Is every expert equal?

An analysis of the differences in performance in structured expert judgement

Jan Harkema

Delft University of Technology

TU Delft

Is every expert equal?

An analysis of the differences in performance in structured expert judgement

by

Jan Harkema

to obtain the degree of Bachelor of Science

at the Delft University of Technology.

Student number: 5116899

Project duration: September 1, 2021 – January 17, 2022

Thesis committee:

Dr. Ir. G. F. Nane,	TU Delft, supervisor
Dr. J. G. Spandaw,	TU Delft
Prof. Dr. R. M. Cooke,	Independent, professor emeritus TU Delft, senior fellow emeritus RFF, supervisor

TU Delft

Abstract

In this thesis the differences in performance scores of experts in the Classical Model for structured expert judgement are analyzed. The underlying assumption in the Classical Model is that variance in performances of experts in a panel is at least partly resultant of the expert's ability to quantify uncertainty. This assumption is tested against the so called Random Expert Hypothesis, that states that these differences are solely resultant of random fluctuations. At the five percent significance level it is concluded that the variation in the combined score of experts cannot exclusively be explained by random fluctuations. When the assumption is tested individually for three different subject fields, health, policy and science, the Random Expert Hypothesis cannot be rejected for both health and policy related studies. Lastly it is shown that the variation in performances between the best and worst expert in a panel strongly correlates with the performance of the best expert against random panels. This indicates that the aggregation of experts according to the scoring rule in the Classical Model may primarily work to diminish the influence of low performing experts.

Contents

1

Introduction

The more we learn about our world, the more we realize how many things are uncertain in our increasingly complex world. Everyday, people have to make decisions where uncertainty plays a role. Be that because the mechanics behind a problem are not fully understood, or because obtaining the definite answers takes too much time or resources. In such cases, people can resort to people deemed experts in a given field or subject. These experts can help making assessments of uncertain quantities, which can then be used in the decision making process.

Whilst at first glance this may not seem like an exact science, one can give a mathematical foundation to the processing of the assessments of the experts. This is called structured expert judgement.

Structured expert judgement is a relatively new field in mathematics. It concerns itself with the aggregation of the assessments of different experts. One wants to do this in a way such that a meaningful answer is obtained which can facilitate the decision making process. One of the main methods of structured expert judgement is known as the Classical Model [3]. This model employs a scoring rule to score the performances of different experts and weight their assessments accordingly. An underlying assumption in the Classical Model is that the performance scores experts achieve on the calibration questions, or seed variables, are a predictor of the performances of the experts on the variable(s) of interest. That is to say that the variances between performances are at least partly non-random. The hypothesis that the variances between performances is resultant of random fluctuations is known as the Random Expert Hypothesis

In this thesis the Random Expert Hypothesis will be tested against the underlying assumption of the Classical Model that the performance scores of the experts are predictors of the performance on the variable(s) of interest. This will be done in order to validate if aggregation of experts in the Classical Model is tenable. In Chapter 2 the theory behind structured expert judgement and the Classical Model will be further elaborated on. After which in Chapter 3 the Random Expert Hypothesis will be explained and tested. This will be done by comparing the maximum score of an expert in a study with the maximum

score in a random panel. Lastly in Chapter 4 the Random Expert Hypothesis will be further elaborated on in more specific cases. So will it be checked if similar results follow for specific fields of study and certain characteristics of the study. The results will be discussed in Chapter 5.

2

Structured expert judgement

In the ever increasingly complex world, the need for the quantification of uncertainty grows. To make complex decisions, concerning for example the risk of an volcano eruption, the probabilities of failure in a nuclear power plant or the impact of human errors in air traffic control, one can turn to people deemed experts and ask for their opinion. The assessments of these experts can then be used to aid in the decision making process.

These experts can be asked to give their assessment of a given problem, a variable, in the form of different quantiles. For example, the experts can be inquired to give a 5, 50 and 95 percent quantile. The aggregation of the these different assessments amounts to structured expert judgement. The process by which the experts provide their assessments, their assessments are evaluated and aggregated and evaluated using objective measures is known as structured expert judgement.

2.1. Elicitation

To apply structured expert judgement, one first needs assessments of experts. The assessments are done with the use of confidence intervals, in the form of quantiles. Experts are asked to give a certain set of quantiles for each variable, or question. Those quantiles correspond with the subjective probability that the true value is equal to or less than the value specified by the expert. In most cases, three different quantiles are used. Those quantiles are the 5, 50 and 95 percent quantiles. In other words, experts are asked to give a 90 percent confidence interval, along with a best guess. They believe that there is a five percent probability of the realization being equal to or less than the 5 percent quantile, an equal probability of the realization being either smaller or larger than the 50 percent quantile and a 5 percent probability of the realization being larger than the 95 percent quantile. One can however use any set of quantiles that one sees fit for the problem.

2.2. Classical Model

Elicitation of experts can be done with only one expert. However, it is advised that multiple experts are gathered to assess the problem at hand. How one then combines the varying assessment of different experts can differ. In this thesis the Classical Model will be employed [1]. This model was developed by Roger M. Cooke at the Delft University of Technology and uses a scoring rule to aggregate the assessments of various experts. The scoring rule aims to evaluate the assessments of a given experts [3]. To do this, in the elicitation process, the experts are not only asked to assess the variable(s) of interest. Additionally they are asked to assess some seed variables, also called calibration variables or questions. Those variables of course ought to be from the same field as the variable(s) of interest. The realizations of those seed variables are to be unknown to the experts. However, they are needed for the evaluation of the expert's assessment. Most often variables are chosen from unpublished studies, or variables which are unknown at the time of elicitation, but become known during the time frame of the elicitation process. These realizations can then be used to score the various experts.

2.2.1. Scoring rules

The scoring rule in the Classical Model consists of two components. The first component, the calibration score, is a measure for the statistical accuracy of an expert. Whilst the second component, the information score, is a measure of information an expert gives, or how well he is able to concentrate the probability mass in a small interval.

2.2.2. Calibration score

The calibration score is calculated over the whole set of seed variables in a given study. For this, one needs a so called probability vector $\mathbf{p}$. The entries of this probability vector are determined by the probability bins determined by the set of quantiles. Take for example the case where experts are asked to assess the 5, 50 and 95 percent quantiles of the variables. This gives four probability bins, the first and the last of size 0.05, and the middle two of size 0.45. Thus the probability vector becomes $\mathbf{p} = (0.05, 0.45, 0.45, 0.05)$. The realization of a given seed variable falls within one of those four bins, depending on the assessment of the expert. Over the set of seed variables, one can count the amount of realizations falling in any given bin of a certain expert. Say, if the realization lies between the 50 and 95 percent quantile, the realization falls into the third bin. Counting the amount of realizations per bin over the seed variables and dividing the number with the total number of seed variables yields the empirical probability vector $\mathbf{s}$. Where each entry corresponds to the proportion of the realizations falling in a specific bin. Using $\mathbf{s}$ and $\mathbf{p}$, one can calculate the Kullback-Leibler divergence $I(\mathbf{s}, \mathbf{p})$ of $\mathbf{s}$ and $\mathbf{p}$, also called the relative entropy. The Kullback-Leibler divergence is defined as follows:

$$I(\mathbf{s}, \mathbf{p}) = \sum_{i=1}^{n} s_i \cdot \ln \frac{s_i}{p_i},$$

where n is the number of probability bins.

It can be shown that

$$2 \cdot m \cdot I(\mathbf{s}, \mathbf{p}),$$

where m is the number of seed variables, asymptotically follows a Chi-squared distribution with $n-1$ degrees of freedom [4]. Following from this result we define the calibration score of expert e in the following way:

$$Cal(e) = 1 - F(2 \cdot m \cdot I(\mathbf{s}, \mathbf{p})),$$

where F is the cumulative distribution function of a χ^2_{n-1} distribution. Note that the calibration score can be any value between zero and one, where the higher the score, the better the expert is deemed to be on the aspect of statistical accuracy.

2.2.3. Information score

Contrary to the calibration score, the information score is calculated separately for each seed variable, after which the average is calculated.

To calculate the information score, first the intrinsic range needs to be defined. The intrinsic range is a range in which the value of a given variable may realistically lie. Sometimes there are clear bounds for such a range. If a variable for example concerns a percentage, one knows that the value has to be at least 0 and at most 100, and this can than be taken as intrinsic range. However, many times their is not such a clear range in which the variable may lie. In that case the intrinsic range is often derived from the assessments of the experts and the realization. For any given variable, one takes the minimum value between all assessments of the lowest quantile by the expert and the realization. This is the lower bound denoted by L. Similarly, for the upper bound, one takes the maximum value between all experts' assessments and the realization for the given seed variable. This is denoted by U. Then an overshoot k is added on both sides of the range defined by L and U. Typically this overshoot is chosen to be 10 percent of the length of the interval $[L, U]$. Which gives the following intrinsic range:

$$[L^*, U^*] = [L - k(U - L), U + k(U - L)].$$

Now let $\mathbf{q}^{ej}$ be the vector with entries the quantile assessments of an expert e for seed variable j. Now the information score for expert e for seed variable j, denoted by $I_j(e)$, is calculated as follows:

$$I_j(e) = p_1 \cdot \ln \frac{p_1}{q_1^{ej} - L^*} + p_n \cdot \ln \frac{p_n}{U^* - q_n^{ej}} + \ln (U^* - L^*) + \sum_{i=2}^{n-1} p_i \cdot \ln \frac{p_i}{q_i^{ej} - q_{i-1}^{ej}},$$

Where $\mathbf{p}$ is again the probability vector. So the information score for expert e becomes:

$$Inf(e) = \sum_{i=1}^{m} \frac{I_i(e)}{m}$$

where again m is the number of seed variables. Also, the information score is a non-negative value, where again the higher the score, the better the expert is believed to be at the ability of concentrating probability mass in a small interval.

2.2.4. Combined score

Both statistical accuracy and informativeness are important in the assessment of experts. Without statistical accuracy, the judgement of an expert is of little value. On the other hand, an expert can increase the probability of realizations falling into one of his central bins by increasing the size of said bins. But it should be obvious that needlessly increasing the bins only to improve ones statistical accuracy is to be discouraged. Therefore the Classical Model uses both the calibration score and the information score to evaluate the experts' assessments. To combine the scores, the calibration score is simply multiplied with the information score to obtain the combined score:

$$CS(e) = Cal(e) \cdot Inf(e).$$

It is important to note, that generally speaking, the statistical accuracy of an expert is deemed more important than the informativeness. This coincides with the fact that the calibration score is a 'quick' function, its value decreases rapidly when more realizations fall outside of the quantiles specified by the expert. So experts are mainly scored on their statistical accuracy, but in cases where those are similar, the information score gives a higher score to the expert that managed to capture a similar amount of realizations on a smaller interval.

In the Classical Model the combined score is used to weight the assessments from the experts on the variable(s) of interest. The weights of the experts are proportional to their combined scores. The weight of expert e_k is defined as follows:

$$w(e_k) = \frac{CS(e_k)}{\sum_{i=1}^{l} CS(e_i)},$$

where l is the number of experts in the study.

2.2.5. Decision maker

The assessments of the expert can be aggregated in a decision maker. The quantiles given by an expert for a certain variable in combination with the intrinsic range of that variable give rise to a probability density function f and cumulative distribution function F. In the Classical Model, a minimal information distribution is used. Thus the probability mass between to quantile assessments of an expert are uniformly distributed. The probability mass in the outer two quantiles are uniformly distributed over the length of the intrinsic range till below or above the respective quantile assessment. The probability distribution of the decision maker for a certain variable is then constructed by summing the distribution functions of the experts multiplied with their respective weights, as defined in Subsection 2.2.4.

3

Random Expert Hypothesis

The Classical Model assigns different weights to different experts, depending on the performance of the experts on the seed variables. It takes into account the statistical accuracy of the experts, and combines that with the information score, the ability of an expert to concentrate the uncertainty in a small interval with high probability. The underlying assumption in the Classical Model is thus that the variance between experts' performance is at least partly resultant of actual differences in the experts' abilities, and not solely due to random fluctuations.

The Random Expert Hypothesis challenges this underlying assumption. It states that the differences in performance on the seed variables by experts is solely the consequence of random fluctuations. In other words, the actual experts can be seen as arbitrary picks from the set of all hypothetical experts that can be formed with the answers to the calibration questions of the actual experts. This would mean that correct method of aggregating experts' assessments is one of equal weighting, where every expert's assessments gets the same weight.

The Random Expert Hypothesis has already been challenged by Marti, Mazzuchi and Cooke [6]. They concluded that the Random Expert Hypothesis, and thus the equal weighting approach, is extremely unlikely. They had drawn there conclusions from a data set of 44 structured expert judgement studies and evaluating the statistical accuracy, or the calibration score, for different experts. Whilst it is important that experts are statistically accurate, it is easy to increase ones statistical accuracy by increasing the length of the different quantiles in the assessments. Therefore, in this chapter the Random Expert Hypothesis is tested with the aid of the combined score, to see if the variance in the performance of experts can be explained solely due to chance.

3.1. Hypothesis

To test if differences between experts could be explained by something other than random variation, a statistical analysis of expert elicitation data was performed. If the variation between experts is solely the result of random fluctuation, then the best expert in the actual panel should not perform any better or worse than the best expert in a scrambled panel. This led to the following null hypothesis:

> H_0: The actual expert panel is arbitrarily picked from the set containing all possible scrambled panels.

This hypothesis was tested against the alternative hypothesis:

> H_1: The actual expert panel is not arbitrarily picked from the set containing all possible scrambled panels.

Here both best expert and scrambled panel should be defined. The best expert in a panel is the expert which achieves the highest combined score as discussed in Subsection 2.2.4. A scrambled panel is a panel where the experts' assessments on different seed variables are randomly reassigned. For example, suppose the original panel has three experts A, B and C. All of which have assessed three seed variables on three quantiles, as shown in Table 3.1.

Table 3.1: Example panel

Expert	Seed 1	Seed 2	Seed 3
A	(3, 6, 9)	(13, 45, 70)	(0.2, 0.5, 0.9)
B	(4, 7.5, 9)	(16, 44, 64)	(0.1, 0.45, 0.8)
C	(2, 5, 8)	(10, 50, 80)	(0.3 , 0.6, 1.0)

When the panel is scrambled, the assessments for each seed variable are reordered, to get three random, or hypothetical experts A', B' and C'. For example, the scrambled panel could be as in Table 3.2.

Table 3.2: Example scrambled panel

Expert	Seed 1	Seed 2	Seed 3
A'	(4, 7.5, 9)	(16, 44, 64)	(0.3, 0.6, 1.0)
B'	(2, 5, 8)	(13, 45, 70)	(0.2, 0.5, 0.9)
C'	(3, 6, 9)	(10, 50, 80)	(0.1 , 0.45, 0.8)

3.2. Data

For this project, the data from 50 different studies have been used in which expert elicitation was used. The data set includes six studies which were not yet available at the time of the research by Marti et al.

[6]. The data was retrieved from the site of Roger M. Cooke [2]. Each study has a number of different experts, each of which has made assessments for a range of different seed variables, along with the variable(s) of interest. Furthermore, the realizations of the seed variables are also included. For this study only the seed variables and their realizations are of interest. An overview of the different studies with their respective number of experts, seed variables and quantiles can be found in Appendix B.

3.3. Methodology

Saying that the actual panel is arbitrarily picked from the set of all possible scrambled panels, and thus that the best expert in the actual panel is not expected to perform any better or worse than the best expert in a scrambled panel, does not mean that one expects to see the exact same score for both. The same fluctuations seen between experts were also expected between the best experts in either the actual panel and a scrambled panel, or two scrambled panels.

So instead for 1,000 scrambled panel the best expert have been calculated. Grouping all those scores gives a range in which the actual best expert was expected to lie. Under the null hypothesis, the percentile in which the actual expert lies in the approximated range of performances scores is an arbitrary pick from a uniform distribution on the interval [0,1]. Repeating this process for multiple studies, read different panels, allows one to make an assertion on the probability that the observed results are realized under the assumption that the null hypothesis is true. This has been tested with two different statistical tests, where a significance level of five percent has been used.

3.3.1. Binomial test

First of all, the binomial test has been applied to calculate the probability of the observed results under the null hypothesis. This is the same test as used by Marti et al. [6]. This test only considered if the actual best expert performs better or worse in comparison with the scrambled panels' best expert. If the actual best expert's score is in the 50th percentile or below, this was considered to be a fail. If the actual best expert's score was above the 50th percentile, this was considered to be a success. Then the probability of the observed number of successes over 50 studies were calculated with the binomial distribution

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Under the null hypothesis, both a fail and a success were equally like, and thus both have a probability of 0.5. Therefore the probability of k successes could also be written as

$$Pr(X = k) = \binom{n}{k} 0.5^n.$$

Thus the probability that one sees at least k successes is

$$Pr(x \geq k) = \sum_{i=k}^{n} \binom{n}{i} 0.5^n$$

Note that the null hypothesis in combination with the alternative hypothesis is two-tailed. So if the null hypothesis were to be rejected with a significance level of five percent, it were to be rejected when the

probability of at most k observed successes is less then 0.025, or the probability of at least k successes is greater than 0.975.

3.3.2. Sum test

The binomial test does not consider the deviance of the actual experts from the median. An expert who ends up at the 98th percentile is considered the same as an expert who ends up at the 52th percentile. Hence the hypothesis was simultaneously tested by the sum test. The sum test uses the fact that the sum of independent random variables has an asymptotic normal distribution by the Central Limit Theorem. The sum of n uniformly distributed variables on the interval [0,1] is thus normally distributed with mean $n \cdot 0.5$ and standard deviation $\sqrt{n/12}$. One can then take the value obtained from summing the percentile scores of the actual best experts over all the studies and calculate the probability of normally distributed variable obtaining a value at least as extreme as this sum of percentile scores, where the normally distributed variable has mean and variance as described. Hence a table of the normal distribution could be used to calculate when the sum of the percentile scores is in the lower or upper 2.5th percentile.

3.4. Code

This study has been carried out in `Python` with help of the `Anduryl` package [8]. `Anduryl` is a library which has function implementations to import structured expert judgement data and carry out different calculations on this data, like calculating the combined score as described in Subsection 2.2.4. All code can be found in Appendix C. The structure of the code will be briefly elaborated.

3.4.1. Code structure

The data has first been imported from the respective folders. The data for each study consisted of two files: a `.dtt` file, which contains the experts assessments for all the variables, and a `.rls` file, which contained the realizations for the different seed variables. Then the data has been evaluated per study. The assessments from the expert have been extracted and scrambled. After which the scores for each scrambled expert was calculated. The score of the best expert has been recorded. This process was repeated 1,000 times for each study. Per study, this gave a list of 1,000 hypothetical best expert scores. This list has been sorted. This allowed one to find the index of the actual expert in the range of scores of the scrambled best experts. Dividing this index by 1,000 gave the percentile in which the score of the actual expert lies. So this yielded a list with the percentiles of the actual best expert for each study. Thereafter the binomial test and the sum test have been easily carried out.

3.5. Results

For every study, the percentiles in which the scores of the actual best experts lay when compared with the scores of the scrambled best experts have been documented. These percentiles can be found in

Table 3.3.

Table 3.3: Percentiles of the best expert for each study

Study	Percentile	Study	Percentile
Arkansas	0.469	Illinois	0.750
Arsenic D-R	0.719	Italy	0.630
ATCEP Error	0.844	IQEarnings	0.049
BFIQ	0.655	Liander	0.162
biol_agents	0.707	Nebraska	0.840
Brexit food	0.735	obesity_ms	0.921
CDC_all	0.865	p6r	0.316
CDC_ROI	0.885	PHAC 2009 T4	0.197
CoveringKids	0.857	PoliticalViolence_March17_CW	0.995
CREATE	0.063	puig-gdp	0.911
CWD	0.603	puig-oil	0.988
Daniela	0.807	Raveem	0.886
dcpn_fistula	0.190	SanDiego	0.764
eBBP	0.907	Sheep Scab	0.943
EffusiveErupt	0.605	Spain	0.502
Erie Carps	0.441	SPEED	0.954
FCEP Error	0.759	Tadini_Clermont_anon	0.731
Florida	0.438	Tadini_Quito_anon	0.914
France	1.00	TdC	1.00
Gerstenberger	0.308	tobacco	0.691
GL-NIS	0.931	Topaz	0.927
Goodheart	0.827	UK	0.394
Hemophilia	0.797	umd_nremoval	0.975
IceSheet2012	0.847	USGSfinal	0.767
ICE_US+EU_June 22 2018	0.576	Washington	0.022

3.5.1. Test results

The number of percentiles found in Table 3.3 which lie above 0.50 is 38. So there are 38 out of 50 successes. Calculating the probability under the null hypothesis, gave a probability of $1.53 \cdot 10^{-4}$, or 0.0153 percent. This was a probability well below the threshold of 2.5 percent needed to reject the null hypothesis.

For the sum test, the sum of the percentile scores needed to be calculated. It turned out to be 34.1. This needed to be compared to a normal distribution with mean $50 \cdot 0.5 = 25$ and standard deviation $\sqrt{50/12}$. It turned out that the probability of a random variable with such distribution attaining a value of at least 34.1, is $4.14 \cdot 10^{-6}$, or 0.000414 percent. This was an even more extraordinary small probability well below the 2.5 percent threshold.

Given that both tests rejected the null hypothesis with a significance level of five percent, the null hypothesis could be rejected. Therefore, it can be stated that the variance between experts can not solely be explained by random fluctuations, confirming the conclusions drawn by Marti et al. [6].

4

Random Expert Hypothesis expansion

In Chapter 3 it has been shown that the variance between expert performances could not solely be explained by random fluctuations. However, not every study is the same. Studies can differ in the amount of experts, the amount of seed variables, the number of quantiles needed to be specified and of course the field of study.

In this chapter, the Random Expert Hypothesis will be investigated further by limiting the set of studies to a specific field. Furthermore, the correlations between different characteristics of the studies, like the amount of experts, and the percentile score of the actual best expert are studied.

4.1. Random Expert Hypothesis for health, policy and science related studies

Whilst the Random Expert Hypothesis was rejected for general cases in Chapter 3, this does not necessarily mean anything for individual studies. Whilst it does not make sense to test this for individual studies, it could be further specified to provide more insight in the underlying mechanisms of the variance in performances between experts. Therefore the set of studies has been split up according to the field of specialization to test if the same conclusions can be reached for more specific studies.

The studies can be grouped in three different fields of expertise, namely studies on health, policy or science related subjects. In Table 4.1 the grouping can be found. Then the Random Expert Hypothesis can be tested for each field of expertise. Remember the null hypothesis from Chapter 3:

> H_0: The actual expert panel is arbitrarily picked from the set containing all possible scrambled panels.

This hypothesis will be tested against the alternative hypothesis:

> H_1: The actual expert panel is not arbitrarily picked from the set containing all possible scrambled panels.

So the hypothesis did not change from the earlier hypothesis, and thus again the scores of the best experts are tested. Therefore, the percentile scores for the actual experts, as found in Table 3.3, could be used to test the hypothesis.

Table 4.1: Grouping of the different studies

Health	Policy	Science
BFIQ	Arkansas	Arsenic D-R
biol_agents	Brexit food	ATCEP Error
CDC_all	CDC_ROI	Daniela
CWD	CoveringKids	EffusiveErupt
dcpn_fistula	CREATE	Erie Carps
eBBP	Florida	FCEP Error
France	Illinois	Gerstenberger
Hemophilia	IQEarnings	GL-NIS
Italy	Nebraska	Goodheart
p6r	obesity_ms	ICE_US+EU_June 22 2018
PHAC	PoliticalViolence_March17_CW	IceSheet2012
SanDiego	Raveem	Liander
Sheep Scab	tobacco	puig-gdp
Spain	Washington	puig-oil
UK		SPEED
		Tadini_Clermont_anon
		Tadini_Quito_anon
		TdC
		Topaz
		umd_nremoval
		USGSfinal

4.1.1. Test results health studies

As can be seen in Table 4.1, there are fifteen different studies which are health related. Of those studies, as can be checked with Table 3.3, eleven had their actual expert score in the top half of the percentiles. The probability of eleven successes out of fifteen under the null hypothesis was 0.0592 when calculated using the binomial test, or 5.92 percent. Therefore the null hypothesis could not be rejected according to the sum test.

The sum of the percentiles of all health related studies is 9.47. For the sum test this number had to

be compared to a normal distribution with mean is $15 \cdot 0.5 = 7.5$ and a standard deviation of $\sqrt{15/12}$. This gave a probability of 0.0390, or 3.90 percent. Thus both test failed to reject the null hypothesis, hence one cannot conclude that the variation between the performances of experts in health related studies is not down due to random fluctuations.

4.1.2. Test results policy studies

In the data set there are fourteen different studies which regard policy, those can be found in Table 4.1. In nine of those studies, the actual best expert scored in the upper fifty percentiles. According to the binomial test, the probability of nine experts scoring in the upper fifty percentiles under the null hypothesis was 0.212, or 21.2 percent. Clearly this probability is too high to reject the null hypothesis.

If the percentile scores of the best expert in policy related studies are summed, this yields 8.60. To apply the sum test this value was to be compared with a normal distribution with a mean of $14 \cdot 0.5 = 7$, and a standard deviation of $\sqrt{14/12}$. This gave a probability of 0.0693, or 6.93 percent. Again both tests failed to reject the null hypothesis. Thus as with health related studies, the variation between experts need not to be explained by anything other than random fluctuations.

4.1.3. Test results science studies

The number of studies in the field of science in the data set is slightly larger, namely twenty-one. Of those studies, the actual best expert's percentile score was higher than 0.5 eighteen times. By the binomial test, the probability of eighteen successes out of twenty-one trials with probability 0.5 was a mere $7.45 \cdot 10^{-4}$, or 0.0745 percent. Which was well below the threshold to reject the null hypothesis.

The percentiles of the best experts summed together produces 16.0. Comparing this with a normal distribution with mean $21 \cdot 0.5 = 10.5$ and standard deviation $\sqrt{21/12}$ gives a probability of a value being at least as large as 16.0 under the null hypothesis of $1.61 \cdot 10^{-5}$, or 0.00161 percent. Thus for studies in the field of science, the null hypothesis could be rejected. Hence the variation between the performances of experts in science related studies cannot solely be explained by random fluctuations.

An overview of all the different p-values for the three different fields of expertise can be found in Table 4.2.

Table 4.2: P-values for both binomial and sum test for each field of specialization

Test	Health	Policy	Science
Binomial	0.0592	0.212	7.45e-4
Sum	0.0390	.0693	1.61e-5

4.2. Variance of percentile scores and the number of quantiles

Whilst the format of the different structured expert judgement studies is the same, there are some factors on which they can differ. One of those factors is the number of quantiles in which the experts have to give their assessment. In the data set there are two different number of quantiles being used, three and five. The experts who have to assess three quantiles, give their assessment for the 5, 50 and 95 percent probabilities. When five quantiles are assessed, those correspond to the 5, 25, 50, 75 and 95 percent probabilities. One could expect that, if there are significant differences between experts' performances which are not down to random fluctuations, that those differences will be more clear if the experts have to give more assessments.

Since there are only two different numbers of quantiles, one has to be mindful about how to check for a correlation. From a visual inspection there is little to go off on, see Figure 4.1. It could be checked with the sum test if the percentile scores obtained when experts have to assess five quantiles are higher than when they have to assess three quantiles. But a more intuitive method is the calculation of the so called ϕ coefficient, also known as the mean square contingency coefficient. It is a measure for the correlation between two binary variables. The data of the studies can be interpreted as two binary variables. For the first variable, the two options are quite clear, one class has three quantiles, whilst the other class has five quantiles. For the second variable, one has to split the percentile scores. This has been done at the 0.5 mark, so the studies in which the actual best expert scored less or equal to the 50th percentile were in the first class, whilst the studies where the best experts had a percentile score above 50 were in the second class. This gave rise to four different groups.

Table 4.3: Binary groups of number of quantiles and percentile scores

	≤ 50	> 50	Total
3	6	27	33
5	6	11	17
Total	12	38	50

	$Y = 0$	$Y = 1$	Total
$X = 0$	a	b	e
$X = 1$	c	d	f
Total	g	h	n

The number of studies in each group is shown in Table 4.3, along with a table for arbitrary groups. The formula for the ϕ coefficient is as follows:

$$\phi = \frac{ad - bc}{\sqrt{efgh}},$$

where each letter corresponds to the number of studies in each group as shown in Table 4.3. Evaluating this formula led to ϕ = -0.190. This could be interpreted as a weak correlation [10]. More notable, the weak correlation was a negative correlation, instead of the expected positive correlation. Therefore a greater number of quantiles assessed does not seem to exacerbate differences in performance of the experts.

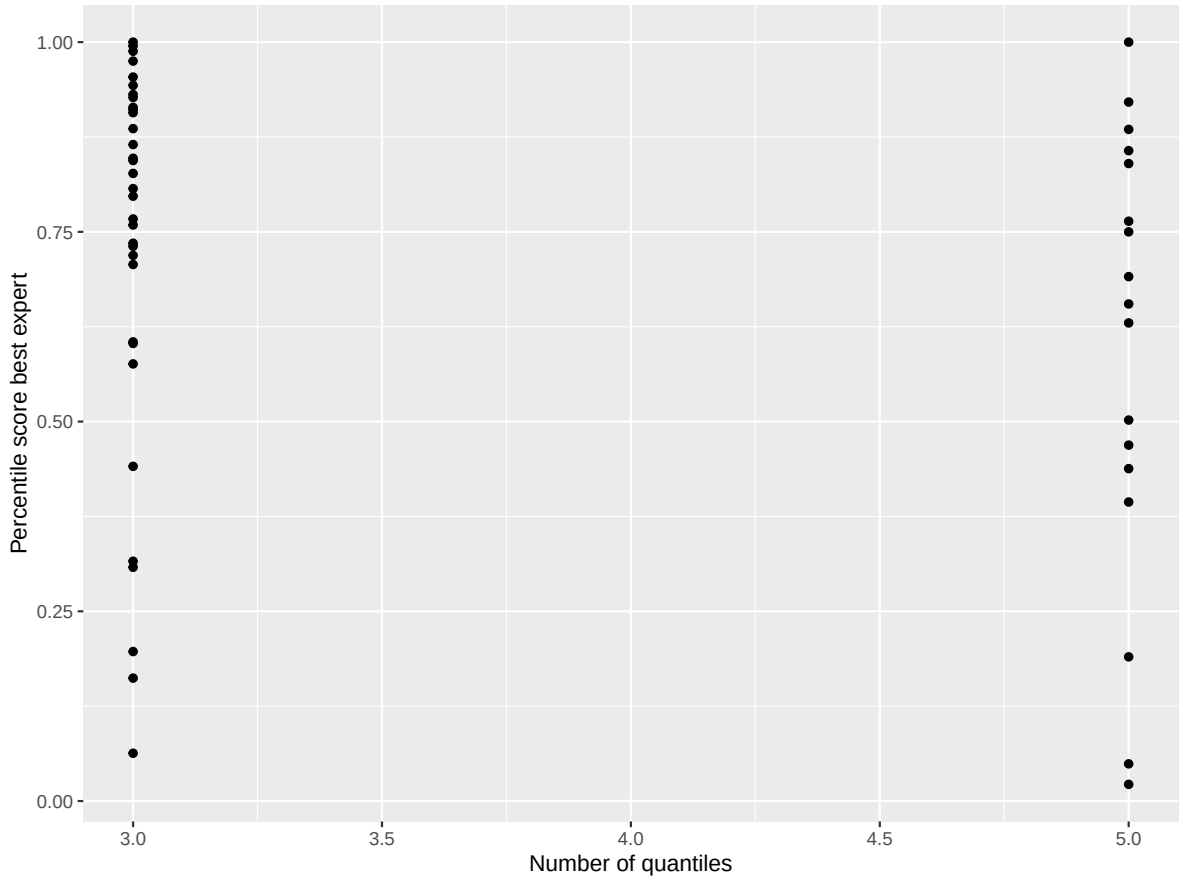


Figure 4.1: Scatter plot of the number of quantiles against the percentile score of the best expert

4.3. Variance of percentile scores and the number of seed variables

Another factor which differs from study to study, is the number of seed variables the experts have to assess and that they will be scored on. In the studies in the data set, the number of seed variables ranges from 7 up to 21. If there are significant differences between the performances of experts not down to random fluctuations, those differences were expected to be more clear when the experts have to assess more seed variables. Assessing more seed variables would diminish the influence of random fluctuations on the variance between combined scores. This should lead to the actual best expert attaining a higher percentile score in comparison with the scrambled panels. To check for such correlation, multiple methods can be used. In this thesis primarily the Kendall rank correlation coefficient, also called Kendall's τ_b [5], was used, together with Spearman's rank correlation coefficient, or Spearman's ρ [9]. An explanation of both coefficient and their correlation can be found in Appendix A.

When checking for one-sided correlations with Kendall's rank correlation coefficient, one implicitly tests the null hypothesis

H_0: The Kendall $\tau_b \leq 0$

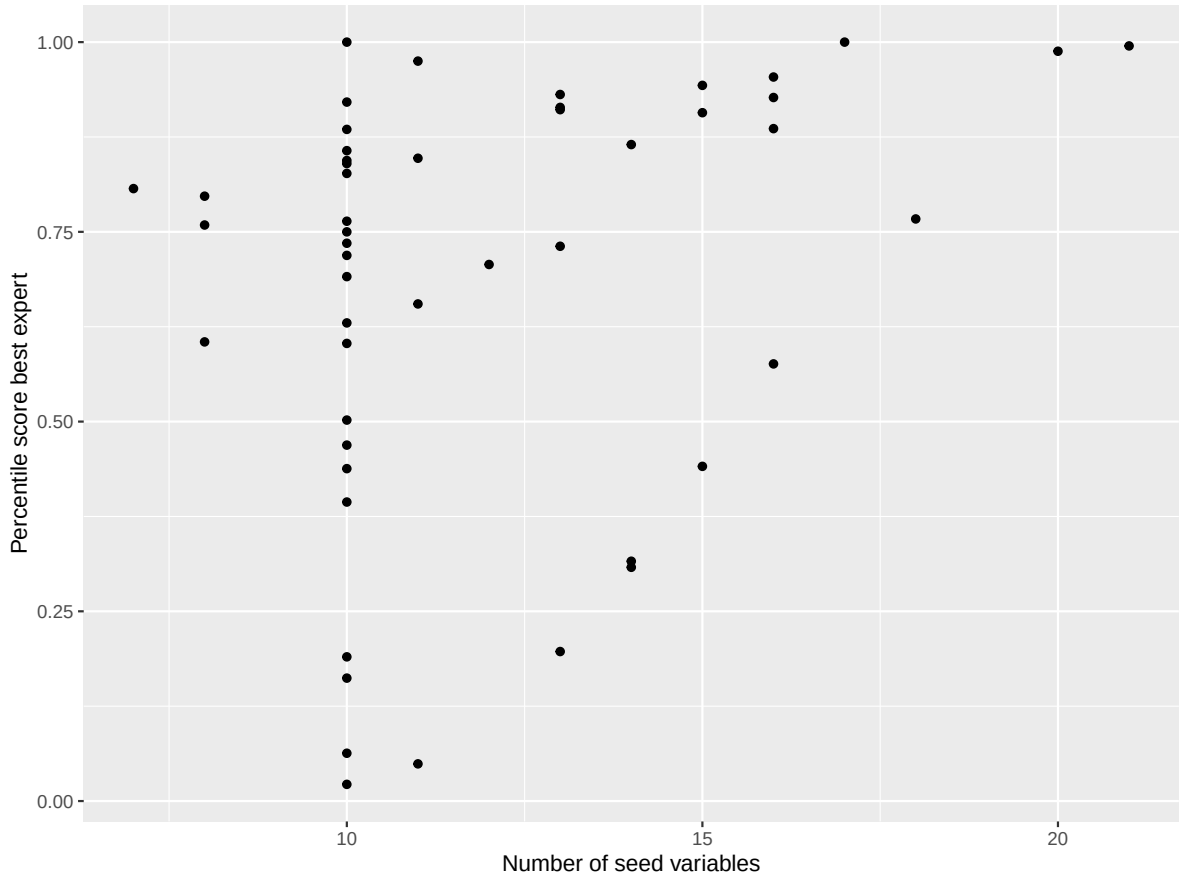


Figure 4.2: Scatter plot of the number of seed variables against the percentile score of the best expert

against

$$H_1\text{: The Kendall } \tau_b > 0.$$

A similar hypothesis is tested when employing Spearman's ρ. In other words, the null hypothesis is that there is no positive correlation between the number of seed variables and the percentile scores of the actual best expert, whilst the alternative hypothesis states that there is indeed such a positive correlation.

In Figure 4.2 the plot of the data points is displayed.

Visually it was not immediately clear if there is a significant positive correlation between the amount of seed variables and the percentile score of the best expert. However, the Kendall τ_b was calculated to be 0.26, with a p-value of 0.000605. Similarly, Spearman's ρ was estimated to be 0.35, with a p-value of 0.00604. Thus there is indeed a positive correlation and the null hypothesis could be rejected at the five percent significance level.

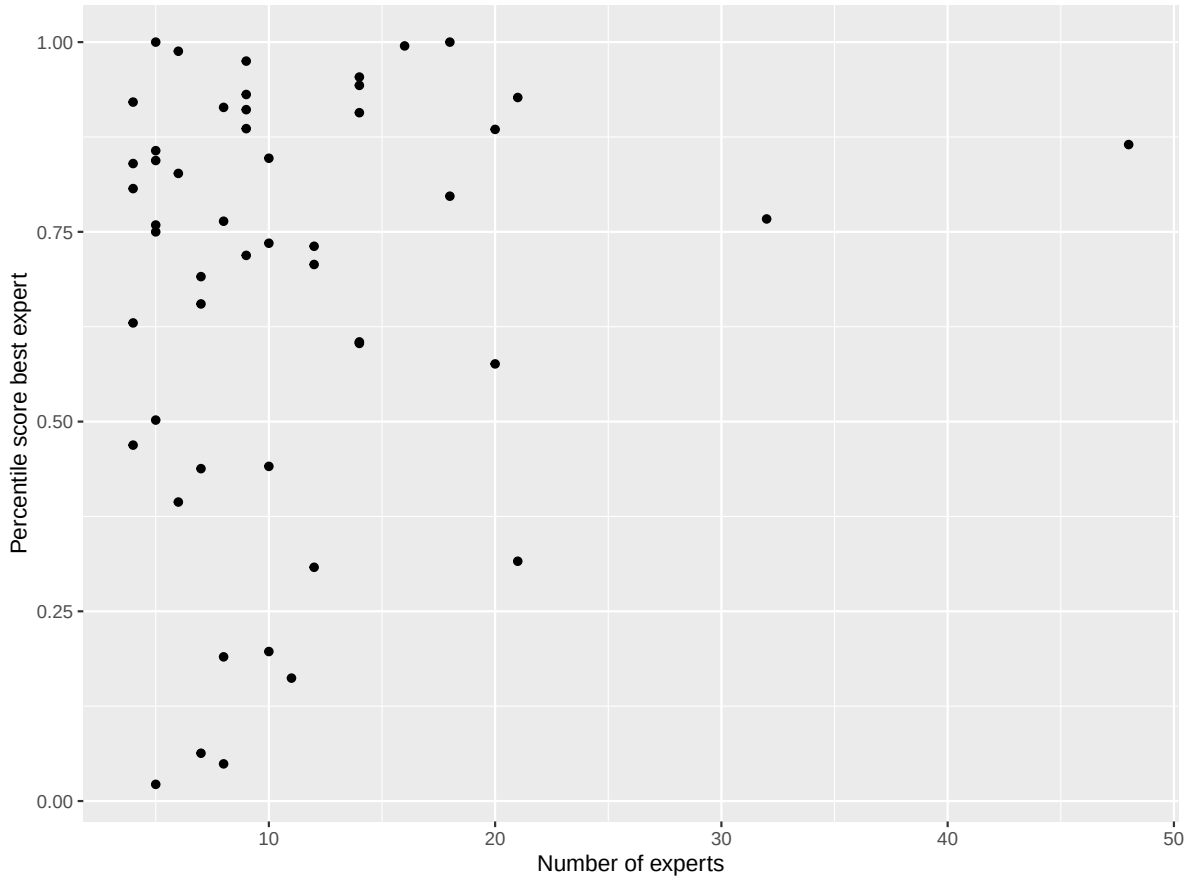


Figure 4.3: Scatter plot of the number of experts against the percentile score of the best expert

4.4. Variance of percentile scores and the number of experts

The different studies also differ in the amount of experts that they selected to for expert elicitation. The number of experts selected in the studies ranges from 4 experts to 48. Similar as with the number of seed variables, the differences in expert performance which are not down to randomness were expected to get clearer when a greater number of experts are considered. This hypothesis has again be checked with Kendall's and Spearman's rank correlation coefficient. In Figure 4.3 the different data points of the fifty studies are shown.

Once again a visual inspection showed little sign of a significant monotonic correlation between the number of experts considered and the performance of the best expert against scrambled panels. The Kendall τ_b was calculated to be only 0.083, so barely a positive correlation, Furthermore, the p-value was 0.205, thus the result is not significant. With Spearman's correlation coefficient the story was similar, it had a value of 0.13, a bit higher than the Kendall τ_b, but with a p-value of 0.190 also this results is not statistically significant. Thus, contrary to the number of seed variables, the number of experts seems not to have any significant correlation with the performances of the best experts against the scrambled panels.

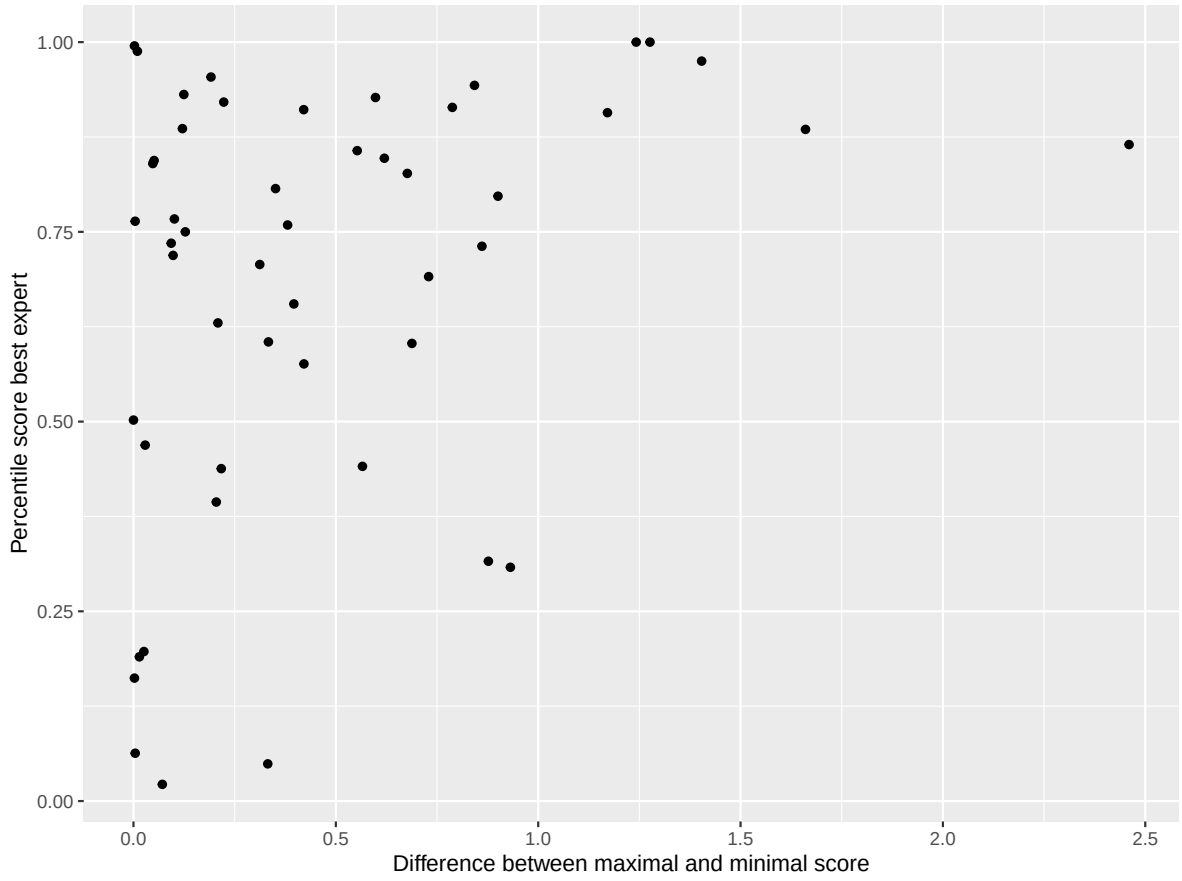


Figure 4.4: Scatter plot of the difference between the maximal and minimal score of the actual experts against the percentile score of the best expert

4.5. Variance of percentile scores and the worst performing expert

When the panels are scrambled, the assessments of experts are mixed together. So if one expert has scored poorly, his performance should impact the score of all experts in the scrambled panels for the worse, where the evaluation of experts is still done with the combined score. Therefore it was theorized that the actual best expert achieves higher percentile scores in panels where the worst performing expert has a lower score, especially when the best expert score is on the higher end. In other words, the larger the difference between the scores of the worst and best performing expert (the minimal and maximal scores) the higher the percentile score of the best performing experts was expected to be against the scrambled panels. In Figure 4.4 one can see the difference between the maximal and minimal score against the percentile score of the best expert.

Calculating the Kendall rank correlation coefficient gave a τ_b of 0.22 with a p-value of 0.0136. Spearman's rank correlation coefficient gave an even higher 0.315, with a similar p-value of 0.0130. Therefore the null hypothesis has been rejected and it can be stated that there is in fact a significant positive correlation between the difference in performance of the best and worst scoring experts and the percentile scores of the best expert.

4.5.1. Minimal scores and variance

This strong correlation raised the question if it is possible to find an even stronger correlation with similar statistics. The variance between the worst and best scoring expert could be simplified by only taking the minimal score, instead of the difference between the best and worst performing experts. Or one could account not only for the variance between the best and worst scoring experts, but for the general variance between the scores of different experts. In the first case one would expect a negative correlation between the minimal score in a panel and the percentile score of the best expert. In the latter case one would expect that a greater variance between experts correlates positively with the percentile score of the best expert. In Table 4.4 one can see the different rank coefficients for both variables along with their p-values.

Table 4.4: Rank correlation coefficients for both the minimal score and the variance against the percentile score of the best expert

Variable	Kendall's τ_b	p-value	Spearman's ρ	p-value
Minimal score	-0.36	0.000133	-0.50	9.64e-05
Variance	0.21	0.0142	0.30	0.0166

From Table 4.4 it is clear that both the minimal score as well as the variance between scores correlate significantly with the percentile score of the best expert. Both the estimates for the coefficients, and the p-values for the correlation between the variance between scores and the percentile score were extremely similar to the values obtained for the correlation between the difference of the maximal and minimal scores and the percentile score. Furthermore, the correlation between the lowest score among experts and the percentile score of the best expert was even stronger. Thus the minimal score seems to be the best predictor out of the three for the performance of the best expert against the scrambled panels.

5

Discussion

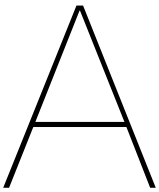
The aim of this thesis was to test the underlying assumption of the Classical Model against the Random Expert Hypothesis. As shown in Chapter 3 the null hypothesis, *the actual expert panel is arbitrarily picked from the set containing all possible scrambled panels*, can be rejected at the five percent significance level by both the binomial test and the sum test. Hence the variance between the combined scores of experts in a panel is not solely the consequence of random fluctuations. Thus the underlying assumption of the Classical Model, that the variance between experts is at least partly down to differences in the experts ability, holds and therefore the aggregation of the experts' assessments based on the scoring rule is justified. This was in line with earlier conclusions of performance weighting versus equal weighting drawn by Marti, Mazzuchi and Cooke [6].

This is however a general conclusion for the studies as a whole. As seen in Section 4.1 and Table 4.2, this conclusion cannot be reached for all different fields in which the studies have been performed. Solely for studies in the field of science the null hypothesis could be rejected at the five percent significance level. It should be stressed however, that this is no indication of the quality of the performances of experts in different fields. Merely, it has been shown that the differences between the combined scores of experts in science related studies were significant enough to conclude that it cannot be down exclusively due to randomness. It could be that experts in health and policies related studies are simply more equal in ability. Another explanation may be that the sample sizes for health and policy related studies were just too small to get significant results, since there were 15 health and 14 policy related studies, against 21 science related studies. Which may not seem like a huge different at first glance, but there are fifty percent more data points for science than for policy. Therefore it would be interesting to see if the Random Expert Hypothesis can be rejected when a greater sample is considered.

Some further explorations were made in Chapter 4 to gain better insights in the variances in performances between experts and when they become clear. This was done by testing for correlations

between different characteristics of the studies against the percentile scores of the best expert in the studies. There was deemed to be no significant correlation between the number of quantiles the assessments were specified in against the percentile score of the best expert. A similar conclusion was reached for the correlation between the number of experts in a study and the percentile score of the best expert. However, there was a significant positive correlation between the number of seed variables on which the experts were evaluated and the performance of the best expert against the scrambled panels. Furthermore, there were also significant correlations between various metrics for the variance in scores between experts and the percentile score of the best expert. Most notably was the negative correlation between the score of the worst performing expert against the percentile score. Which seems to indicate that the scoring rule may not necessarily be as important to increase the influence of the best expert as it is to negate the impact of the worst expert on the aggregation of assessments.

As stated before, the Random Expert Hypothesis and the underlying assumption of the Classical Model only concerns the variances in performances according to the scoring rule of the Classical Model of the different experts in a panel. It does not offer any insight in the quality of the assessments, as their assessments are only compared with the assessments of other experts in the panel and not evaluated on how accurate their assessments are in general. To gain a better understanding of the differences in performances of experts in various studies, the Random Expert Hypothesis should be tested again for different fields of specialization, but with a larger data set. Another interesting question for future research would be to extend on the correlation between the worst performing expert and the percentile score of the best performing experts, to see if the same conclusion can be reached once one corrects for, or removes, low performing experts from the panel.



Technical appendix

A.1. Kendall rank correlation coefficient

The Kendall rank correlation coefficient, also called Kendall's τ, is a statistical measure of rank correlation between two variables. This is to say, it measure the dependence between two variables, only not by the actual value of those variables, but by the order, or rank, of those values. The Kendall correlation coefficient takes values between -1 and 1. This happens when there exists a strictly monotone function mapping one variable onto the other. A stronger correlation will lead to Kendall's τ being closer to -1 if the correlation is negative, and 1 if the correlation is positive. Whilst the absence of any correlation will lead to a value of zero.

There are multiple variants of Kendall's τ. Here Kendall's τ_b will be discussed, which makes adjustments for ties in rank within the data. For the calculation of the Kendall τ_b, concordant and discordant pairs need to be defined.

A.1.1. Concordant and discordant pairs

Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ is set of observations of random variables X and Y. Then any pair of two observation for which it holds that either $x_i > y_i$ and $x_j > y_j$ or $x_i < y_i$ and $x_j < y_j$ is said to be concordant. Conversely, if $x_i > y_i$ and $x_j < y_j$ or $x_i < y_i$ and $x_j > y_j$, the pair is said to be discordant. When one of the values is tied, the pair is neither concordant or discordant.

A.1.2. Kendall's τ_b

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of observations of X and Y, two random variables. Then the Kendall τ_b coefficient is defined as follows:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

where

$$n_0 = n(n-1)/2$$
$$n_1 = \sum_i t_i(t_i - 1)/2$$
$$n_2 = \sum_j u_j(u_j - 1)/2$$
$$n_c = \text{Number of concordant pairs}$$
$$n_d = \text{Number of discordant pairs}$$
$$t_i = \text{Number of tied values in the group of ties } i \text{ for the first variable}$$
$$u_j = \text{Number of tied values in the group of ties } j \text{ for the second variable.}$$

A.1.3. Correlation testing

To see if two variables are correlated, one can calculate the value for the Kendall τ_b coefficient and check its significance. The correlation can be tested either one-sided or two-sided. In the first case, one tests the null hypothesis that $\tau_b \leq 0$ or $\tau_b \geq 0$ against the alternative hypothesis that $\tau_b > 0$ or $\tau_b < 0$ respectively. In the second case, one test the null hypothesis that $\tau_b = 0$ against the alternative hypothesis that $\tau_b \neq 0$.

For the calculation of τ_b and the p-value, R will be used. When the variables checked for correlation have fifty or more observations, as is the case for the data used in this thesis, τ_b is scaled so it is asymptotically standard normal distributed. It is scaled as follows:

$$z = \frac{3\tau_b\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}.$$

In R using the function `cor.test()` this z is then compared with a standard normal distribution to calculate the p-value.

A.2. Spearman's rank correlation coefficient

Spearman's rank correlation coefficient, also called Spearman's ρ and denoted by r_s, is a statistical measure of rank correlation. This is to say, it measures the dependence between two variables, only not by the actual values of those variables, but by the order, or relative position, of those values. Spearman's correlation coefficient, takes values between -1 and 1. A value of -1 or 1 is also called a perfect Spearman correlation. This happens when there exists a strictly monotone function mapping one variable onto the other. A stronger correlation will lead to Spearman's correlation coefficient being

closer to either -1 or 1, depending on whether the relation is decreasing or increasing. Whilst the absence of any correlation will lead to a value of zero.

Spearman's rank correlation coefficient r_s is calculated with the same formula as the Pearson correlation coefficient. Only the variables X and Y are replaced with the ranks $R(X)$, $R(Y)$. Hence the formula is as follows:

$$r_s = \rho_{R(X),R(Y)} = \frac{\mathrm{Cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}.$$

If all n observations are distinct, i.e. no two observations have the same value, then this formula simplifies to

$$r_s = 1 - \frac{6\sum_i R(X_i) - R(Y_i)}{n(n^2-1)}.$$

A.2.1. Correlation testing

To check the correlation between two variables, one can calculate the value for Spearman's ρ and its p-value, which can be either done one-sided or two-sided. So in the first case, one test the null hypothesis that $r_s \leq 0$ or $r_s \geq 0$ against the alternative hypothesis that $r_s > 0$ or $r_s < 0$ respectively.

For the calculation of the p-value, r_s can be used to calculate the test statistic

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}}$$

where n is the number of observations. This t is asymptotically t_{n-2} distributed [7]. In R this test statistic is used in the function `cor.test()` with the option `exact = FALSE`.

B

Data

Name of study	Number of experts	Number of seed variables	Number of quantiles
Arkansas	4	10	5
Arsenic D-R	9	10	3
ATCEP Error	5	10	3
BFIQ	7	11	5
biol_agents	12	12	3
brexit food	10	10	3
CDC_all	48	14	3
CDC_ROI	20	10	5
CoveringKids	5	10	5
CREATE	7	10	3
CWD	14	10	3
Daniela	4	7	3
dcpn_fistula	8	10	5
eBBP	14	15	3
EffusiveErupt	14	8	3
Erie Carps	10	15	3
FCEP Error	5	8	3
Florida	7	10	5
France	5	10	5
Gerstenberger	12	14	3
GL-NIS	9	13	3
Goodheart	6	10	3
Hemophilia	18	8	3

Name of study	Number of experts	Number of seed variables	Number of quantiles
ICE_US+EU_June 22 2018	20	16	3
IceSheet2012	10	11	3
Illinois	5	10	5
IQEarnings	8	11	5
Italy	4	10	5
Liander	11	10	3
Nebraska	4	10	5
obesity_ms	4	10	5
p6r	21	14	3
PHAC	10	13	3
PoliticalViolence_March17_CW	16	21	3
puig-gdp	9	13	3
puig-oil	6	20	3
Raveem	9	16	3
SanDiego	8	10	5
Sheep Scab	14	15	3
Spain	5	10	5
SPEED	14	16	3
Tadini_Clermont_anon	12	13	3
Tadini_Quito_anon	8	13	3
TdC	18	17	3
tobacco	7	10	5
Topaz	21	16	3
UK	6	10	5
umd_nremoval	9	11	3
USGSfinal	32	18	3
Washington	5	10	5

C

Python code

C.1. Run code

```
import REHsource as reh
from scipy.stats import binomtest

# For colored printing
cyel = '\33[93m'
cblu = '\33[94m'
cend = '\33[0m'


# Print start run
print(cblu + '---------------- New run ---------------' + cend)


folderpaths = ['folder']
results = reh.REH(folderpaths, 1000, .1, 0, 1)

# Count number of percentile scores above 0.5
a = 0
for i in range(len(results)):
    if results[i][0] > 0.5:
        a += 1

bin = binomtest(a, len(results), 0.5, 'greater')

# Save percentile scores per stuy in a .txt file, one line per study, scores seperated from
    study name with ;
text = open('percentilescores.txt', 'w')
for r in results:
    text.write(str(r[0]) + '; ' + r[1] +'\n')

```

```
text.close()
```

C.2. Source code

```

# Initializing imports
import sys
import os
sys.path.append('path_to_anduryl')
import anduryl
import numpy as np
from random import shuffle


# Takes a project and calculates the combined scores for the experts, then returns the highest score
def best_expert(project: anduryl.Project, overshoot: float, alpha: float, calpower: float):
    project.experts.calculate_weights(overshoot, alpha, calpower)
    return np.amax(project.experts.weights)


# Takes a path to a folder and extracts all .dtt files from that folder and puts their filenames with location in a list
def import_studieslist(folderpath: str):
    results = []
    for i in os.listdir(folderpath):
        # Check if file is in .dtt format (Excalibur)
        if i.endswith('.dtt'):
            # Append path + file name without extension
            results.append(folderpath + '/' + os.path.splitext(i)[0])
    return results


# Given a filename and location, it creates an Anduryl project from that file
def init_project(filepath: str):
    project = anduryl.Project()
    project.io.load_excalibur(f'{filepath}.dtt', f'{filepath}.rls')
    return project


# Reshapes the assessments array and realizations array to only include the seed questions
def purge_target(project: anduryl.Project):

    # Obtain indices for seed variables
    idx = project.items.get_idx('seed')

    # Redefines the assessments and realizations array to only contain the seed variables
    project.assessments.array = project.assessments.array[:, :, idx]
    project.items.realizations = project.items.realizations[idx]
    return project


```

```
# Takes an Anduryl project, scrambles the assessments per question
def randomization(project: anduryl.Project):

    # Initialize the function
    orig_assessment = project.assessments.get_array('both')
    nexperts = np.size(orig_assessment, 0)
    nquantiles = np.size(orig_assessment, 1)
    nquestions = np.size(orig_assessment, 2)
    orderlist = [i for i in range(nexperts)]

    # Create array for the randomized answers
    new_assessment = np.zeros((nexperts, nquantiles, nquestions))

    # Randomization: loop over the questions and randomize per question
    for i in range(nquestions):
        shuffle(orderlist)
        for j in range(nexperts):
            new_answer = orderlist[j]
            for k in range(nquantiles):
                new_assessment[j, k, i] = np.copy(orig_assessment[new_answer, k, i])

    # Rewrite assessments in project to scrambled assessments
    project.assessments.array = new_assessment
    return project

# Takes a list of folderpaths, extracts all excalibur files and scrambles each project N
    times, whilst recording the best expert score in the scrambled panels.
# Uses these scores to determine the percentile in which the orginal project's best expert
    lies compared to the scrambled projects' best experts.
def REH(folderpaths: list, N: int, overshoot: float, alpha: float, calpower: float):

    # Create a list with all the study paths
    studieslist = []
    for folder in folderpaths:
        studieslist += import_studieslist(folder)

    overall_results = []

    # Repeat for each study
    for studypath in studieslist:

        # Initialize Anduryl project, remove expert with unanswered 'seed' questions, and
            initialize results list
        print(studypath)
        tproject = init_project(studypath)
        project = remove_incomplete_expert(tproject)
        actual_best = best_expert(project, overshoot, alpha, calpower)
        results = np.zeros(N)

        # Carry out N randomizations, record the results
        for i in range(N):
            rproject = randomization(project)
            rres = best_expert(rproject, overshoot, alpha, calpower)
            results[i] = rres
```

```
98
99          # Sort the results, then find the index of the original expert to calculate its
                percentile
100         results.sort()
101         a = np.searchsorted(results, actual_best)
102         percentile = a/N
103
104         # Record percentile
105         overall_results.append([percentile, studypath])
106
107     return overall_results
108
109
110 # Removes expert which has not assessed all seed variables
111 def remove_incomplete_expert(project: anduryl.Project):
112
113     # Get experts ids and initialize set for experts to be removed
114     ids = project.experts.get_exp('both')
115     removeset = set()
116
117     # For each expert, check all seed questions, break if an unanswered question is found and
            add to removeset
118     for exp in project.experts.get_idx('actual'):
119         for q in project.items.get_idx('seed', where=True):
120             if np.isnan(project.assessments.array[exp, 0, q]):
121                 removeset.add(ids[exp])
122                 break
123
124     # Remove experts which have not assessed every seed variable
125     for exp in removeset:
126         project.experts.remove_expert(exp)
127
128     return project
```

D

R code

```r
library(ggplot2)
library(psych)


# Loading the data
SEJdata <- read.csv2("file.csv")

# Section 4.2.
ggplot(SEJdata, aes(x = Nquantiles, y = REHpercent)) +
  geom_point() +
  labs(x = 'Number of quantiles', y = 'Percentile score best expert' )

phidata <- matrix(c(6,6,27,11), nrow = 2)
phi(phidata)


# Section 4.3.

ggplot(SEJdata, aes(x = Nseed, y = REHpercent)) +
  geom_point() +
  labs(x = 'Number of seed variables', y = 'Percentile score best expert' )

cor.test(SEJdata$Nseed, SEJdata$REHpercent, method = 'kendall', alternative = 'greater')
cor.test(SEJdata$Nseed, SEJdata$REHpercent, method = 'spearman', alternative = 'greater',
    exact = F)

# Section 4.4.

ggplot(SEJdata, aes(x = Nexperts, y = REHpercent)) +
  geom_point() +
  labs(x = 'Number of experts', y = 'Percentile score best expert' )

cor.test(SEJdata$Nexperts, SEJdata$REHpercent, method = 'kendall', alternative = 'greater')
```

```r
cor.test(SEJdata$Nexperts, SEJdata$REHpercent, method = 'spearman', alternative = 'greater',
    exact = F)

# Section 4.5

ggplot(SEJdata, aes(x = Difminmax, y = REHpercent)) +
  geom_point() +
  labs(x = 'Difference between maximal and minimal score', y = 'Percentile score best expert'
       )


cor.test(SEJdata$Difminmax, SEJdata$REHpercent, method = 'kendall', alternative = 'greater')
cor.test(SEJdata$Difminmax, SEJdata$REHpercent, method = 'spearman', alternative = 'greater',
     exact = F)

cor.test(SEJdata$Min, SEJdata$REHpercent, method = 'kendall', alternative = 'less')
cor.test(SEJdata$Min, SEJdata$REHpercent, method = 'spearman', alternative = 'less', exact =
    F)

cor.test(SEJdata$Variance, SEJdata$REHpercent, method = 'kendall', alternative = 'greater')
cor.test(SEJdata$Variance, SEJdata$REHpercent, method = 'spearman', alternative = 'greater',
    exact = F)
```

References

[1] T. J. Bedford and R. M. Cooke. *Probabilistic risk analysis: foundations and methods*. Cambridge University Press, 2001.

[2] R. M. Cooke. *Data from structured expert judgement*. URL: https://rogermcooke.net.

[3] R. M. Cooke. *Experts in Uncertainty: opinion and subjective probability in science*. Oxford University Press, 1991.

[4] P. G. Hoel. *Introduction to mathematical statistics*. Wiley, 1971.

[5] M. G. Kendall. "A New Measure of Rank Correlation". In: *Biometrika* 30.1-2 (June 1938), pp. 81–93. ISSN: 0006-3444. DOI: 10.1093/biomet/30.1-2.81. eprint: https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf. URL: https://doi.org/10.1093/biomet/30.1-2.81.

[6] D. Marti, T. A. Mazzuchi, and R. M. Cooke. "Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis". In: *Expert Judgement in Risk and Decision Analysis*. Ed. by A. M. Hanea et al. 2021. Chap. 3, pp. 53–82.

[7] W. H. Press and W. T. Vetterling. *Numerical recipes in c: the art of scientific computing (2nd ed.)* Cambridge Universtiy Press, 1992.

[8] G. Rongen et al. *Anduryl*. Version 1.2. 2020. URL: https://github.com/grongen/anduryl.

[9] C. Spearman. "The proof and measurement of association between two things". In: *The American Journal of Psychology* 15 (1904), pp. 72–101.

[10] L. Todorova, P. Vassilev, and J. Surchev. "Using Phi Coefficient to Interpret Results Obtained by InterCriteria Analysis". In: *Novel Developments in Uncertainty Representation and Processing*. Ed. by K. T. Atanassov et al. Cham: Springer International Publishing, 2016, pp. 231–239. ISBN: 978-3-319-26211-6.