

Bayes Factors and the Observational Method

Gregory B. BAECHER^a and John T. CHRISTIAN^b

^aDepartment of Civil and Environmental Engineering, University of Maryland, USA

^bConsulting Engineer, Burlington, MA, USA

Abstract. The statistics course most engineers took in college—seldom more than one—introduced them to a particular species of statistics, which, regrettably, is of limited use in geotechnical practice: frequentist sampling theory. That species of statistical thinking arose to address problems of agricultural experimentation, biology, and economics. It is mostly applicable to narrow domains in medical trials, political polls, and the like, where carefully planned experiments lead to large databases and *p*-value tests of hypotheses. These are not the problems facing the geotechnical engineer. He or she faces extremely limited numbers of observations (maybe only one), measurements of differing types and quality, a blend of qualitative and quantitative information, and a need to make sequential decisions as data arrive.

Keywords. Bayesian, observational approach, risk screening, measurements

1. Strong Inferences from Sparse Data

Bayesian methods are widely considered suited to geotechnical problems, but much of current Bayesian developmental work in geotechnical reliability tends toward applications involving large numbers of data (*e.g.*, Bayesian Belief Nets, or MCMC) and to sampling problems for which frequentist methods are already well suited. The real power of Bayesian thinking lies in making sense of very few data, combining different sorts of data, and making real-time decisions.

A *Bayes Factor* is the Likelihood Ratio of a known set of observations conditioned on a set of hypotheses over which one's degrees-of-belief are imperfect; it is the ratio of the conditional probability of the data given an hypothesis to the conditional probability of the same data given the complement of the hypothesis. The Bayes Factor is used to update prior degrees-of-belief to posterior degrees-of-belief. This sounds simple and obvious to anyone familiar with Bayesian methods, but in practice it is an immensely powerful approach to thinking about practical problems.

2. Bayes Factors

How does the weight of new information influence inferences about probability? It does not do

so through a simple sum. Inferences of probability of some event based on new information need to reflect two things: (1) How probable the event was thought to be before this new information arrived, and (2) how knowing the new information logically changes that prior probability.

The relationship among these things is,

$$P(X | \text{data}) \propto P(X)P(\text{data}|X) \quad (1)$$

in which, $P(X | \text{data})$ is the updated probability of some event, condition, parameter, etc., x , after the new “data” are taken into account; $P(X)$ is the probability before the new data; and $P(X | \text{data})$ is the conditional probability of observing the new data were x true. This last conditional probability has a special name in statistical theory, it is called the *Likelihood*.

So, the updated probability of x is logically proportional to the prior probability of x multiplied by how likely the data would be if that value of x obtained. The more likely the data that were observed, the greater the weight of evidence they provide for x being true. The normalizing constant for this equation is the sum over all possible values that x might take on—that is, how probable the observed data are irrespective of the true value of x —making the sum of the $P(x|\text{data})$ over all x equal to 1.0,

$$P(X | data) = \frac{P(X)P(data|X)}{\sum P(x_i)P(data|x_i)} \quad (2)$$

This derives from the *Total Probability Theorem*, and is referred to as *Bayes Rule* or *Bayes Theorem*.

Dividing the updated probability in favor of x by the updated probability in favor of *not- x* yields,

$$\frac{P(X | data)}{P(\bar{X} | data)} = \frac{P(X) P(data|X)}{P(\bar{X}) P(data|\bar{X})} \quad (3)$$

in which $\bar{X} = \text{not-}x$. The normalizing constant is the same in the numerator and denominator, and cancels.

The term $\frac{P(X)}{P(\bar{X})}$, is the odds in favor of x before the data are taken into account. More precisely, this term is the *prior odds*, sometimes loosely called the base rate or ratio of base rates. The second term,

$$LR = \frac{P(data|X)}{P(data|\bar{X})} \quad (4)$$

is the *Likelihood Ratio (LR)*; it is the weight of evidence contained in the observations. It contains all the statistical information in the data relative to x . Both pieces of information are important. The Likelihood Ratio alone (*i.e.*, the data alone) is not sufficient to draw conclusions on the updated probabilities. The prior odds and the weight of information are multiplied, not added.

Table 1. Jeffreys (1998) strength of evidence table

LR	Strength of evidence
< 1:1	Negative
1:1 to 3:1	Barely worth mentioning
3:1 to 10:1	Substantial
10:1 to 30:1	Strong
30:1 to 100:1	Very strong
> 100:1	Decisive

The weight of information contained in the *LR* depends on how different it is from 1.0. If $LR=1.0$, then $P(data|X) = P(data|\bar{X})$, and the

association of the data we observed with x is the same as their association with \bar{x} , so the observation does not alter the probability of x . If $LR>1.0$, the data increase the probability of x , and if $LR<1.0$, the data decrease the probability of x . Table 1 is Jeffreys' (1998) qualitative interpretation of the *LR*.

In our experience, people tend not to make such quantitative calculations from small numbers of observations, and when they do, they tend to make significant errors. This is supported by the literature (Puhan et al. 2005).

2.1. Creating an Additive Score

To make the relationship of Equation (3) additive, one can take the logarithm of both sides.

$$\ln \frac{P(X | data)}{P(\bar{X} | data)} = \ln \frac{P(X)}{P(\bar{X})} + \ln \frac{P(data|X)}{P(data|\bar{X})} \quad (5)$$

In words, the logarithm of the updated odds can be found from the sum of the logarithm of the prior odds and the log-Likelihood Ratio.

2.2. Correlations Among the Indicators

For observations, (z_1, z_2, \dots, z_k) the likelihood becomes the *joint likelihood* of all the observations. In other words, this is the joint probability of observing the set of things that we did observe, were x true. It is a joint conditional probability, so if the observations are correlated, this must be accounted for,

$$LR = \frac{P(z_k, \dots, z_1 | x)}{P(z_k, \dots, z_1 | \bar{x})} = \frac{P(z_1 | x)}{P(z_1 | \bar{x})} \dots \frac{P(z_k | z_{k-1}, \dots, z_1, x)}{P(z_k | z_{k-1}, \dots, z_1, \bar{x})} \quad (6)$$

In the desirable case that the observations are mutually independent, the joint likelihood ratio reduces to the product,

$$LR = \prod_i \frac{P(z_i | x)}{P(z_i | \bar{x})} \quad (7)$$

3. Screening for Levee Safety

Three case studies are presented and discussed to illustrate the power and simplicity of Bayes-Factor thinking in observational problems. The first involves condition assessments of flood levees and the use of qualitative inspector data in drawing conclusions about the safety of levee reaches when few or no instrumental data are available. How can a limited number of visual inspection data be used to draw quantitative inferences about safety?

Table 2 shows data on the association of two indicators—embankment cracking and embankment settlement—with levee sections that later failed or did not fail, within some specified interval of time (Baecher and Christian 2013). These data are hypothetical and only for illustration. For simplicity, we have taken a set of 1000 levee sections, and assumed that 100 of them later failed. So, the base rate of levee failure is 10%.

The Likelihood of observing cracking for a levee section that later fails is estimated by taking the number of levee sections that later fail and counting the fraction of these that exhibited cracking during the preceding inspection. In the present case, the first line of Table 2 shows this is 70 of the 100 levee sections, or 70%. Thus, the Likelihood of observing cracking for a failing levee section is 0.70. Similarly, 200 of the 900 unfailed sections exhibited cracking, so the corresponding Likelihood is 0.22 and $LR=(0.70/0.22)=3.18$. The posterior odds are $(0.1/0.9)(3.18)=0.35$; the posterior probability of failure is 0.26.

The second row of the table is for the observation of settlement alone. The third row is for the observation of both cracking and settlement, presuming that the conditional probabilities of the two data types are mutually independent. In fact, the data suggest that the information contained in the two observations is not independent. That is, if we know that cracking exists, that changes the probability (and thus, the Likelihood) of settlement. In other words, cracking and settlement tend to occur together, thus, if we know that one exists, the fact that the other also exists does not carry as much weight of evidence as it would if the probabilities were independent. The fourth row shows that the updated probability of failure indeed decreases from 0.56 to 0.38

on account of the dependence between the two indicators.

Table 2. Simple calculation of the updated probability of levee failure given the observation of cracking or settlement during an inspection. Hypothetical data for 100 failures and 900 non-failures.

case	failing		non-failing		LR	post. prob.
	n	p	n	p		
cracking	70	0.70	200	0.22	3.15	0.35
settlement	40	0.40	100	0.11	3.60	0.40
both (ind.)		0.28		0.02	11.34	1.26
both (dep.)	30	0.30	50	0.06	5.00	0.56

The data of Table 2 intend to suggest the association of various indicators observed during a periodic inspection with the later occurrence of levee failures. In principle, these likelihoods arise from the historical frequency with which indicators have been observed at failed and at well-performing levees.

In application, inspectors observe many indicators of performance. For simplicity, assume that the condition of each of the indicators is judged by the inspector to be either satisfactory or unsatisfactory. It is straightforward to allow for an ordinal ranking of scores (*e.g.*, low-medium-high) for any of the indicators, but this makes the analysis more complicated. Second, assume that the information contained in observing the indicators is mutually independent from one indicator to another. We saw above that this will not always be the case, but again the assumption can be relaxed in practice at the expense of a somewhat more complicated table.

As an example let there be eight (8) indicators that the inspector judges. For each, he or she rates the indicator as either acceptable (A) or unacceptable (U). The (additive) numerical score associated with each rating is set equal to the log-Likelihood Ratio for that indicator. Table 3 lists the values derived from historical judgment about these indicators for a system of levees. The $\ln(LR)$ is the natural logarithm of the Likelihood Ratio for failed levee sections, and $\ln(\bar{LR})$ is the corresponding term for the non-failed sections.

Table 3. Hypothetical historical values for indicators.

Indicator	LR	\bar{LR}	$\ln(LR)$	$\ln(\bar{LR})$
Unwanted Vegetation	0.80	1.20	-0.22	0.18
Encroachments	0.83	1.07	-0.18	0.07
Settlement	0.28	1.48	-1.29	0.39
Cracking	0.31	2.60	-1.16	0.96
Animal Control	1.00	1.00	0.00	0.00
Culverts / Pipes	0.83	1.25	-0.18	0.22
Relief Wells / Drainage	0.40	1.60	-0.92	0.47
Seepage	0.50	1.75	-0.69	0.56

Table 4 shows the results of a new survey for a particular section of levee. For “acceptable” indicators $\ln(LR)$ is used as the score; for “unacceptable” indicators $\ln(\bar{LR})$ is used. (The appropriate scores are in bold type in Table 4.) Once all the indicators are judged, the eight scores are added to give the total score of 0.66. The updated probability of levee failure associated with this score is simply calculated from Equation (5) to yield a value of $P(F|data) \approx 0.18$. This can all be automated in a spreadsheet.

Table 4. Results from a specific survey of the eight indicators.

Indicator	Value	$\ln(LR)$	$\ln(\bar{LR})$	Score
Unwanted Vegetation	A	-0.22	0.18	-0.22
Encroachments	U	-0.18	0.07	0.07
Settlement	U	-1.29	0.39	0.39
Cracking	U	-1.16	0.96	0.96
Animal Control	A	0.00	0.00	0.00
Culverts / Pipes	A	-0.18	0.22	-0.18
Relief Wells / Drainage	A	-0.92	0.47	-0.92
Seepage	U	-0.69	0.56	0.56
Sum of scores				0.66
ln-prior odds of failure				-2.20

The advantages of this scoring scheme are that (1) it is firmly founded in the logic of probability theory (*i.e.*, it is not *ad hoc*) and (2) it can be fully informed by historical statistical data (*i.e.*, the numbers need not be based on expert opinion). However, the same scheme works when some of the LR 's are empirical and some subjective: the two types of information are easily fused through the respective LR 's.

4. Competing Predictions

Two groups of analysts have been evaluating slope stability issues along a section of highway using different modeling approaches. The goal in each case is to predict the age at which individual sections of slope might become unstable due to weathering. Their respective predictions differ substantially for particular sections. How can observations of the current age of the sections inform the evaluation of these models?

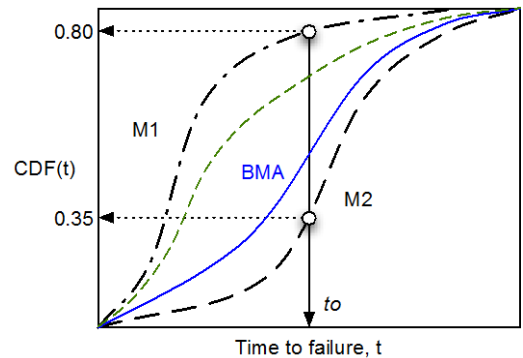


Figure 1. The curves are the predictions of failure before time, t , made by the two models, M_1 and M_2 . Probabilistic predictions of age to failure for the same highway slope section based on two competing models. CDF=cumulative probability. Bold solid curve shows case of slope failure; bold dashed curve shows case of no slope failure.

For a given section of highway, the slopes are currently stable, and have been in service for 20 years. The predictions of the two modeling approaches are shown in Figure 1. The curves are the predictions of failure before time, t , made by the two models, M_1 and M_2 .

M_1 predicts 0.80 probability of failure before 20 years, while M_2 predicts 0.35 probability. We know that the age at failure is at least 20 years since the slope remains standing. Thus, the posterior odds of model M_1 being correct compared to M_2 can be inferred to be (Hoeting et al. 1999),

$$\frac{P(M_1|data)}{P(M_2|data)} = \frac{P(M_1)}{P(M_2)} \times \frac{P(data|M_1)}{P(data|M_2)} \quad (8)$$

If *a priori* the two models are considered equally valid, the prior odds are 1.0, so the updated odds on M_1 is,

$$\frac{P(M_1|data)}{P(M_2|data)} = (1.0) \times \frac{(1-0.8)}{(1-0.35)} = 0.3 \quad (9)$$

which corresponds to a probability $P(M_1|data) = 0.23$. A similar calculation can be made if more than one slope section has been observed.

A Bayesian Model Average (BMA) is formed by weighting the respective predictions by their probabilities, given the data (Barnard 1963),

$$F(t|data) = \frac{P(M_1|data)F(t|M_1) + P(M_2|data)F(t|M_2)}{P(M_1|data) + P(M_2|data)} \quad (10)$$

as shown by the bold solid curve in the figure.

In contrast, presume that the slope in question failed at time t_0 . This should give more weight of evidence to the model that predicts an earlier failure. Equation (1) still obtains, but Equation (2) becomes

$$\frac{(P(M_1 | data))}{(P(M_2 | data))} = (1.0) \times ((0.8)/((0.35))) = 2.23 \quad (11)$$

which corresponds to a probability of 0.70. This generates the dotted BMA closer to the M_1 prediction.

BMA is a simple hypothesis weighting. It does not account for the fact that models may be based on different principles and therefore their predictions may not be combinable, but such issues are beyond the present scope.

5. Liquefaction Probabilities and Likelihoods

Empirical methods for evaluating the potential of a saturated soil to liquefy during an earthquake usually compare a cyclic stress ratio (*CSR*) representing the stress caused by the earthquake and a cyclic resistance ratio (*CRR*) representing the soil's resistance to liquefaction. Almost all the criteria for relating *CRR* and *CSR* are derived from observations at sites where liquefaction either did or did not occur during earthquakes.

Figure 2 from Idriss and Boulanger (2008) is typical for SPT results. The horizontal axis consists of a parameter representing the measured soil strength. The vertical axis describes the es-

timated load imposed by the earthquake. Liquefaction cases are plotted as filled-in circles, and cases without liquefaction are open circles.

The first researchers used visual approximations to separate the regions, but more recent work has employed frequentist methods such as discriminant analysis and logistic regression analysis. The probabilities in those analyses are the probabilities of observing data in one of the regions, given that the site has or has not liquefied.

What the engineer needs is the reverse of this. The engineer wants to know, given a set of data that indicate a safe site, what is the probability that the site will liquefy. To address this problem, many recent studies have applied Bayesian methods. This requires a prior estimate of the probability of liquefaction. Since researchers usually do not have an estimate of the prior probability for the abstract exercise of creating general plots, a common procedure is to assume a non-informative prior (Jeffreys 1998). This is, in fact, what was done in most of the recently published analyses, and it means that the analysts assumed equal prior probability of liquefaction and non-liquefaction, of 0.50.

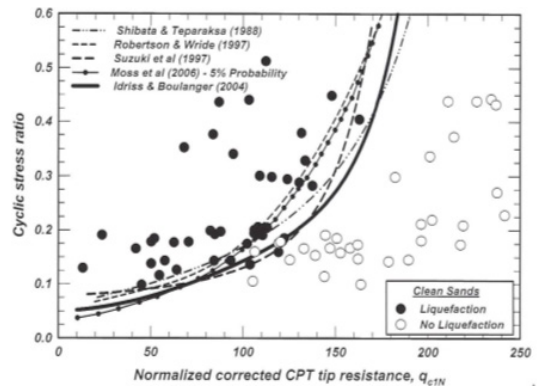


Figure 2. Typical Plots of CSR versus Field SPT Data, from Idriss and Boulanger (2008)

The probability results are usually presented in a plot like Figure 3. This is a generic figure, similar to those presented by most researchers, showing the shape of the typical probability curves, without specifying the parameters for either axis, but representing Bayesian updating with non-informative priors. Since the prior probabilities are non-informative, the prior odds are 1.00. Therefore, the posterior odds must

equal the LR . The plot is presented in terms of the posterior probability p , so the $LR = p/(1-p)$. For example, points in Figure 3 for which $p=0.2$, will have $LR=0.25$.

This is of more than academic interest when dealing with a case for which there is an informed prior. For example, Baise and her colleagues have developed methods for estimating the probability of liquefaction on the basis of geological, meteorological, and historical data (Brankman and Baise, 2008). Consider what happens for a site where their procedure identified an overall probability of liquefaction of 0.8. This is then a prior probability of liquefaction. If the field data for this site fall on the 20% line in Figure 4, corresponding to $LR = 0.25$, the posterior odds of liquefaction become $(0.8/0.2)(0.25) = 1.00$. The posterior probability is 0.50. This compares with the values of 0.20 for the non-informative prior. The prior probability has a big effect.

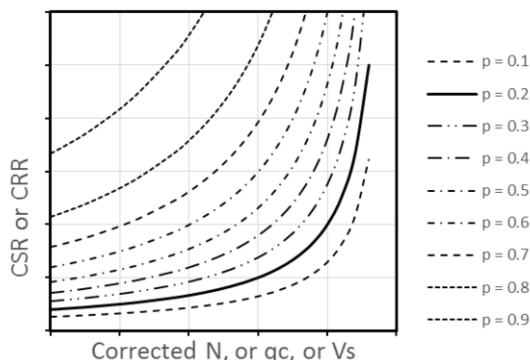


Figure 3. Generic Plot of Probabilities of Liquefaction Reflecting Bayesian Updating

6. Conclusions

Geotechnical engineers face problems with extremely limited numbers of observations (maybe only one), measurements of differing types and quality, a blend of qualitative and quantitative information, and the need to make sequential decisions as data arrive. Frequentist methods are useless in such situations, but Bayesian approaches significantly improve our understanding the weight of evidence.

The first of the above cases demonstrates how a Bayesian formulation leads to a rational

way to process qualitative observations to identify levees sections most likely to require remediation. Unlike other techniques now employed, the method is consistent with probability theory.

The second case describes a rational way to select between or to weight the predictions of different models on the basis of limited observations of field behavior. The example involves slope stability, but the same approach can be extended to other problems such as the dynamic response to earthquake excitation.

The third case shows how generalized empirical methods for liquefaction analysis can be combined with prior estimates of susceptibility to obtain updated probabilities of liquefaction. Publication of curves of Likelihood Ratios would be an improvement over current practice.

Finally, one of the major advantages of Bayesian approaches is that they encourage clear distinction between data and observations on the one hand and field performance on the other.

References

- Baecher, G. B., and Christian, J. T. (2013). Screening Geotechnical Risks. *Foundation Engineering in the Face of Uncertainty*, American Society of Civil Engineers, 215–224.
- Barnard, G. A. (1963). New methods of quality control. *Journal of the Royal Statistical Society Series A*, **126**(2), 255–258.
- Brankman, C. M., and Baise, L. G. (2008) Liquefaction susceptibility mapping in Boston, Massachusetts, *Environmental and Engineering Geoscience*, **14**, 1–16.
- Hoeting, J. A., Madigan, D., Raftery, A., and Volinsky, C. T. (1999). “Bayesian Model Averaging.” *Statistical Science*, **14**(4), 382–417.
- Idriss, I. M., and Boulanger, R. W. (2008) *Soil Liquefaction during Earthquakes*, Earthquake Engineering Research Institute, MNO-12, Oakland, CA, 243 pp.
- Jeffreys, H. (1998). *Theory of probability*. Clarendon Press ; Oxford University Press, Oxford Oxfordshire New York.
- Kayen, R. E., Moss, R. E. S., Thompson, E. M., Seed, Cetin, K. O., Der Keureghian, A., Tanaka, Y., and Tokimatsu, K. (2013). Shear wave velocity-based probabilistic and deterministic assessment of seismic soil liquefaction potential, *Journal of Geotechnical and Geoenvironmental Engineering*, **CE**, **139**(3): 407–419.
- Puhan, M. A., Steuer, J., Bachmann, L. M., and ter Riet. (2005). A randomized trial of ways to describe test accuracy: the effect on physicians’ post-test probability estimates. *Annals of Internal Medicine*, **134**(3), 184–189