{+j}) = \sum_{j=1}^{m} p_{+j} \log \frac{1}{p_{+j}},$$

are the information entropies (25.1) of the probability distributions (23.1), (23.2), and (23.3), respectively.

   Combining (23.6) and (23.7), we may get the relevance of knowing the true A_i for the prediction of the true B_j, in terms of information entropy (25.1):

$$d(\mathbf{A} | \mathbf{B}) = \frac{H(p_{i+}) + H(p_{+j}) - H(p_{ij})}{H(p_{+j})}. \qquad (23.9)$$

Likewise, the relevance of knowing the true B_j for the prediction of the true A_i is found to be

$$d(\mathbf{B} | \mathbf{A}) = \frac{H(p_{i+}) + H(p_{+j}) - H(p_{ij})}{H(p_{i+})}. \qquad (23.10)$$

   We now demonstrate how a relevance of zero and one imply, respectively, in terms of predictability, absolute irrelevancy and absolute relevancy. Say, we have two bivariate distributions $p(A_i, B_j)$, Tables 23.1 and 23.2, respectively. In

	B_1	B_2	
A_1	$p_{11} = \frac{1}{4}$	$p_{12} = \frac{1}{4}$	$p_{1+} = \frac{1}{2}$
A_2	$p_{21} = \frac{1}{4}$	$p_{22} = \frac{1}{4}$	$p_{2+} = \frac{1}{2}$
	$p_{+1} = \frac{1}{2}$	$p_{+2} = \frac{1}{2}$	$p_{++} = 1$

Table 23.1: A and B unrelated

	B_1	B_2	
A_1	$p_{11} = \frac{1}{2}$	$p_{12} = 0$	$p_{1+} = \frac{1}{2}$
A_2	$p_{21} = 0$	$p_{22} = \frac{1}{2}$	$p_{2+} = \frac{1}{2}$
	$p_{+1} = \frac{1}{2}$	$p_{+2} = \frac{1}{2}$	$p_{++} = 1$

Table 23.2: A and B absolutely dependent

distribution 1 the variables A_i and B_j are independent, whereas in distribution 2 they are highly dependent.

By way of (25.1), we have for the unrelated distribution, Table 23.1:

$$H(p_{i+}) = \frac{1}{2} \log \frac{1}{1/2} + \frac{1}{2} \log \frac{1}{1/2} = \log 2$$

$$H(p_{+j}) = \frac{1}{2} \log \frac{1}{1/2} + \frac{1}{2} \log \frac{1}{1/2} = \log 2 \tag{23.11}$$

$$H(p_{ij}) = \frac{1}{4} \log \frac{1}{1/4} + \frac{1}{4} \log \frac{1}{1/4} + \frac{1}{4} \log \frac{1}{1/4} + \frac{1}{4} \log \frac{1}{1/4} = 2 \log 2.$$

Substituting (23.11) into (23.9), we find that variable A_i holds no relevance for variable B_j, that is, knowing which value variable A_i holds, gives us no information whatsoever as to what the value of variable B_j might be:

$$d(\mathbf{A}|\mathbf{B}) = \frac{\log 2 + \log 2 - 2 \log 2}{2 \log 2} = 0. \tag{23.12}$$

Likewise, substituting (23.11) into (23.10), we find that variable B_j holds no relevance for variable A_i, that is, knowing which value variable B_j holds, gives us no information, whatsoever, as to what the value of variable A_i might be:

$$d(\mathbf{B}|\mathbf{A}) = \frac{\log 2 + \log 2 - 2\log 2}{2\log 2} = 0. \tag{23.13}$$

So, the independence of A_i and B_j in Table 23.1 implies, (23.12) and (23.13):

$$d(\mathbf{A}|\mathbf{B}) = d(\mathbf{B}|\mathbf{A}) = 0. \tag{23.14}$$

By way of (25.1), we have for the absolutely dependent distribution, Table 23.2:

$$H(p_{i+}) = \frac{1}{2}\log\frac{1}{1/2} + \frac{1}{2}\log\frac{1}{1/2} = \log 2$$

$$H(p_{+j}) = \frac{1}{2}\log\frac{1}{1/2} + \frac{1}{2}\log\frac{1}{1/2} = \log 2 \tag{23.15}$$

$$H(p_{ij}) = \frac{1}{2}\log\frac{1}{1/2} + 0\log 0 + \frac{1}{2}\log\frac{1}{1/2} + 0\log\frac{1}{1/2} = \log 2.$$

Substituting (23.15) into (23.9), we find that variable A_i holds maximal relevance for variable B_j, that is, knowing which value variable A_i holds, gives us absolute information as to what the value of variable B_j will be:

$$d(\mathbf{A}|\mathbf{B}) = \frac{\log 2 + \log 2 - \log 2}{\log 2} = 1. \tag{23.16}$$

Likewise, substituting (23.15) into (23.10), we find that variable B_j holds maximal relevance for variable A_i, that is, knowing which value variable B_j holds, gives us absolute information as to what the value of variable A_i will be:

$$d(\mathbf{B}|\mathbf{A}) = \frac{\log 2 + \log 2 - \log 2}{\log 2} = 1. \tag{23.17}$$

So, for the dependence of A_i and B_j in Table 23.2, where both A_1 and B_1, and A_2 and B_2 imply each other, we find the maximum relevancies, (23.16) and (23.17):

$$d(\mathbf{A}|\mathbf{B}) = d(\mathbf{B}|\mathbf{A}) = 1. \tag{23.18}$$

Now, (23.14) and (23.18) are limit cases of relevancy in probability distributions. We give an intermediate case a third probability distribution in which variable A_i is absolutely relevant for the prediction of B_j, but not the other way around, Table 23.3.

	B_1	B_2	B_3	
A_1	$p_{11} = 0$	$p_{12} = 0$	$p_{13} = \frac{1}{4}$	$p_{1+} = \frac{1}{4}$
A_2	$p_{21} = \frac{1}{4}$	$p_{22} = 0$	$p_{23} = 0$	$p_{2+} = \frac{1}{4}$
A_3	$p_{31} = \frac{1}{4}$	$p_{32} = 0$	$p_{33} = 0$	$p_{3+} = \frac{1}{4}$
A_4	$p_{41} = 0$	$p_{42} = \frac{1}{4}$	$p_{33} = 0$	$p_{4+} = \frac{1}{4}$
	$p_{+1} = \frac{1}{2}$	$p_{+2} = \frac{1}{4}$	$p_{+3} = \frac{1}{4}$	$p_{++} = 1$

Table 23.3: A and B semi-absolutely dependent

By way of (25.1), we have for the semi-absolutely dependent distribution, Table 23.3:

$$H(p_{i+}) = \frac{1}{4}\log\frac{1}{1/4} + \frac{1}{4}\log\frac{1}{1/4} + \frac{1}{4}\log\frac{1}{1/4} + \frac{1}{4}\log\frac{1}{1/4} = 2\log 2$$

$$H(p_{+j}) = \frac{1}{2}\log\frac{1}{1/2} + \frac{1}{4}\log\frac{1}{1/4} + \frac{1}{4}\log\frac{1}{1/4} = \frac{3}{2}\log 2$$

$$\begin{aligned}
H(p_{ij}) = {}& 0\log 0 + 0\log 0 + \frac{1}{4}\log\frac{1}{1/4} \\
& + \frac{1}{4}\log\frac{1}{1/4} + 0\log 0 + 0\log 0 \\
& + \frac{1}{4}\log\frac{1}{1/4} + 0\log 0 + 0\log 0 \\
& + 0\log 0 + \frac{1}{4}\log\frac{1}{1/4} + 0\log 0 \\
= {}& 2\log 2.
\end{aligned} \tag{23.19}$$

Substituting (23.19) into (23.9), we find that variable A_i holds maximal relevance for the prediction of variable B_j:

$$d(\mathbf{A}\,|\,\mathbf{B}) = \frac{2\log 2 + \frac{3}{2}\log 2 - 2\log 2}{\frac{3}{2}\log 2} = 1. \tag{23.20}$$

But, substituting (23.19) into (23.10), we find that variable B_j is not maximally

relevant for the prediction of variable A_i:

$$d(\mathbf{B}|\,\mathbf{A}) = \frac{2\log 2 + \frac{3}{2}\log 2 - 2\log 2}{2\log 2} = \frac{3}{4}. \tag{23.21}$$

So, for the semi-absolutely dependent distribution, knowing which value variable A_i holds gives us absolute information, that is, certainty as to what the value of variable B_j will be, (23.20); i.e., A_1 implies B_3, A_2 and A_3 both imply B_1, and A_4 implies B_2. The other way around, in the B_j to A_i direction, absolute relevance is not achieved, (23.21); i.e., B_3 implies A_1, B_2 implies A_4, whereas B_1 is undecided as to whether A_2 or A_3 will occur.

Note that the relevance measure is much like a correlation coefficient for contingency tables. Though it is much more general than the ordinary correlation coefficient in that it can capture non-linear associations[1], like those given in Table 23.3. Moreover, it can capture the asymmetry between the relevancies in, respectively, the A_i to B_j direction and the B_j to A_i direction, (23.20) and (23.21).

23.2   Relevancies, Entropy, and Venn Diagrams

We now will take a closer look at the entropy and how it may be partitioned.

Say, we again have a bivariate distribution:

$$p(A_i, B_j) = p_{ij}, \tag{23.22}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, having marginal distributions

$$p(A_i) = \sum_{j=1}^{m} p(A_i, B_j) = \sum_{j=1}^{m} p_{ij} = p_{i+}, \tag{23.23}$$

and

$$p(B_j) = \sum_{i=1}^{n} p(A_i, B_j) = \sum_{i=1}^{n} p_{ij} = p_{+j}. \tag{23.24}$$

Then, in the case of dependencies, we may represent the entropy of this distribution as a Venn diagram of two overlapping circles, Figure 23.1.

The total entropy present in $p(A_i, B_j)$ is the aggregate of all three pieces of entropy in the Venn diagram:

$$H(p_{ij}) = a + b + c. \tag{23.25}$$

---

[1]Stated more forcefully, the relevance measure is a measure of true statistical independence, whereas concepts like decorrelation only describe independence up to second-order. Two variables can be uncorrelated, yet still dependent. This fact is usually poorly understood and it stems from the confusion between the common meaning of the word "uncorrelated", which we usually take to mean "independent", and the precise mathematical definition of the word "uncorrelated", which means that the covariance is of diagonal form [67].
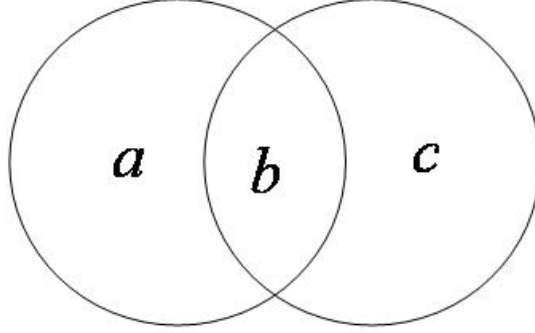
Figure 23.1: Venn Diagram of Entropy in Distribution p_{ij}

The entropy in the univariate marginal distributions, $p(A_i)$ and $p(B_j)$, then is:

$$H(p_{i+}) = a + b, \quad \text{and} \quad H(p_{+j}) = b + c. \quad (23.26)$$

The entropy wich results from an application of the sum rule, is the entropy shared by both $p(A_i)$ and $p(B_j)$:

$$H(p_{i+}) + H(p_{+j}) - H(p_{ij}) = a + 2b + c - (a + b + c) = b. \quad (23.27)$$

Substituting (23.26) and (23.27) in (23.9), we see that the relevance $d(\mathbf{A}|\mathbf{B})$ is just the ratio of the entropy shared by $p(A_i)$ and $p(B_j)$, and the total entropy in $p(B_j)$:

$$d(\mathbf{A}|\mathbf{B}) = \frac{H(p_{i+}) + H(p_{+j}) - H(p_{ij})}{H(p_{+j})} = \frac{b}{b+c}, \quad (23.28)$$

which is intuitive enough. If the circles in the Venn diagram do not overlap, that is, if $p(A_i)$ and $p(B_j)$ are independent, then $b = 0$ and

$$d(\mathbf{A}|\mathbf{B}) = \frac{0}{c} = 0,$$

and if the circles overlap totally, that is, if $p(A_i)$ and $p(B_j)$ are totally dependent, then $c = 0$ and

$$d(\mathbf{A}|\mathbf{B}) = \frac{b}{b} = 1.$$

23.3   Transfer Entropy

In the previous section, we discussed that relevancies can be used as measures of association on contingency tables, not much unlike a correlation coefficient. We will now show that the transfer entropy [87] is the information theoretical equivalent of the partial correlation coefficient.

In order to discuss transfer entropy, we need to introduce the trivariate distribution

$$p(A_i, B_j, C_k) = p_{ijk}, \tag{23.29}$$

which has marginal distributions

$$p(A_i, B_j) = \sum_{k=1}^{l} p(A_i, B_j, C_k) = \sum_{k=1}^{l} p_{ijk} = p_{ij+},$$

$$p(A_i, C_k) = \sum_{j=1}^{m} p(A_i, B_j, C_k) = \sum_{j=1}^{m} p_{ijk} = p_{i+k}, \tag{23.30}$$

$$p(B_j, C_k) = \sum_{i=1}^{n} p(A_i, B_j, C_k) = \sum_{i=1}^{n} p_{ijk} = p_{+jk},$$

and

$$p(A_i) = \sum_{j=1}^{m}\sum_{k=1}^{l} p(A_i, B_j, C_k) = \sum_{j=1}^{m}\sum_{k=1}^{l} p_{ijk} = p_{i++},$$

$$p(B_j) = \sum_{i=1}^{n}\sum_{k=1}^{l} p(A_i, B_j, C_k) = \sum_{i=1}^{n}\sum_{k=1}^{l} p_{ijk} = p_{+j+}, \tag{23.31}$$

$$p(C_k) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(A_i, B_j, C_k) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ijk} = p_{++k},$$

Then, in the case of dependencies, we may represent the entropy of this trivariate distribution as a Venn diagram of three overlapping circles, Figure 23.1.

The total entropy in the trivariate distribution (23.29) is the aggregate of all the seven pieces of entropy in the Venn diagram:

$$H(p_{ijk}) = a + b + c + d + e + f + g. \tag{23.32}$$

The entropies in the marginal distributions (23.30) and (23.31) are:

$$
\begin{aligned}
H(p_{i++}) &= a + b + d + e,\\
H(p_{+j+}) &= b + c + e + f,\\
H(p_{++k}) &= d + e + f + g,\\
H(p_{ij+}) &= a + b + c + d + e + f,\\
H(p_{i+k}) &= a + b + d + e + f + g,\\
H(p_{+jk}) &= b + c + d + e + f + g.
\end{aligned}
\tag{23.33}
$$
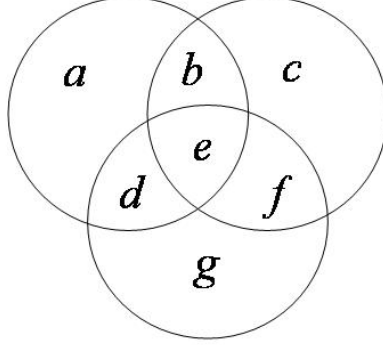
Figure 23.2: Venn diagram of entropy in dependent trivariate distribution p_{ijk}

The entropy shared by A_i, B_j, and C_k, that is, e, may be found as follows:

$$e = H(p_{i++}) + H(p_{+j+}) + H(p_{++k})$$
$$- H(p_{ij+}) - H(p_{i+k}) - H(p_{+jk}) + H(p_{ijk}), \qquad (23.34)$$

where the right hand side of 23.34 is the generalized sum rule of entropies [61].

Inspecting the Venn diagram in Figure 23.2, we see that the unique contribution of A_i in the prediction of C_k is represented by element d. Now, from the Venn diagram and (23.33), we may deduce the two following relations

$$a + d = H(p_{ij+}) - H(p_{+j+}) \qquad (23.35)$$

and

$$a = H(p_{ijk}) - H(p_{+jk}). \qquad (23.36)$$

From (23.35) and (23.36), it follows that element d, which may be interpreted as the unique contribution of A_i in the prediction of C_k, can be found through the relation:

$$d = -H(p_{+j+}) + H(p_{ij+}) + H(p_{+jk}) - H(p_{ijk}). \qquad (23.37)$$

This unique contribution of A_i in the prediction of C_k is also known as the transfer entropy [87]:

$$T(A|B,C) = -H(p_{+j+}) + H(p_{ij+}) + H(p_{+jk}) - H(p_{ijk}). \qquad (23.38)$$

We now define the mutual informations:

$$I(B,C) = e + f = H(p_{+j+}) + H(p_{++k}) - H(p_{+jk}) \qquad (23.39)$$

and

$$I(AB, C) = e = H(p_{ij+}) + H(p_{++k}) - H(p_{ijk}) \,. \qquad (23.40)$$

From (23.39) and (23.40), we have that the transfer entropy (23.38) may also be written as [67]

$$T(A|\, B, C) = d = I(AB, C) - I(B, C) \,. \qquad (23.41)$$

Switching to the notation of the previous section, we may define **C** to be the issue of interest. By doing so, we may obtain the relevancies of the anti-chained **AB** and **B** relative to this issue:

$$d(\mathbf{AB}|\, \mathbf{C}) = \frac{I(AB, C)}{H(p_{++k})} \,. \qquad (23.42)$$

and

$$d(\mathbf{B}|\, \mathbf{C}) = \frac{I(B, C)}{H(p_{++k})} \,. \qquad (23.43)$$

So, we see that the scaled transfer entropy may be interpreted as the unique relevance of **A** in regards to the prediction of **C**, since it is the relevance of **AB** for the prediction of **C** minus the relevance of **B** for the prediction of **C**, (23.4), (23.42), (23.43), and (23.41):

$$d\big(\mathbf{A}^{unique}\big|\, \mathbf{C}\big) = \frac{T(A|\, B, C)}{H(p_{++k})}$$

$$= \frac{I(AB, C)}{H(p_{++k})} - \frac{I(B, C)}{H(p_{++k})}$$

$$= d(\mathbf{AB}|\, \mathbf{C}) - d(\mathbf{B}|\, \mathbf{C}) \,.$$

23.4   Discussion

In this chapter we have computed relevancies for known probability distributions, as given in Table 23.4.

	b_1	b_2	
a_1	θ_1	θ_3	$\theta_1 + \theta_3$
a_2	θ_2	θ_4	$\theta_2 + \theta_4$
	$\theta_1 + \theta_2$	$\theta_3 + \theta_4$	$\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$

Table 23.4: Distribution 1

For example, if we wish to compute the relevance of b_j for a_i in Table 23.4, then, by way of the product and sum rule of information theory, the conditional

relevance is a function u of the theta's:

$$d(a \mid b) = u(\theta_1, \theta_2, \theta_3, \theta_4)$$
(23.44)
$$= \frac{H(\theta_1 + \theta_3, \theta_2 + \theta_4) + H(\theta_1 + \theta_2, \theta_3 + \theta_4) - H(\theta_1, \theta_2, \theta_3, \theta_4)}{H(\theta_1 + \theta_2, \theta_3 + \theta_4)},$$

where H is the Shannon's entropy:

$$H(\theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} \theta_i \log \frac{1}{\theta_i}.$$
(23.45)

However, we typically only have indirect access, by way of our count data, to the probabilities θ, Table 23.4.

	b_1	b_2	
a_1	r_1	r_3	$r_1 + r_3$
a_2	r_2	r_4	$r_2 + r_4$
	$r_1 + r_2$	$r_3 + r_4$	$r_1 + r_2 + r_3 + r_4 = n$

Table 23.5: Count data

Let $D = (r_1, r_2, r_3, r_4)$ be the observed count data in Table 23.4 and let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ be the vector of the unknown probabilities in Table 23.4. Then the likelihood function of the unknown probabilities is assumed to follow multinomial distribution:

$$L(\boldsymbol{\theta}) = p(D \mid \boldsymbol{\theta}) = \frac{(r_1 + r_2 + r_3 + r_4)!}{r_1! r_2! r_3! r_4!} \theta_1^{r_1} \theta_2^{r_2} \theta_3^{r_3} \theta_4^{r_4}$$
(23.46)

As a prior for the unknown probabilities we assign the uninformative Dirichlet prior

$$p(\boldsymbol{\theta}) \propto \theta_1^{-1} \theta_2^{-1} \theta_3^{-1} \theta_4^{-1},$$
(23.47)

which, if marginalized, collapses to the uninformative Beta prior. Combining the likelihood (23.46) and prior (23.47), by way of the product rule, and normalizing, by way of the sum rule, we obtain the multivariate Dirichlet posterior distribution of the theta's given the observed count data:

$$p(\boldsymbol{\theta} \mid D) = \frac{(r_1 + r_2 + r_3 + r_4 - 1)!}{(r_1 - 1)! (r_2 - 1)! (r_3 - 1)! (r_4 - 1)!} \theta_1^{r_1 - 1} \theta_2^{r_2 - 1} \theta_3^{r_3 - 1} \theta_4^{r_4 - 1}$$
(23.48)

Each realization of a relevance $u(\boldsymbol{\theta})$, (23.44), maps onto a corresponding probability $p(\boldsymbol{\theta} \mid D) \, d\theta$, (23.44). By arranging the values $u(\boldsymbol{\theta})$ on the x-axis and the corresponding $p(\boldsymbol{\theta} \mid D) \, d\boldsymbol{\theta}$ on the y-axis, we may obtain the univariate probability distribution of the relevance $u(\boldsymbol{\theta})$. This probability distribution of $u(\boldsymbol{\theta})$ takes the uncertainty into account we have with regard to the unknown $\boldsymbol{\theta}$ and, consequently, lets us put confidence bounds on this relevance.

If we only have four unknown theta's we may use brute computational force to partition the domain of $\boldsymbol{\theta}$ and compute of for each partitioning the corresponding pair $[u(\boldsymbol{\theta}), p(\boldsymbol{\theta}|D)\,d\boldsymbol{\theta}]$, after which we then order the $u(\boldsymbol{\theta})$ and plot them together with their corresponding probabilites $p(\boldsymbol{\theta}|D)\,d\boldsymbol{\theta}$. This can be done by way of Nested Sampling Monte Carlo sampling framework [91], even for large distributions that have many unknown $\boldsymbol{\theta}$'s, the curse of dimensionality notwithstanding. One possible implementation of the Nested Sampling framework is by way of the Inner Nested Sampling algorithm [108].

Chapter 24

Risk Communication

If there is the possibility of some danger, then information theory allows us to assign in a rational manner relevancies to statements made by officials in regards to that danger. It is found that the competence and trustworthiness of that official source is directly related its relevance. High competence and trustworthiness imply a corresponding high relevance, and low competence trustworthiness imply a low relevance. This mathematical derived result is in close correspondence with previous social scientific findings [95].

Since competence and trustworthiness turn out to be the necessary boundary conditions for relevance, we find that risk communicators, in order for their message to be effective (i.e., relevant), should not only focus themselves on the message itself, but also should take great care to manage their perceived competence and trustworthiness with the public at large. As it is rationality itself which, given a low perceived competence and trustworthiness of the risk communicators, dictates the public to disregard that which is communicated to them.

24.1 The Importance of Unbiasedness

Using information theory, or, equivalently, inquiry calculus, we will demonstrate the importance of a source of information being unbiased in order for it to be relevant.

Say, some disaster has occurred, like, for example, the nuclear accident in

Fukushima. Then the propositions we will use are

$$D = \text{Danger},$$
$$\overline{D} = \text{No danger},$$

$$W = \text{Warning},$$
$$\overline{W} = \text{All-clear},$$

$$U = \text{Unbiased},$$
$$\overline{U} = \text{Biased}.$$

The set $\mathbf{D} = \{D, \overline{D}\}$ constitutes the question whether there is danger or not, $Q_{\mathbf{D}}$; this question $Q_{\mathbf{D}}$ is the issue of interest. The set $\mathbf{W} = \{W, \overline{W}\}$ constitutes the warning signal[1] $\mathbf{W}$. The central issue $\mathbf{I}$ is the question $\mathbf{DW}$, which corresponds with the set of propositions:

$$\mathbf{DW} = \left\{ DW, D\overline{W}, \overline{D}W, \overline{D}\,\overline{W} \right\}.$$

If our source of information is unbiased, a warning will be given in case of danger, and an all-clear if there is no danger. Then we have the following probabilities:

$$p(W \,|\, DU) = 1,$$

$$p\big(\overline{W} \,\big|\, DU\big) = 0,$$

$$p\big(W \,\big|\, \overline{D}U\big) = 0,$$ (24.1)

$$p\big(\overline{W} \,\big|\, \overline{D}U\big) = 1.$$

If our source of information is biased, in that an all-clear will be given even if

---

[1]Questions and signals are equivalent entities, by way of the Cox definition of a question, [19]. A signal, just like a question, is defined by the set of possible messages it will communicate.

there is a clear and present danger, then we have the following probabilities:

$$p\big(W\,\big|\,D\overline{U}\big) = 0,$$

$$p\big(\overline{W}\,\big|\,D\overline{U}\big) = 1,$$

$$p\big(W\,\big|\,\overline{D}U\big) = 0,$$ (24.2)

$$p\big(\overline{W}\,\big|\,\overline{D}\,\overline{U}\big) = 1.$$

Personal prior probabilities are assigned to the possibility of there being a dangerous situation:

$$p(D) = d,$$

$$p\big(\overline{D}\big) = 1 - d.$$ (24.3)

Personal prior probabilities are also assigned to the possibility of the source of information being unbiased:

$$p(U) = u,$$

$$p\big(\overline{U}\big) = 1 - u.$$ (24.4)

Combining (24.1) through (24.4), by way of the product rule of probability theory (4.1), we find[2]:

$$p(WDU) = du, \qquad\qquad p\big(WD\overline{U}\big) = 0,$$

$$p\big(\overline{W}DU\big) = 0, \qquad\qquad p\big(\overline{W}D\overline{U}\big) = d\,(1-u),$$

$$p\big(W\overline{D}U\big) = 0, \qquad\qquad p\big(W\overline{D}\,\overline{U}\big) = 0,$$ (24.5)

$$p\big(\overline{W}\,\overline{D}U\big) = (1-d)\,u, \qquad p\big(\overline{W}\,\overline{D}\,\overline{U}\big) = (1-d)\,(1-u).$$

---

[2] For example, $p\,(WDU) = p(\,W\,|\,DU)\,p(D)\,p(U) = du$

By way of (24.5) and the sum rule of probability theory (4.2), we find[3] :

$$p(WD) = du,$$

$$p(\overline{W}D) = d\,(1 - u)\,,$$

$$p(W\overline{D}) = 0,$$ (24.6)

$$p(\overline{W}\,\overline{D}) = (1 - d)\,.$$

The central issue is $\mathbf{I} = \mathbf{DW}$. We want to find to relevance of the warning signal in relation to the dangerousness of the current situation. So, the issue of interest is $\mathbf{D}$, and we are looking for the conditional relevance of the signal $\mathbf{W}$, given by our source of information, in relation to this issue of interest. We have, (25.20),

$$d(\mathbf{W}|\,\mathbf{D}) = \frac{d(\mathbf{W} + \mathbf{D}|\,\mathbf{I})}{d(\mathbf{D}|\,\mathbf{I})},$$ (24.7)

where, (25.18),

$$d(\mathbf{W} + \mathbf{D}|\,\mathbf{I}) = d(\mathbf{W}|\,\mathbf{I}) + d(\mathbf{D}|\,\mathbf{I}) - d(\mathbf{WD}|\,\mathbf{I})\,.$$ (24.8)

Substituting (24.8) into (24.7), we obtain

$$d(\mathbf{W}|\,\mathbf{D}) = \frac{d(\mathbf{W}|\,\mathbf{I}) + d(\mathbf{D}|\,\mathbf{I}) - d(\mathbf{WD}|\,\mathbf{I})}{d\,(\mathbf{D}|\,\mathbf{I})}.$$ (24.9)

The right-hand relevancies in (24.9) can be found to be, (25.15),

$$d(\mathbf{W}|\,\mathbf{I}) = \frac{H\big[p(W)\,,p(\overline{W})\big]}{H\big[p(WD)\,,\ldots,p(\overline{W}\,\overline{D})\big]},\qquad d(Q_{\mathbf{D}}|\,Q_{\mathbf{I}}) = \frac{H\big[p(D)\,,p(\overline{D})\big]}{H\big[p(WD)\,,\ldots,p(\overline{W}\,\overline{D})\big]},$$

$$d(\mathbf{W} + \mathbf{D}|\,\mathbf{I}) = \frac{H\big[p(WD)\,,\ldots,p(\overline{W}\,\overline{D})\big]}{H\big[p(WD)\,,\ldots,p(\overline{W}\,\overline{D})\big]},$$ (24.10)

where (25.1), (24.3), (24.7), and (24.8),

$$H\big[p(W)\,,p(\overline{W})\big] = du\log\frac{1}{du} + (1 - du)\log\frac{1}{1 - du},$$

$$H\big[p(D)\,,p(\overline{D})\big] = d\log\frac{1}{d} + (1 - d)\log\frac{1}{1 - d},$$ (24.11)

$$H\big[p(WD)\,,\ldots,p(\overline{W}\,\overline{D})\big] = du\log\frac{1}{du} + (1 - d)\log\frac{1}{1 - d} + d\,(1 - u)\log\frac{1}{d\,(1 - u)}.$$

---

[3]For example, $p(WD) = p(WDU) + p\left(WD\overline{U}\right) = du$

Substituting (24.10) and (24.11) into (24.9), and making use of the logarithmic property,

$$c \log \frac{1}{c} = -c \log c,$$

we obtain, after some algebra,

$$d(\mathbf{W}|\mathbf{D}) = \frac{d \log d + (1-du) \log (1-du) - d(1-u) \log d (1-u)}{d \log d + (1-d) \log (1-d)}. \quad (24.12)$$

Inspecting (24.12), we see that as the probability of unbiasedness goes to one, that is, $u \to 1$, then the relevance of the source of information goes to one as well,

$$d(\mathbf{W}|\mathbf{D}) \to \frac{d \log d + (1-d) \log (1-d)}{d \log d + (1-d) \log (1-d)} = 1. \quad (24.13)$$

And as the probability of unbiasedness goes to zero, that is, $u \to 0$, then the relevance of the source of information goes to zero as well,

$$d(\mathbf{W}|\mathbf{D}) \to \frac{d \log d + \log 1 - d \log d}{d \log d + (1-d) \log (1-d)} = 0. \quad (24.14)$$

24.2   What Does It Mean?

Both the warning and the all-clear, that is, W and $\overline{W}$, are signals send by the risk communicators to the public at large. Let

$$p(D) = d \quad (24.15)$$

be the initial danger assessment, prior to receiving the signal, of a receiver. Then, in case of an optimal risk communication, this signal will modify the belief of the receiver to the extent that a warning will imply with certainty the presence of danger, that is,

$$p(D|W) = 1, \quad (24.16)$$

whereas an all-clear, in the case of an optimal risk communication, will imply with certainty the absence of danger, that is,

$$p(D|\overline{W}) = 0, \quad (24.17)$$

irrespective of the initial danger assessments (24.15).

The scenario where there is a possible biasedness in the direction of not giving a warning, even if there is a clear and present danger, is expressed in the probabilities (24.1) through (24.8). By way of (24.7), (24.8) and the product rule of probability theory, we find that, for such a bias, a warning is always communicated successfully[4]:

$$p(D|W) = \frac{p(WD)}{p(W)} = \frac{du}{du} = 1, \quad (24.18)$$

---

[4]For example, if the tobacco industry tells us that smoking may cause lung cancer, then we will be very inclined to believe them. We expect them to be biased, but only to the extent that they will try to deny any causal connection between smoking and lung cancer. So, if the tobacco industry says that smoking may cause lung cancer, then it is safe to assume that this is indeed the case.

as (24.18) satisfies the communication ideal (24.16).

However, in the case of an all-clear, by way of (24.7), (24.8) and the product rule of probability theory, we find that the risk communication is dependent upon the perceived unbiasedness of the source of information u:

$$p\big(D|\,\overline{W}\big) = \frac{p\big(\overline{W}D\big)}{p\big(\overline{W}\big)} = \frac{d\,(1-u)}{1-du}. \tag{24.19}$$

As the probability of unbiasedness goes to one, that is, $u \to 1$, then the relevance of the source of information goes to one, (24.13), while, (24.19),

$$p\big(D|\,\overline{W}\big) \to 0, \tag{24.20}$$

thus, satisfying the communication ideal (24.17). But as the probability of unbiasedness goes to zero, that is, $u \to 0$, then the relevance of the source of information goes to zero, (24.14), while, (24.19),

$$p\big(D|\,\overline{W}\big) \to d, \tag{24.21}$$

or, equivalently, (24.21) and (24.15),

$$p\big(D|\,\overline{W}\big) \to p\,(D). \tag{24.22}$$

And we see that for this particular scenario, a relevance of zero implies the inability of the all-clear signal to move the danger perception of the receiver away from its prior pre-signal state and in the direction of the communication ideal (24.17).

24.3   Truth or Dare

In *Truth or Dare in Japan*, Correspondents Report, November 5, 2011, by Mark Willacy, we can read the following account:

> "A Japanese government official was dared by a journalist to drink a glass of water taken from a puddle inside the Fukushima nuclear plant. But this wasn't some glowing green liquid concoction that would turn the hapless official into the Incredible Hulk. It was water from the basement of reactors 5 and 6 at Fukushima, both of which were shut down successfully after the tsunami hit the plant in March unlike three of the other reactors which each suffered meltdowns. The water had been purified but because of fears it was still slightly contaminated it was deemed too unsafe to release outside the grounds of the Fukushima plant [by the journalists gathered at the press conference]. The TV cameras showed Yasuhiro Sonoda's hands shaking as he poured the water into the glass. The government MP and parliamentary secretary to the cabinet looked like a man about to drink poison. He took a gulp, held the glass out once more for the assembled press, and then sculled the rest."

How relevant was Mr. Sonoda gesture for our risk assessment in terms of alleviating our fears regarding the dangers of the decontaminated water? Common sense would suggest not very much. And indeed, we can read in the same article:

> "The ABC North's Asia correspondent Mark Willacy watched this feat of daring and wonders whether it was a publicity stunt, or just a dare between testosterone-fuelled men."

We now will look at this incident from an information theoretic perspective.

The drinking of decontaminated water is just another way of giving the all-clear signal $\overline{W}$. If we look at the BBC report then it is stated that Mr. Sonoda had been challenged repeatedly in the course of a five hour press conference to prove that what he was saying was true, that the decontaminated water was safe for use around the Fukushima plant.

Now, seeing that Mr. Sonoda had committed himself to the standpoint that the decontaminated was safe and had subsequently been called out to prove so, we suspect that had Mr. Sonoda not drunk the water this would have led to a serious loss of face. Moreover, this also would have amounted to giving a warning signal as to the safety of the water, forcing the public to update their danger assessment, irrespective of their initial danger assessment, to (24.18),

$$p(D) = d \qquad \text{updated to} \qquad p(D|W) = 1, \tag{24.23}$$

which probably would have been detrimental to his political career.

So, we assign a low probability of Mr. Sonoda being unbiased, say, (24.4):

$$p(U) = u = 0.20. \tag{24.24}$$

Not being an expert on the decontamination process, however, we are on the fence in regards to the short term safety of the decontaminated water. Stated differently, it may or may not be safe. We simply do not know. So, we assign as the probability of danger

$$p(D) = d = 0.50. \tag{24.25}$$

Then, by substituting (24.24) and (24.25) into (24.12), we find the relevance of the act of either drinking or not drinking the decontaminated water to be

$$d(\mathbf{W}|\mathbf{D}) = 0.108. \tag{24.26}$$

And after we have observed Mr. Sonoda drinking the water, that is, giving us the all-clear signal $\overline{W}$, our prior danger assessment is updated somewhat, (24.19):

$$p(D) = 0.50 \qquad \text{is updated to} \qquad p(D|\overline{W}) = 0.444. \tag{24.27}$$

Note that had we assigned no credibility to the unbiassedness of Mr. Sonoda, that is, $p(U) = 0$, then the relevance $d(\mathbf{W}|\mathbf{D})$ would have been zero, (24.14), as

the reassurance of no danger $\overline{W}$ would have failed to modulate our prior danger assessments in any way whatsoever, (24.22).

Based on the little knowledge on radiation we do have, we are less doubtful in regards the long term dangers of the decontaminated water. As even small amounts of radiation will accumulate over time. So, we assign a much larger probability to the long term dangers[5]:

$$p(D) = d = 0.90. \tag{24.28}$$

The corresponding relevance then drops further to, (24.12):

$$d(\mathbf{W}|\mathbf{D}) = 0.065. \tag{24.29}$$

And our prior danger assessment is updated somewhat, (24.19):

$$p(D) = 0.90 \qquad \text{is updated to} \qquad p\left(D|\overline{W}\right) = 0.878. \tag{24.30}$$

A low relevance of a source of information implies the inability of that source to modulate the prior beliefs of those to which the information is communicated.

Conversely, if a source of information is unable to modulate the prior beliefs, then this implies low relevance for that source of information. The fact that the relevancies (24.26) and (24.27) do not approach 0 much faster, given the small effects on the respective posteriors, (24.27) and (24.30), can be explained by the fact that these relevance values also reflect the possibility that Mr. Sonado could have chosen not the drink the decontaminated water. This then would have amounted to the sending of a warning signal W, which would have been highly informative, (24.23).

Based on this information theoretical analysis, we see that Mr. Sonoda had maneuvered himself into a very unfortunate position. Had he not drunk the decontaminated water, then a clear danger flag would have been raised. On the other hand, Mr. Sonoda could hardly ever hope to convince the reporters of the safety of the water, even if he drank it himself, given the very real possibility in their minds of Mr. Sonoda being biased.

24.4   How to Be Relevant

The relevancies $d\left(\mathbf{W}|\mathbf{D}\right)$ for the biasedness scenario may be plotted as a function of the prior perceived danger $p\left(D\right)$. For a high probability of unbiasedness $p\left(U\right) = 0.90$, we find the relevance function (24.12), Figure 24.1. It can be seen that even in the case of great trust, or, equivalently, a high perceived likelihood of unbiasedness, the relevance of the source of information drops off as the perceived likelihood of danger $p\left(D\right)$ increases.

---

[5]As an aside, on 6 November, 2011, it was reported that Otsuka Norikazu (63), one of the main newscasters on Fuji TV had been diagnosed with acute lymphoblastic leukemia. In his morning program, Mr. Norikazu had been promoting Fukushima produce by eating them in the show. The Japanese annual incidence of adult cases of lymphoblastic leukemia is 1 in 100,000.
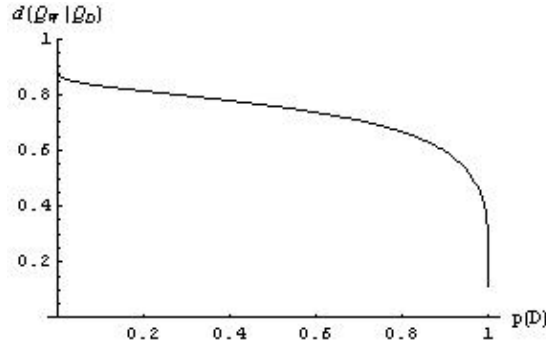
Figure 24.1: Relevance function for very high confidence in the unbiasedness of the source

The source of information remains relevant for small $p(D)$. But as the perceived likelihood of danger grows, the information theoretic equivalent of a (rational) panic sets in, and all reassurances of safety made by the trusted source of information are bound to fall on deaf ears, as the relevancy of these assurances eventually diminish to zero in the face of imminent danger.

The modified danger assessments $p\left(D|\,\overline{W}\right)$ may also be plotted as a function of the prior perceived danger $p(D)$. In Figure 24.2 we see that even in the case of great trust, that is, of a high perceived likelihood of unbiasedness, the modified danger assessment goes away from the communication ideal $p\left(D|\overline{W}\right)=0$, (24.17), as the perceived likelihood of danger $p(D)$ increases.
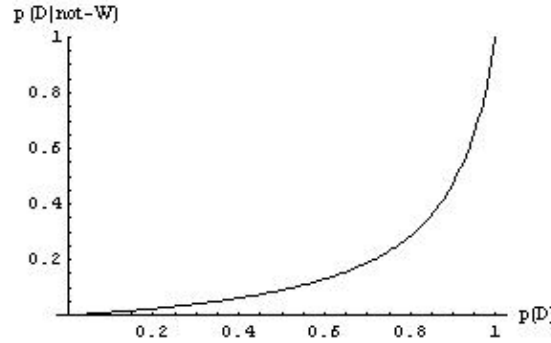
Figure 24.2: Danger assessment modification for very high confidence in the unbiasedness of the source

In Figures 24.3 and 24.4 there are given for a low probability of unbiasedness $p(U) = 0.10$, the relevance function (24.12) and the danger assessment modification function (24.19), respectively. In Figure 24.4 we see a demonstration of the fact that low relevancies, as those observed in Figure 24.3, imply the inability of the all-clear signal to move the danger perception of the receiver away from

its prior pre-signal state $p(D) = d$, and in the direction of the communication ideal $p\left(D|\,\overline{W}\right) = 0$.
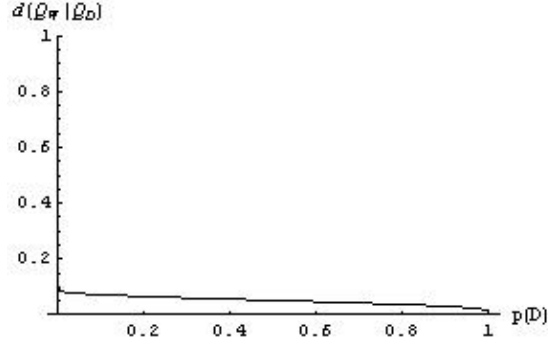
Figure 24.3: Relevance function for very low confidence in the unbiasedness of the source
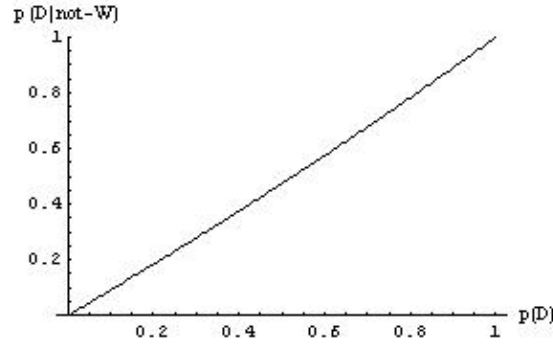

Figure 24.4: Danger assessment modification for very low confidence in the unbiasedness of the source

Comparing Figures 24.1 and 24.3, we see that there are two factors which determine the relevance of the source of information. The first factor is the perceived likelihood of unbiasedness $p(U)$. The second factor is the perceived likelihood of danger $p(D)$. Because even for a high trust in the unbiasedness of the source of information, large perceived likelihoods of danger may render that source of information irrelevant as panic sets in, that is, as $p(D) \to 1$.

The perceived likelihood of unbiasedness of the source of information $p(U)$ will typically be influenced by the actions of the source of information. These actions may entail giving full disclosure, taking full responsibility, and distancing oneself from any suggestion of a conflict of interest, and so on. And the perceived likelihood of danger $p(D)$ will typically be a function of the state of knowledge one has regarding the specifics of the danger.

For example, in the Fukushima nuclear accident the public expressed some

doubts on the unbiasedness of the government. It was said that the government had been slow to give full disclosure. Also, there were perceived to be strong ties between the nuclear industry and the government. Furthermore, as there is a relatively large dread for the dangers of radiation [96], most people will typically assign by default a high plausibility to the proposition that a nuclear accident entails a severe danger. Both these factors made it difficult for the Japanese government to effectively communicate to the public at large that the dangers were not that catastrophic.

24.5 The Importance of a Competent Source

Using information theory, or, equivalently, inquiry calculus, we will demonstrate the importance of a source of information being competent in order for it to be relevant.

We operationalize competence by way of the concept of false positives and false negatives. Let α be the probability of a false positive and β be the probability of a false negative. Then we have that the conditional probability of either a warning or an all-clear, given either danger or no danger, is

$$p(W \mid D) = 1 - \beta,$$

$$p(\overline{W} \mid D) = \beta,$$

$$p(W \mid \overline{D}) = \alpha,$$

$$(24.31)$$

$$p(\overline{W} \mid \overline{D}) = 1 - \alpha.$$

The personal prior probabilities which assigned to the possibility of there being a dangerous situation are again, (24.3),

$$p(D) = d,$$

$$(24.32)$$

$$p(\overline{D}) = 1 - d.$$

Combining (24.31) and (24.32), by way of the product rule of probability theory,

we have

$$p(WD) = (1 - \beta)\, d,$$

$$p(\overline{W}D) = \beta d,$$

$$p(W\overline{D}) = \alpha\,(1 - d)\,, \tag{24.33}$$

$$p(\overline{W}\,\overline{D}) = (1 - \alpha)\,(1 - d)\,,$$

and

$$p(W) = (1 - \beta)\, d + \alpha\,(1 - d)\,, $$

$$p(\overline{W}) = \beta d + (1 - \alpha)\,(1 - d)\,. \tag{24.34}$$

Note that the false positive and false negative scenario may pertain to medical tests, which are validated by their false positives and false negatives; weather and terror alarms, which tend to favor false positives over false negatives, by way of the fact that it is better to be safe than sorry; and so on. Furthermore, if we set the probability of a false positive to zero, that is,

$$\alpha = 0, \tag{24.35}$$

and if we interpret the probability of a false negative β as the probability of biasedness, that is,

$$\beta = 1 - u, \tag{24.36}$$

then the competence scenario collapses to the previous scenario in which biasedness played an important role. This can be seen by substituting both (24.35) and (24.36) into (24.33) and (24.34) and comparing the result with (24.7) and (24.8). So, biasedness and incompetence admit the same inference and relevance structure.

It was found for the biasedness scenario that as the probability of unbiasedness goes to zero, that is, $u \to 0$, then the corresponding relevance goes to, (24.14),

$$d\,(\mathbf{W}|\,\mathbf{D}) \to 0.$$

This result can be restated in terms of false positive and negatives, as follows. If the probability of a false positive is zero and the probability of a false negative goes to one, that is, $\beta \to 0$, the source of information will flat-line, in that only no warnings $\overline{W}$ can be given out, and hence the relevance of zero.

By going through the steps (24.7) through (24.12), we may compute the relevance of the of false positives and false negatives scenario, and may be checked, that if the source of information is as informative as a coin toss, that is, $\alpha = \beta = 0.5$, then its relevance goes to zero.

Just as in the case of $\alpha = 0$ and $\beta = 1$, where there is a bias to never give out a warning. Or, for that matter, $\alpha = 1$ and $\beta = 0$, which is the scenario in which there is a bias to over exaggerate the possibility of danger, by always giving out a warning.

Both cases $\alpha = \beta = 0$ and $\alpha = \beta = 1$ will result in relevancies of one. The latter being the case where it is known that the source of information always lies, which in the case of dichotomous propositions, then implies perfect information.

All other configurations of false positives α and false negatives β will result in relevancies greater than zero and smaller than one, that is,

$$0 < d\left(\mathbf{W}|\mathbf{D}\right) < 1.$$

24.6 Discussion

Slovic states that the limited effectiveness of risk-communication can be attributed to the lack of trust. If you trust the risk-manager, communication is relatively easy. If trust is lacking, no form or process of communication will be satisfactory [95]. In information theoretic terms this translates to the statement that if the probability of some source of information being unbiased and competent is low, then its relevance will also be low, where a low relevance implies an a priori inability to modulate our prior beliefs regarding some issue interest.

However, information theory shows us that there is a second factor at play in risk-communication, other than trustworthiness. This second factor is the perceived likelihood of danger. Because even for a high trust in the unbiasedness of a source of information, large perceived likelihoods of danger may render that trusted source of information still irrelevant, as one's own sense of danger overrides all the assurances of safety[6].

The perceived likelihood of unbiasedness of the source of information will typically be influenced by the actions of the source of information. If the source of information is the government itself, then these actions may entail giving full disclosure, taking full responsibility, and distancing oneself from any suggestion of a conflict of interest, and so on. And the perceived likelihood of danger will typically be a function of the state of knowledge one has regarding the specifics of the danger which is in play.

If a full and thorough understanding of the dangers involved requires some form of scientific training, then the plausibility that a lay person assigns to the proposition of there being a danger, typically, will be diffuse. Such diffuse plausibilities may be swayed either way, if some authoritative and unbiased source of information has some pertinent information to offer regarding these dangers. Examples of such 'opaque' dangers would be, for example, the dangers that flow from exposure to radiation or the dangers associated with climate change. In such cases, scientists from the respective fields, typically, fulfill the role of being the authoritative and unbiased sources of information.

---

[6]See for example Figure 24.1.

However, today we witness a rising general skepticism as to the unbiasedness of these experts. So, even if their perceived competence is high, by way of their scientific credentials, scientists now also will need to manage the perceived likelihood of them being unbiased in order to be heard by the public, that is, be relevant. Just like the governments have to do. Scientists may do this by giving full disclosure, distancing themselves from any suggestion of a conflict of interest, and by refraining from committing scientific fraud.

Chapter 25

A Short Historical Overview

Information theory is still a very young scientific discipline. The first rudimentary building blocks of information theory were laid in 1948, with Shannon's work on Information Entropy, and only very recently, in 2013, with the derivation of the information theoretic equivalent of Bayes' Theorem, do we have a information theoretic framework of any generality. We will give here a brief overview of information theory by introducing the most important information theoretical concepts in their chronological order of discovery.

25.1 The first phase, 1948-1951

Information theory started in 1948 with Shannon's formal derivation of the *information entropy* [88],

$$H(p_1 \cdots p_m) = p_1 \log \frac{1}{p_1} + \ldots + p_m \log \frac{1}{p_m}, \qquad (25.1)$$

as a measure of the amount of uncertainty in the probability distribution $(p_1 \cdots p_m)$.

In order to get some feeling for what constitutes the amount of uncertainty in a probability distribution, we need to discuss the building blocks that make up the information entropy H. The probability distribution

$$\mathbf{p} = (p_1, \ldots, p_m)$$

represents our state of knowledge in regards to the plausibility of each of the m possible outcomes we are considering, and the terms $\log \frac{1}{p_j}$, $j = 1, \ldots, m$, in (25.1) are called the 'surprises' [99]. If the kth outcome is almost certain, that is, if $p_k \to 1$, then the corresponding surprise goes to zero as we observe this outcome:

$$\text{surprise } k\text{th outcome} = \log \frac{1}{p_k} \to \log \frac{1}{1} = \log 1 = 0.$$

To illustrate, if the sun rises in the morning, then we will not be that surprised, being that the sun always rises. Alternatively, if the kth outcome is almost impossible, that is, if $p_k \to 0$, then the corresponding surprise will go to infinity:

$$\text{surprise } k\text{th outcome} = \log \frac{1}{p_k} \to \log \frac{1}{0} = \log \infty = \infty,$$

as we can imagine our surprise should the sun not rise. If the probability distribution $\mathbf{p}$ is our state of knowledge in regards to m possible outcomes of some event. Then Shannon's entropy H is the mean surprise we are expected to experience after the outcome of that event is presented to us, (25.1):

$$H(\mathbf{p}) = \sum_{j=1}^{m} p_j \log \frac{1}{p_j} = E\left(\log \frac{1}{p}\right) = \text{mean (surprise)} .$$

This is what is meant with measuring the amount of uncertainty in a probability distribution, (25.1).

For example, if the chances of recovery of a terminal patient are negligible with a probability of effectively 0 and the chance of death is 1, then the amount of uncertainty in regards to the outcome is zero:

$$H = 0 \log \frac{1}{0} + 1 \log \frac{1}{1} = \log 1 = 0. \tag{25.2}$$

Now, if the chances of recovery are fifty-fifty then the amount of uncertainty in regards to the outcome of this event is

$$H = \frac{1}{2} \log \frac{1}{1/2} + \frac{1}{2} \log \frac{1}{1/2} = \log 2 = 0.693. \tag{25.3}$$

For two-outcome events, (25.2) and (25.3) give, respectively, the minimum and maximum possible entropies H. The latter is in accordance with our intuition of uncertainty; the statement 'fifty-fifty' is our way of saying that we do not have any clue whatsoever as to the eventual outcome of some two-valued event.

In general, if we have m equally probable outcomes, then the amount of certainty becomes

$$H = \frac{1}{m} \log \frac{1}{1/m} + \ldots + \frac{1}{m} \log \frac{1}{1/m} = \log m. \tag{25.4}$$

So, as the number m of equally probable outcomes of some event goes to infinity, our uncertainty in regards to the specific outcome will also go to infinity:

$$H = \log m \to \log \infty = \infty, \quad \text{as } m \to \infty. \tag{25.5}$$

Property (25.5) is in accordance with our intuition on uncertainty. As the number of equally probable possible outcomes increases, the more uncertain we become in regards to the eventual outcome.

25.2 The second phase, 1951-2002

Information theory may be used to quantify the information gain of performing some experiment, or, equivalently, asking some question or performing some medical test. This is done by generalizing the uncertainty measure H, (25.1), to the *cross-entropy* measure H' [72]:

$$H'(\mathbf{p}|\,\boldsymbol{\pi}) = p_1 \log \frac{p_1}{\pi_1} + \cdots + p_m \log \frac{p_m}{\pi_m}. \qquad (25.6)$$

The cross-entropy is a measure of the distance between the prior distribution, $(\pi_1 \cdots \pi_m)$, describing our initial ignorance, and the posterior distribution, $(p_1 \cdots p_m)$, describing the information which the data of the experiment has given us.

The information gain of performing some experiment may be quantified as follows: the experiment that promises to be the most informative is the one which is expected to generate the greatest mean cross-entropy $E(H')$. In order to better illustrate the concept of information gain in an experiment, we now derive the mean cross-entropy by way of a medical example.

Let

$$\mathbf{A} = \{A_1, \ldots, A_m\}$$

be a set of competing medical diagnoses from which we must choose the correct one and let

$$\mathbf{D}^{(k)} = \left\{ D_1^{(k)}, \ldots, D_{n_k}^{(k)} \right\}$$

be the set of possible test results for the kth medical test.

Cox's mathematical operationalization of a question is that a question is the set of all possible statements that will answer that question [19]. So, the question as to which diagnosis to make may be represented by the set $\mathbf{A}$. By analogy, the mathematical operationalization of a medical test is that a medical test is the set of all possible test results that may be returned by that test. It follows that the kth medical test may be represented by the set $\mathbf{D}^{(k)}$.

Let $P(A_j)$ be the prior probability we assign to the diagnosis A_j and let $P\left(A_j|\,D_i^{(k)}\right)$ be the posterior probability of diagnosis A_j in case we observe the test result $D_i^{(k)}$ after having administered the kth medical test $\mathbf{D}^{(k)}$. Then the cross-entropy associated with the specific test result $D_i^{(k)}$ is given as, (25.6),

$$H'\left(D_i^{(k)}\right) = \sum_j p\left(A_j|\,D_i^{(k)}\right) \log \frac{p\left(A_j|\,D_i^{(k)}\right)}{p(A_j)}. \qquad (25.7)$$

Note that in the above notation, we emphasize that H' is the cross-entropy that results from the posterior which follows from observing the specific test result $D_i^{(k)}$. But it is to be understood that entropies are measures on probability distributions which, as a consequence, take as their arguments whole probability distributions.

Let $p\left(D_i^{(k)}\middle|A_j\right)$ be the likelihood of A_j, or, equivalently, the probability of the test results $D_i^{(k)}$ of the kth medical test, given that the diagnosis A_j should hold. Then the marginal probabilities of the test results are given as

$$p\left(D_i^{(k)}\right) = \sum_j p\left(A_j D_i^{(k)}\right) = \sum_j p(A_j)\, p\left(D_i^{(k)}\middle|A_j\right) \qquad (25.8)$$

and the mean cross-entropy for a given medical test $\mathbf{D}^{(k)}$ is given as the probability weighted sum of cross-entropies for this medical test, (25.7) and (25.8):

$$E\left[H'\left(D^{(k)}\right)\right] = \sum_{i=1}^{n_k} p\left(D_i^{(k)}\right) H'\left(D_i^{(k)}\right). \qquad (25.9)$$

If we compute (25.9) for all the K medical tests, then we get K mean cross-entropies:

$$E\left[H'\left(D^{(1)}\right)\right], E\left[H'\left(D^{(2)}\right)\right], \ldots, E\left[H'\left(D^{(K)}\right)\right],$$

and the medical test $\mathbf{D}^{(k)}$ which has the largest mean cross-entropy is the test that promises to give us, on average, the most information regarding the medical diagnoses $\mathbf{A}$ [89].

The mean cross-entropy (25.9) may be rewritten as a sum of Shannon information entropies (25.1). By way of (25.7), (25.8), (25.9), we have

$$E\left[H'\left(D^{(k)}\right)\right] = \sum_i p\left(D_i^{(k)}\right) \sum_j p\left(A_j\middle|D_i^{(k)}\right) \log \frac{p\left(A_j\middle|D_i^{(k)}\right)}{p(A_j)}$$

$$= \sum_i \sum_j p\left(A_j D_i^{(k)}\right) \log \frac{p\left(A_j D_i^{(k)}\right)}{p\left(D_i^{(k)}\right) p(A_j)}$$

$$= \sum_i \sum_j p\left(A_j D_i^{(k)}\right) \log \frac{1}{p(A_j)} + \sum_i \sum_j p\left(A_j D_i^{(k)}\right) \log \frac{1}{p\left(D_i^{(k)}\right)}$$

$$- \sum_i \sum_j p\left(A_j D_i^{(k)}\right) \log \frac{1}{p\left(A_j D_i^{(k)}\right)}. \qquad (25.10)$$

Furthermore, we also have

$$\sum_i \sum_j p\left(A_j D_i^{(k)}\right) \log \frac{1}{p(A_j)} = \sum_j \left[\sum_i p\left(D_i^{(k)}\Big| A_j\right)\right] p(A_j) \log \frac{1}{p(A_j)}$$

$$= \sum_j p(A_j) \log \frac{1}{p(A_j)} \qquad (25.11)$$

and

$$\sum_i \sum_j p\left(A_j D_i^{(k)}\right) \log \frac{1}{p\left(D_i^{(k)}\right)} = \sum_i \log \frac{1}{p\left(D_i^{(k)}\right)} \sum_j p\left(A_j D_i^{(k)}\right)$$

$$(25.12)$$

$$= \sum_i p\left(D_i^{(k)}\right) \log \frac{1}{p\left(D_i^{(k)}\right)}.$$

So, we may rewrite (25.10), by way of (25.11) and (25.12), as

$$E\left[H'\left(D^{(k)}\right)\right] = \sum_j p(A_j) \log \frac{1}{p(A_j)} + \sum_i p\left(D_i^{(k)}\right) \log \frac{1}{p\left(D_i^{(k)}\right)}$$

$$- \sum_i \sum_j p\left(A_j D_i^{(k)}\right) \log \frac{1}{p\left(A_j D_i^{(k)}\right)}. \qquad (25.13)$$

Then by way of (25.1) and (25.13), we have

$$E\left[H'\left(D^{(k)}\right)\right] = H[p(A_j)] + H\left[p\left(D_i^{(k)}\right)\right] - H\left[p\left(A_j D_i^{(k)}\right)\right]. \qquad (25.14)$$

The mean cross-entropy in the form of (25.14) is also known as the Mutual Information (MI) [63].

We summarize, the more general definition of entropy, (25.6), through equations (25.7), (25.8), (25.9), (25.10), has given us expression (25.14), which is a function of the initial information entropies (25.1). In the next section we will give a further generalization. This generalization will allow us to find (25.9) in an algorithmic manner, by combining information entropies.

25.3 The third phase, 2002-now

Information theory has entered a new phase with the work of Knuth, [58, 59, 61, 62, 63, 64, 65, 104]. In Knuth's Inquiry Calculus relevancies are assigned to questions, or, equivalently, tests or experiments. These relevancies are always defined to relative to some central issue **I**. The central issue is the baseline in

that it is the question which, when answered, will fill in all the unknowns. Also, the central issue is not necessarily the issue of interest.

For example, in the previous paragraph the central issue was the joint question $\mathbf{AD}^{(k)}$:

Which of the states $\left\{ A_1 D_1^{(k)}, A_1 D_2^{(k)}, \dots, A_m D_{n_k}^{(k)} \right\}$ will be true if we perform the kth medical test?

and the issue of interest was the question $\mathbf{A}$:

Which of the medical diagnoses $\{A_1, \dots, A_m\}$ is true?

Let $\mathbf{D}^{(k)}$ be the question:

Which of the test results $\left\{ D_1^{(k)}, \dots, D_{n_k}^{(k)} \right\}$ will we observe if we perform the kth medical test?

Then the relevancies of the questions $\mathbf{A}$, $\mathbf{D}^{(k)}$, and $\mathbf{AD}^{(k)}$, relative to the central issue $\mathbf{I} = \mathbf{AD}^{(k)}$, are defined as follows [63]:

$$d(\mathbf{A} \,|\, \mathbf{I}) = d\left(\mathbf{A} \,\middle|\, \mathbf{AD}^{(k)} \right) = \frac{H[p(A_j)]}{H\left[p\left(A_j \, D_i^{(k)} \right) \right]} \tag{25.15}$$

and

$$d\left(\mathbf{D}^{(k)} \,\middle|\, \mathbf{I} \right) = d\left(\mathbf{D}^{(k)} \,\middle|\, \mathbf{AD}^{(k)} \right) = \frac{H\left[p\left(D_i^{(k)} \right) \right]}{H\left[p\left(A_j \, D_i^{(k)} \right) \right]}, \tag{25.16}$$

and

$$d\left(\mathbf{AD}^{(k)} \,\middle|\, \mathbf{I} \right) = d\left(\mathbf{AD}^{(k)} \,\middle|\, \mathbf{AD}^{(k)} \right) = \frac{H\left[p\left(A_j \, D_i^{(k)} \right) \right]}{H\left[p\left(A_j \, D_i^{(k)} \right) \right]}, \tag{25.17}$$

where it is to be understood that the information entropies H, (25.1), are measures on probability distributions; for example,

$$H[p(A_j)] = \sum_j p(A_j) \log \frac{1}{p(A_j)}.$$

Questions are the collection of all the statements that will answer that question [59]. relevancies d may be assigned to questions by computing entropies H of the probability distributions p which assign probabilities to all the statements that will answer these questions. These entropies H are then scaled by the entropy of the central issue $\mathbf{I}$, which is the joint question that is constructed by taking the meet of all the questions. But how do we operate on these relevancies? Thanks to the efforts of Knuth, sum and product rules for relevancies of questions have been derived, as the sufficient and necessary operators of relevancies. The sum rule for relevancies was already given in 2002 [59]. But the

product rule, though already conjectured in 2002 [59], was only formally derived in 2013 [104].

Ever since 1774, with Laplace's discovery of the Bayes' Theorem, Bayesians have been able to effortlessly operate on probabilities by way of the product and sum rules [46]. This ease of use and generality has now been extended to information theory, as we now will demonstrate. The sum rule of relevancies is given as [59]

$$d\left(\mathbf{A} + \mathbf{D}^{(k)} \middle| \mathbf{I}\right) = d(\mathbf{A} | \mathbf{I}) + d\left(\mathbf{D}^{(k)} \middle| \mathbf{I}\right) - d\left(\mathbf{A}\mathbf{D}^{(k)} \middle| \mathbf{I}\right), \tag{25.18}$$

where the '+' is the join of the questions. The product rule of relevancies is given as [104]

$$d(\mathbf{A} | \mathbf{I}) \, d\left(\mathbf{D}^{(k)} \middle| \mathbf{A}\right) = d\left(\mathbf{D}^{(k)} \middle| \mathbf{I}\right) d\left(\mathbf{A} | \mathbf{D}^{(k)}\right) = d\left(\mathbf{A} + \mathbf{D}^{(k)} \middle| \mathbf{I}\right), \tag{25.19}$$

From this product rule then follows the Bayes' Theorem of information theory:

$$d\left(\mathbf{D}^{(k)} \middle| \mathbf{A}\right) = d\left(\mathbf{D}^{(k)} \middle| \mathbf{I}\right) \frac{d\left(\mathbf{A} | \mathbf{D}^{(k)}\right)}{d(\mathbf{A} | \mathbf{I})} = \frac{d\left(\mathbf{A} + \mathbf{D}^{(k)} \middle| \mathbf{I}\right)}{d(\mathbf{A} | \mathbf{I})}, \tag{25.20}$$

where it is to be noted that $d\left(\mathbf{D}^{(k)} \middle| \mathbf{A}\right)$ is the relevance we are looking for, namely, the relevance of the kth medical test $\mathbf{D}^{(k)}$ for the answering of the diagnosis question A.

By substituting the relevancies (25.15), (25.16), and (25.17), into the sum rule (25.18), we find

$$d\left(\mathbf{A} + \mathbf{D}^{(k)} \middle| \mathbf{I}\right) = \frac{H[p(A_j)] + H\left[p\left(D_i^{(k)}\right)\right] - H\left[p\left(A_j \, D_i^{(k)}\right)\right]}{H\left[p\left(A_j \, D_i^{(k)}\right)\right]}. \tag{25.21}$$

Substituting (25.15) and (25.21) into (25.20), that is, the reshuffled product rule (25.19), we obtain

$$d\left(\mathbf{D}^{(k)} \middle| \mathbf{A}\right) = \frac{H[p(A_j)] + H\left[p\left(D_i^{(k)}\right)\right] - H\left[p\left(A_j \, D_i^{(k)}\right)\right]}{H[p(A_j)]}. \tag{25.22}$$

Substituting (25.10) into (25.22), we see that the relevance of interest is just the scaled mean cross-entropy, which was earlier obtained by going through a rather torturous line of reasoning, equations (25.7) through (25.9):

$$d\left(\mathbf{D}^{(k)} \middle| \mathbf{A}\right) = \frac{E\left[H'\left(D^{(k)}\right)\right]}{H[p(A_j)]}. \tag{25.23}$$

Note that the scale $H[p(A_j)]$ is the same for all the K tests. So, the ordering between the expected values of cross-entropy,

$$E\left[H'\left(D^{(1)}\right)\right], E\left[H'\left(D^{(2)}\right)\right], \ldots, E\left[H'\left(D^{(K)}\right)\right],$$

is equivalent to an ordering in conditional relevancies,

$$d\left(\mathbf{D}^{(1)}\middle|\mathbf{A}\right), d\left(\mathbf{D}^{(2)}\middle|\mathbf{A}\right), \ldots, d\left(\mathbf{D}^{(K)}\middle|\mathbf{A}\right).$$

It follows that we may reinterpret (25.10) as choosing that test $\mathbf{D}^{(k)}$, which has the highest relevance relative to the issue of interest $\mathbf{A}$. This then concludes our introduction into the new and extended information theory.

25.4   Discussion

We have demonstrated how each phase in the development of information theory has been a steady generalization of the principles that were found earlier. Cross entropy is a generalization of the information entropy that came before it. While experimental design by way of mean cross entropy maximization is equivalent to choosing that experiment that has the highest relevancy for some issue of interest. It has also been demonstrated that the maximization of the mean cross entropy is equivalent to choosing that experiment which has the highest relevancy. And it follows that Knuth's inquiry calculus provides an alternative theoretical validation for the approach in which experiments are designed so as to maximize the mean cross entropy.

For a practical example of the traditional mean cross entropy maximization approach, we refer the interested reader to [66].

Chapter 26

Discussion

An important impetus for the development of the new inquiry calculus was that this calculus promised to enable intelligent machines to ask questions, as an automation of inquiry will allow robots to perform science in the far reaches of our solar system and in other star systems by enabling them to decide which question to ask, which experiment to perform, or which measurement to take given what they have learned and what they are designed to understand [60].

But whether or not the the new inquiry calculus formulation, in terms of experimental design, actually adds any new content to the traditional mean cross entropy maximization approach remains to be seen. In this regard we are not much unlike the first students of Laplace's probability theoretical work. We would imagine that those students could not possibly conceive the incredible range and sophistication of the applications which those two simple and modest rules of probability calculus, the product and sum rules, would give rise to. So, as to the total breath of the Bayesian information theory, we think it very likely to have only scratched the surface in this thesis.

For example, we might conceive in which we assign bin sizes to multivariate histograms on count data by way of a maximization of relevances conditional on some issue of interest, rather than by way of model selection [67]. The rational for such an approach would be provided by the fact that the multivariate histogram that has the greatest relevancy for some issue interest is the histogram that is conditionally the most informative for that issue. But either way, we can already state with certainty that Knuth's inquiry calculus provides us with a vastly more intuitive and, consequently, deeper understanding into the nature of information theory, as we patiently await the future.

Bibliography

[1] Aczel J.: Lectures on Functional Equations and Their Applications, Academic Press, New York, (1966).

[2] Allais M.: Fondements d'une Theorie Positive des Choix Comportant un Risque et Critique des Postulates et Axiomes de l'Ecole Americaine, Colloque Internationalle du Centre National de la Recherche Scientifique, No. 36, (1952).

[3] Allais M.: Le Comportement de l'Homme Rationel devant le Risque, Critique des Postulates et Axiomes de l'Ecole Americaine, Econometrica 21,503-546, (1953).

[4] Allais M.: L'Extension des Theories de l'Equilibre Economique General et du Rendement Social au Cas du Risque, Econometrica, 21, 269-290, (1953) .

[5] Allais M.: An Outline of My Main Contributions To Economic Science, Nobel Lecture, December 9, (1988).

[6] Bernoulli D.: Exposition of a New Theory on the Measurement of Risk. Translated from Latin into English by Dr Louise Sommer from 'Specimen Theoriae Novae de Mensura Sortis', Commentarii Academiae Scientiarum Imperialis Petropolitanas, Tomus V, 175-192, (1738).

[7] Bernstein P.L.: Against the gods: The Remarkable Story of Risk. John Wiley & Sons, Inc, (1996).

[8] Bernstein W.: A Splendid Exchange; How Trade Shaped the World, Grove Atlantic Ltd., (2008).

[9] Bleichrodt H. and Pinto J.L.: A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis. Management Science, 46, 1485-1496, (2000).

[10] Bloomfield R.: Behavioural Finance. In S.N. Durlauf & L.E. Blume (eds.), The new Palgrave dictionary of economics (2nd Ed). Macmillan Publishers, Ltd, (2008).

[11] Booij A.S., van Praag B.M.S. and Van de Kuilen G.: A Parametric Analysis of Prospect Theorys Functionals for the General Population. Vol. 68, Issue 1, 115148, (2010).

[12] Brentano F.: Psychologie vom Empirischen Standpunkt, Vol. 1, Dunker and Hunblot, Leipzig, (1874).

[13] Bretthorst L.G.: Bayesian Spectrum Analysis and Parameter Estimation. Springer-Verlag: New York, NY, USA; Volume 48, (1988).

[14] Bucholz R.H. and Ralph H.: Perfect Pyramids, Bulletin Australian Mathematical Society, vol. 45, no. 3, (1992).

[15] Camerer C. and Weber M.: Recent Developments in Modeling Preferences: Uncertainty and Ambiguity. Journal of Risk and Uncertainty, 5, 325370, (1992).

[16] Center J.L.: Inquiry Calculus and Information Theory. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Oxford MS, USA 2009 (eds. Goggans P and Chun CY) AIP Conference Proceedings, American Institute of Physics, New-York. 69-78, (2010).

[17] Cox R.T.: Probability, Frequency and Reasonable Expectation, American Journal of Physics, 14, 1-13, (1946).

[18] Cox R.T.: The Algebra of Probable Inference, John Hopkins University Press, Baltimore MD, (1961).

[19] Cox R.T.: Of Inference and Inquiry, an Essay in Inductive Logic. In The Maximum Entropy Formalism, ed by Levine R.D. and Tribus M., The MIT Press, Cambridge, MA, pp.119-167, (1979).

[20] De Finetti B.: Theory of Probability, 2 vols., J. Wiley and Sons, Inc., New York, (1974).

[21] Durlauf S.N. and Blume L.E., eds.: Behavioural and Experimental Economics. Palgrave Macmillan, (2010).

[22] Edwards W.: The Theory of Decision Making, Psychological Bulletin, Vol 51., No. 4, (1954).

[23] Ellis P.L.: On the Foundations of the Theory of Probability, Camb. Phil. Soc. viii, (1842).

[24] Ellsberg D.: Risk, Ambiguity, and the Savage Axioms, Quarterly Journal of Economics 75, 643-699, (1961).

[25] Fancher R.E.: Pioneers of Psychology, W. W. Norton and Company, London, (1990).

[26] Fechner G.J.: Elemente der Psychophysik, 2 vols.; Vol. 1 translated as Elements of Psychophysics, Boring, E.G. and Howes, D.H., eds. Holt, Rinehart and Winston, New York, (1966).

[27] Fennema H. and Wakker P.: Original Cumulative Prospect Theory: a Discussion of Empirical Differences, Journal of Behavioral Decision Making, Vol. 10, 53-64, (1997).

[28] Georgescu-Roegen N.: Utility, Expectations, Measurability, and Prediction, Paper read Econometric Soc., September (1953).

[29] Go S.C.: Marine Insurance in the Netherlands 1600-1870: A Comparative Institutional Approach, Groningen, (2009).

[30] Goldstein W.M. and Einhorn H.J.: Expression Theory and the Preference Reversal Phenomena. Psychological Review, 94, 236-254, (1987).

[31] Good I.J.: Probability and the Weighing of Evidence, C. Griffin and Co., London, (1950).

[32] Good I.J.: The Contributions of Jeffreys to Bayesian Statistics, in Zellner, A. ed., Bayesian Analysis in Econometrics and Statistics, North-Holland Pub. Co., Amsterdam, (1980).

[33] Guala F. (2008). Experimental Economics, History of. In S.N. Durlauf & L.E. Blume (eds.), The new Palgrave dictionary of economics (2nd Ed). Macmillan Publishers, Ltd.

[34] Gul, F. (2008). Behavioural Economics and Game Theory. In S.N. Durlauf & L.E. Blume (eds.), The new Palgrave dictionary of economics (2nd Ed). Macmillan Publishers, Ltd.

[35] Heath C. and Tversky A.: Preference and Belief: Ambiguity and Competence in Choice Under Uncertainty, Journal of Risk and Uncertainty 4, 5-28, (1991).

[36] Herrnstein R.J.: Relative and Absolute Strengths of Response as a Function of Frequency of Reinforcement, Journal of the Experimental Analysis of Behavior, 4, 267-72, (1961).

[37] Hogg R.V. and Tanis E.A.: Probability and Statistical Inference, Prentice Hall, Upper Saddle River, New Jersey, (2001).

[38] Hudson M.: The New Road to Serfdom; An Illustrated Guide to the Coming Real Estate Collapse, Harper's Magazine, (May 2006).

[39] Hurwicz L.: What Has Happened to the Theory of Games, Cowles Foundation Paper 75b, reprinted from American Economic Review, 43(2), (1953).

[40] Jaynes E.T.: How Does the Brain Do Plausible Reasoning?, Stanford University Microwave Laboratory Report 421, (1957).

[41] Jaynes E.T.: Foundations of Probability Theory and Statistical Mechanics, in Bunge M., ed., Delaware Seminar in Foundations of Physics, Springer-Verlag, Berlin, (1967).

[42] Jaynes E.T.: Prior Probabilities, IEEE Transactions on Systems Science and Cybernetics, SSC-4, pp. 227-241, (1968).

[43] Jaynes E.T.: The Well-Posed Problem, Found. Phys. 3, 477, (1973).

[44] Jaynes E.T.: Confidence Intervals vs Bayesian Intervals, W.L. Harper and C.A. Hooker, eds. Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Reidel Publishing Co., Dordrecht, Holland, (1976).

[45] Jaynes E.T.: Where Do We Stand on Maximum Entropy?, in Levine R.D. and Tribus M., eds., The Maximum Entropy Formalism, M.I.T. Press, Cambridge MA, (1978).

[46] Jaynes E.T.: Bayesian Methods: General Background. In the Proceedings Volume, Maximum Entropy and Bayesian Methods in Applied Statistics, J. H. Justice, Editor, Cambridge University Press, pp. 1-25, (1985).

[47] Jaynes E.T.: Probability Theory; the Logic of Science, Cambridge University Press, (2003).

[48] Jeffreys H.: Theory of Probability Theory, Clarendon Press, Oxford, (1939).

[49] Johnson-Laird P.N. and Wason P.C.: A Theoretical Analysis of Insight into a Reasoning Task, Cognitive Psychology, 1, 134-148, (1970).

[50] Jorion P.: Value at Risk: The New Benchmark for Managing Financial Risk, McGraw-Hill, (2006).

[51] Kahneman D. and Tversky A.: Subjective Probability: a Judgment of Representativeness, Cognitive Psychology 3, 430-454, (1972).

[52] Kahneman D. and Tversky A.: On the Psychology of Prediction, Psychological Review, Vol. 80, No. 4, (1973).

[53] Kahneman D. and Tversky A.: Prospect Theory: an Analysis of Decision Under Risk, Econometrica, 47(2), 263-291, (1979).

[54] Kahneman D: Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice, Nobel Lecture, December 8, (2002).

[55] Kahneman D.: Thinking Fast and Slow, Penguin Random House, UK, (2011).

[56] Keynes J.M.: A Treatise on Probability, MacMillan, London, (1921); Reprinted by Harper and Row, New York (1962).

[57] Kim E., Morse A. and Zingales L.: What Has Mattered to Economics Since 1970. Journal of Economic Perspectives, 20(4), 189-202, (2006).

[58] Knuth K.H.: Inductive Logic: From Data Analysis to Experimental Design. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Baltimore ML, USA, (ed. Fry RL) AIP Conference Proceedings 617, American Institute of Physics, New-York. 392-404, (2002).

[59] Knuth K.H.: What is a Question? In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Moscow ID, USA 2002 (ed. Williams C) AIP Conference Proceedings 659, American Institute of Physics, New-York. 227-242, (2003).

[60] Knuth K.H.: Intelligent Machines in the Twenty-First Century: Foundations of Inference and Inquiry. Phil. Trans. Roy. Soc. Lond. A, 361, 28592873 (2003).

[61] Knuth K.H.: Deriving Laws from Ordering Relations. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Jackson Hole WY, USA 2003 (eds. Erickson GJ and Zhai Y) AIP Conference Proceedings 707, American Institute of Physics, New-York. 204-235, (2004).

[62] Knuth K.H.: Lattice Duality: the Origin of Probability and Entropy. Neurocomputing 67;245-274, (2004).

[63] Knuth K.H.: The Origin of Probability and Entropy. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Sao Paulo, Brazil 2008 (eds. Lauretto MS, Pereira CAB) AIP Conference Proceedings, American Institute of Physics, New-York. 35-48, (2008).

[64] Knuth K.H.: Measuring on Lattices. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Oxford, MS, USA, P. Goggans and C.Y. Chan, eds., AIP Conf. Proc. 1193 AIP, New York, 2009, pp. 132144 (arXiv:0909.3684v1 [math.GM]), (2009).

[65] Knuth K.H. and Skilling J.: Foundations of Inference, arXiv: 1008.4831v1 [math.PR], (2010).

[66] Knuth K.H. and Center J.L.: Autonomous Science Platforms and Question-Asking Machines. Cognitive Information Processing 2010 (CIP 2010), Italy, (2010).

[67] Knuth K.H., Gotera A., Curry C.T., Huyser K.A., Wheeler K.R. and Rossow B.W.: Revealing Relationships Among Relevant Climate Variables with Information Theory, arXiv:1311.4632v1 [physics.data-an], (2013).

[68] Knuth K.H.: Information-Based Physics: An Observer-Centric Foundation, arXiv:1310.1667v1 [quant-ph], (2013).

[69] Knuth K.H. and Walsh J.L.: *Personal Communication*, (2016).

[70] Kok M., Vrijling J.K., van Gelder P.H.A.J.M. and Vogelsang M.P.: Risk of Flooding and Insurance in the Netherlands, Flood Defence 2002, Wu et al. eds., Science Press, New York Ltd, (2002).

[71] Kolgomorov A.N.: Foundations of the Theory of Probability, 2nd English ed., Chelsea, New York, (1956).

[72] Kullback S. and Leibler R.A.: On information and Sufficiency. Ann. Math. Statist. 22:79-86, (1951).

[73] Laplace P.S.: Essai Philosophique sur les Probabilites, Courcier Imprimeur, Paris, (1819).

[74] Lindgren B.W.: Statistical Theory, Chapman & Hall, Inc., New York, (1993).

[75] Loomes G.: Prospect Theory. In S.N. Durlauf & L.E. Blume (eds.), The new Palgrave dictionary of economics (2nd Ed). Macmillan Publishers, Ltd, (2008).

[76] MacKay D.J.C.: Information Theory, Inference, and Learning Algorithms, Cambridge University Press, Cambridge, (2003).

[77] Masin S.C., Zudini V., and Antonelli M.: Early Alternative Derivations of Fechner's Law, Journal of Behavioral Sciences, 45, 56-65, (2009).

[78] McGlothlin W.H.: Stability of Choices Among Uncertain Alternatives, American Journal of Psychology, 69, 604-615, (1956).

[79] Medvedev Z.A. (1990): The Legacy of Chernobyl (paperback ed.). W. W. Norton and Company.

[80] Mongin P.: Expected Utility Theory, Prepared for the Handbook of Economic Methodology (slightly longer version than the published one), Davis J., Hands W., and Maki U., eds., London, Edward Elgar, (1997).

[81] Murray D.J.: A Perspective for Viewing the History of Psychophysices, Behavioral and Brain Sciences, 16, 115-137, (1993).

[82] Phillips L.D. and Von Winterfeldt D.: Reflections on the Contributions of Ward Edwards to Decision Analysis and Behavioral Research, Working Paper LSEOR 06.86, Operational Research Group, Department of Management London School of Economics and Political Science, (2006).

[83] Plateau J.A.F.: Sur la Mesure des Sensations Physiques et Sur la Loi Qui Lie l' Intensite de ces Sensations a l' Intensite de la Cause Excitante. Bulletins de l' Academie Royale des Sciences, des Lettres, et des Beaux-Arts de Belgique. 33, 376-388, (1872).

[84] Polya G.: How to Solve It, Princeton University Press, (1945). Second paperbound edition by Doubleday Anchor Books, (1957).

[85] Polya G.: Mathematics and Plausible Reasoning, 2 vols: Induction and Analogy in Mathematics, Patterns of Plausible Inference, Princeton Press, (1954).

[86] Rosenkrantz R.D.: Inference, Method, and Decision: Towards a Bayesian Philosophy of Science, D. Reidel Publishing Co., Boston, (1977).

[87] Schreiber T.: Measuring Information Transfer. Phys Rev Lett Vol. 85, pp. 461-464, (2000).

[88] Shannon C.E.: A Mathematical Theory of Communication. Bell Syst. Tech. J. 27:379-623, (1948).

[89] Sivia D.S.: Data analysis A Bayesian Tutorial. Clarendon Press, Oxford, (1996).

[90] Skilling J.: The Axioms of Maximum Entropy, in Erickson G.J. and Smith C.R., eds., Maximum entropy and Bayesian Methods in science and engineering, Vol. 1, 173-87, Kluwer, Dordrecht, (1988).

[91] Skilling J.: Nested Sampling, In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, (eds. Erickson G., Rychert J.T., and Smith C.R.) AIP Conference Proceedings, American Institute of Physics, New-York. 395-405, (2004).

[92] Skilling J.: Bayesics, In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, (eds. Knuth K.H., Abbas A.E., Morris R.D., and Castle J.P.) AIP Conference Proceedings, American Institute of Physics, New-York.

[93] Skilling J.: Nested Sampling for General Bayesian Computation, Bayesian Analysis 4, pp. 833-860, (2006).

[94] Skilling J.: The Canvas of Rationality, In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Sao Paulo, Brazil 2008 (eds. Lauretto MS, Pereira CAB) AIP Conference Proceedings, American Institute of Physics, New-York. 67-79, (2008).

[95] Slovic P.: Trust, Emotion, Sex, Politics, and Science: Surveying the Risk-Assessment Battlefield, Risk Analysis 19(4), (1999).

[96] Slovic P. and Weber E.U. : Perception of Risk Posed by Extreme Events, paper prepared for discussion at the conference 'Risk Management strategies in an Uncertain World', Palisades, New York, April 12-13, (2002).

[97] Slovic P., Finucane M.L., Peters E. and MacGregor D.G.: Risk as Analysis and Risk as Feelings: Some Thoughts About Affect, Reason, Risk and Rationality, Risk Analysis 24(2), (2004).

[98] Stevens S.S.: To Honor Fechner and Repeal His Law, Science, New Series, Vol. 133, No. 3446, 80-86, (1961).

[99] Tribus M.: Thermostatics and Thermodynamics. New York: van Nostrand, (1961).

[100] Tribus M.: Rational Descriptions, Decisions, and Designs, Pergamon Press, New York, (1969).

[101] Tversky A and Kahneman D.: The Framing of Decisions and the Psychology of Choice. Science, 211, 453-458, (1981).

[102] Tversky A. and Kahneman D.: Rational Choice and the Framing of Decisions. Journal of Business, 59, S251-S278, (1986).

[103] Tversky A. and Kahneman D.: Advances in Prospect Theory: Cumulative Representation of Uncertainty, Journal of Risk and Uncertainty, 5: 297-323, (1992).

[104] Van Erp H.R.N.: Uncovering the Specific Product Rule for the Lattice of Questions. (arXiv:1308.6303 [stat.ME]), (2013).

[105] Van Erp H.R.N., Linger R.O. and Van Gelder P.H.A.J.M.: An Outline of the Bayesian Decision Theory. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conf. Proc. 1757, (2016).

[106] Van Erp H.R.N., Linger R.O. and Van Gelder P.H.A.J.M.: Bayesian Decision Theory: A Simple Toy Problem. In Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conf. Proc. 1757, (2016).

[107] Van Erp H.R.N., Linger R.O. and Van Gelder P.H.A.J.M.: Deriving Proper Uniform Priors for Regression Coefficients, Part I, II, and III, Entropy Journal, (2017).

[108] Van Erp H.R.N., Linger R.O. and van Gelder P.H.A.J.M.: Introducing Inner Nested Sampling. (arXiv:1704.02207v1 [stat.ME]), (2017).

[109] Vanslette K.M.: *Personal Communication*, (2016).

[110] Von Neumann J. and Morgenstern O.: Theory of Games and Economic Behavior, Princeton, Princeton Univer. Press, (1944).

[111] Weber E.H.: Der Tastsinn und das Gemeingefuhl. In R. Wagner (Ed.), Handworterbuch der Physiologie mit Rucksicht auf Physiologische Pathologie (v.3 part 2, pp. 481-588). Braunschweig: Vieweg, (1846).

[112] Zellner A.: An Introduction to Bayesian Inference in Econometrics, J. Wiley & Sons, Inc., New York, (1971).

Main Thesis Findings

Primary Finding

The work of behavioural economists is unified by a substantial project of revision of economic theory, by replacing the *homo oeconomicus* with a psychological model that better fits the empirical data of (hypothetical) betting experiments. The Bayesian decision theory in its turn endeavours to replace the psychological model of the behavioural economists with a slightly adjusted *homo oeconomicus*. The Bayesian decision theory is a neo-Bernoullian decision theory. For it adopts Bernoulli's original utility function, on the strength of a consistency derivation. But it differs from Bernoulli's original expected utility theory in that it proposes that the most likely trajectory (i.e., the expected value) is not the only information upon which we typically base our decisions. That is, it is postulated in the Bayesian decision theory that worst- and best-case scenarios (i.e., the lower and upper bounds of our outcome probability distributions) can be expected to intrude themselves upon our decision making as well.

Secondary Finding

In the inquiry calculus relevances are assigned to sources of information. By way of the information theoretic product and sum rules, relevances may then be operated upon in order to determine the relevancy of a source of information in regards to some issue of interest. For example, if there is the possibility of some danger, then the inquiry calculus allows us to assign relevances to statements made by officials in regards to that danger. It is then found that the relevance of that official source is directly related to its unbiasedness and competence. A high probability of unbiasedness and competence imply a corresponding high relevance, and a low probability of unbiasedness and competence imply a low relevance. This mathematical derived result is in close correspondence with social scientific findings on the importance of trust. So, if the probability measure of probability theory assigns numerical values to the plausibilities of our propositions, then the relevance measure of information theory assigns, for some issue of interest, numerical values to the potential pertinence of the variables we are considering to include in our inference. And the fact that the extended information theory admits a product rule and a sum rule puts this theory on

the same footing as the Bayesian probability theory, which is why we use the qualifier 'Bayesian' in connection with this information theory.

Acknowledgements

Short Curriculum Vitae

Drs. van Erp is a researcher at Delft University of Technology. He has been involved in several European projects as a researcher. He has done research on subjects ranging from the sophisticated, i.e., C- and B-spline interpolation techniques and Nested Sampling, to the basic, i.e., uninformative priors for regression coefficients and Beta-like distributions, to the fundamental, i.e., specific product rule of the inquiry calculus and the neo-Bernoullian decision theory. He is currently working on the development of Bayesian Deep Learning techniques.