The dilemma of statistics[a]

Rigorous mathematical methods cannot compensate messy interpretations and lousy data.

Peter Schuster[b]

Statistics although being indispensible in present day science and society has a bad reputation in particular in public. This can hardly be expressed in a better way than in the famous well-known quotation:

"*There are three kinds of lies: Lies, damned lies, and statistics*".[c]

It would be unfair not to make an attempt to restore the image of statistics and I try to do this in part by means of another citation.

"*While it is easy to lie with statistics, it is even easier to lie without them.*"

This quote is attributed to Frederick Mosteller.[1] Both citations are built undoubtedly upon the association of statistics with telling lies and it is worth asking why statisticians have such a bad image. I feel there are two main reasons for it:
(i) Statistics can never be better than the underlying data, and the sometimes articulated belief of laymen that lousy data can be compensated by the application of highly elaborate statistical tools is simply wrong. What can be done at best is filtering, and the analogy from speech or music is useful as an illustrative example: Proper audio tuning tools can reduce noise and cut out unimportant frequencies but they cannot make an information generating input.
(ii) Terms used in the language of statisticians are frequently misunderstood by the public. One of the problematic notions is "significance": To be "significant" is often understood as a confirmation of the interpretation of data but it is little more than a test of mathematical consistency.
Even in the latter case the appropriate methodology is a matter of ongoing dispute. Here we shall not engage in the diverse problems of data quality and sufficient sample sizes. Instead the issue of attributing significance to hypotheses will be in the focus.

This essay will be dealing with hypothesis testing in mathematical statistics. A paper by Jan Sprenger[2] that appeared in December 2013 prompted me to write on this subject. The review article analyzes the problems of the two major schools in probability theory, the frequentists' approach and the Bayesian method, when they are confronted with statistical tests of the significance of hypotheses. He writes in the summary of his analysis,

> "*A close analysis of the paradox*[d] *reveals that both the Bayesians and frequentists fail to satisfactorily resolve it.*",

---

[a] The essay has been published in *Complexity* **19**/6: x-xx (2014).

[b] Peter Schuster, Editor-in-chief of *Complexity* is at the Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, 1090 Wien, Austria. E-mail: pks@tbi.univie.ac.at

[c] Commonly and probably falsely this quotation has been attributed by Mark Twain to Benjamin Disraeli, Earl of Beaconsfield. For a detailed historical analysis of the origin of this citation see: http://www.york.ac.uk/depts/maths/histstat/lies.htm. Retrieved April 29, 2014.

[d] Jan Sprenger uses Lindley's paradox[3] as a test case for hypothesis testing.

and he proposes an alternative. Why should such a special case be relevant for the image of statistics in the public? The answer is straightforward: When the experts cannot agree on a clear-cut example, the non-experts have all freedom of interpretations and this is particularly problematic if political interest or high risks are involved.

In 1957 Dennis Lindley[3] used an example from Harold Jeffreys book[4] in order to illustrate a general problem in significance tests that became known under the name Lindley's paradox. The paradox points out a substantial flaw of the common significance tests at large sample sizes and moreover, the two approaches, frequentist and Bayesian, inevitably come to opposite conclusions. In a nut shell Lindley has shown that the requirements for the significance of hypotheses may depend on sample size and in particular, two contradictory results obtained by the two methods illustrate the paradox: A sharp null hypothesis H_0, for example exactly equal numbers of males and females in populations ($n_m = n_f$) or a uniform distribution of zeros and ones in a binary sequence leading to the fraction $\theta_0 = ½$, is tested against a diffuse alternative hypothesis H_1, $n_m \neq n_f$ or $\theta \neq \theta_0$, and the test is expected to lead to support or rejection of the null hypothesis. The two proven lemmas say that for any predefined level of significance,[e] $0 \leq \gamma \leq 1$, there exists a population size n such that: (i) in the frequentist language the sample mean m is significantly different from the prediction of H_0 ($p < \gamma$), and (ii) the Bayesian posterior probability $P(H_0|x)$ expressing the probability of H_0 in the light of the data x,[f] for example the probability that $\theta = \theta_0$, is at least as big as $1-\gamma$. In other words there is a sample size above which the Bayesian result contradicts the frequentist result in the sense that the Bayesians confirm the null hypothesis and the frequentists reject it, and this happens inevitably, no matter what the null hypothesis was. The discrepancy becomes larger with increasing sample size n.

Jan Sprenger illustrates Lindley's paradox by means of an example that is taken from experiments dealing with spooky extrasensory capacities,[5] which are already a conflicting issue as such: A test person claims that his extrasensory power could affect a sequence of zeros and ones created by a random number generator using a radioactive source, and consequently his presence would introduce a bias into the otherwise perfectly uniform distribution. In other words, we expect to find a one to one fraction of zeros and ones encapsulated in the null hypothesis H_0 of $\theta = \theta_0 = ½$, and any statistically significant deviation from the ratio $\theta \neq ½$ is attributed to the alternative being the unspecified hypothesis H_1 that would be interpreted as a result extrasensory influence. In the particular case a very large data set of n = 104 490 000 Bernoulli trials was created under the null hypothesis of a binomial distribution with $\theta_0 = (1-\theta_0) = ½$, and yielded $x_1 = x$ = 52 263 471 ones and x_0 = 52 226 529 = 1- x zeros in presence of the test person who claims to use his extrasensory influence to increase the number of ones. The question is

---

[e] The level of significance (γ) expresses the maximal difference between the calculated expectation value of the null hypothesis (μ) and the mean of the measured (experimental) sample (m) that was recorded for testing the null hypothesis. It is commonly expressed as p-value, which measures the fraction of cases, which are more extreme or further away from the null hypothesis than the one under consideration. The result $p>\gamma$ confirms the null hypothesis whereas the null hypothesis is rejected as a consequence of $p<\gamma$.

[f] The expression $P(A|B)$ is a conditional probability, the probability of event A provided B is fulfilled.

now whether or not the result deviates significantly from the uniform distribution assumed under the null hypothesis. First we adopt the frequentists' approach and calculate the *z*-score, which is commonly used for samples with $n > 30$,

$$z(x) = \frac{\frac{x}{n} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} = 3.614 \ ,$$

and then we calculate the *p*-value from the normal distribution and find for this *z*-value: $p = 1 - (3.614) = 0.000151$, where $F_{\mathcal{N}}$ is the cumulative standard normal probability distribution. The null hypothesis is conventionally rejected for $p < \gamma = 0.05$, and in the common frequentist approach the data would be interpreted as evidence for the presence of extrasensory influence on the random number generator. I guess quite a few of my friends from physics are now inclined to stop reading this essay, but I promise this would be premature since we are not yet at the end.

How would a Bayesian analyze the data?[6] The basis of the Bayesian approach is Bayes' theorem

$$P(H \mid x) = \frac{P(x \mid H) \cdot P(H)}{P(x)}$$

where $P(H)$ is the so-called prior probability of hypothesis H, x are the data, in our example $x = x_1$, and $P(H|x)$ is the posterior probability, the probability of H under the condition given by the data x. For two hypotheses, H_0 and H_1, we calculate the so-called Bayes-factor expressing the significance of hypothesis H_0 versus hypothesis H_1:

$$B_{01} = \frac{P(x \mid H_0)}{P(x \mid H_1)} = \frac{P(H_0 \mid x)}{P(H_0)} \cdot \frac{P(H_1)}{P(H_1 \mid x)}$$

For the interpretation of Bayes factors Harold Jeffreys gave a scale (ref.4, p.432):

$B_{01}<1$ support for H_1,
$1<B_{01}<3$ insignificant support for H_0,
$3<B_{01}<10$ substantial support for H_0,
$10<B_{01}<30$ strong support for H_0,
$30<B_{01}<100$ very strong support for H_0, and
$B_{01}>100$ decisive support for H_0.

The calculation in case of our example occurs as follows: A positive probability is assigned to the null hypothesis, $P(H_0) = \varepsilon > 0$, and a uniform prior distribution is attributed to the alternative hypothesis H_1. A Bayes-factor of $B_{01} \approx 12$ is then calculated from the conditional probabilities, which expresses strong support for the null hypothesis that was no extrasensory force in action. The majority of scientists will take note of this result with satisfaction but "*don't celebrate just yet*", because as predicted by Lindley's paradox you need only take a sufficiently large sample and the Bayes factor will favor the null hypothesis.

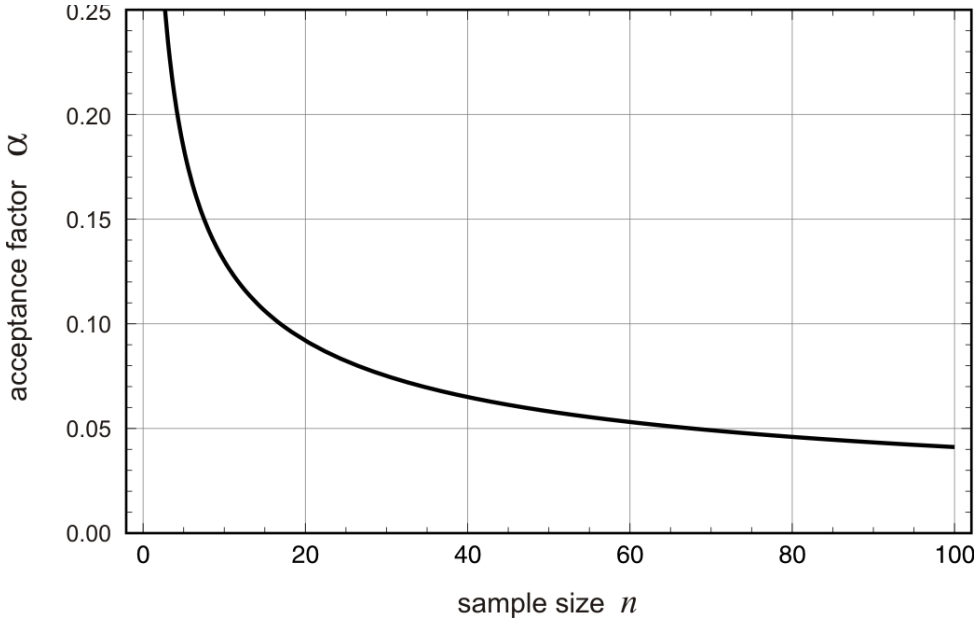

Figure 1: **Acceptance factor α as a function of sample size**. The acceptance factor α defined in equation (1) is plotted against the sample size *n*. It converges asymptotically with $n^{-½}$ to value $\alpha=0$ meaning certain rejection of the null hypothesis H_0. Parameter choice: $\theta_0=½$ and $z(x)=1.645$ corresponding to $\gamma=0.05$.

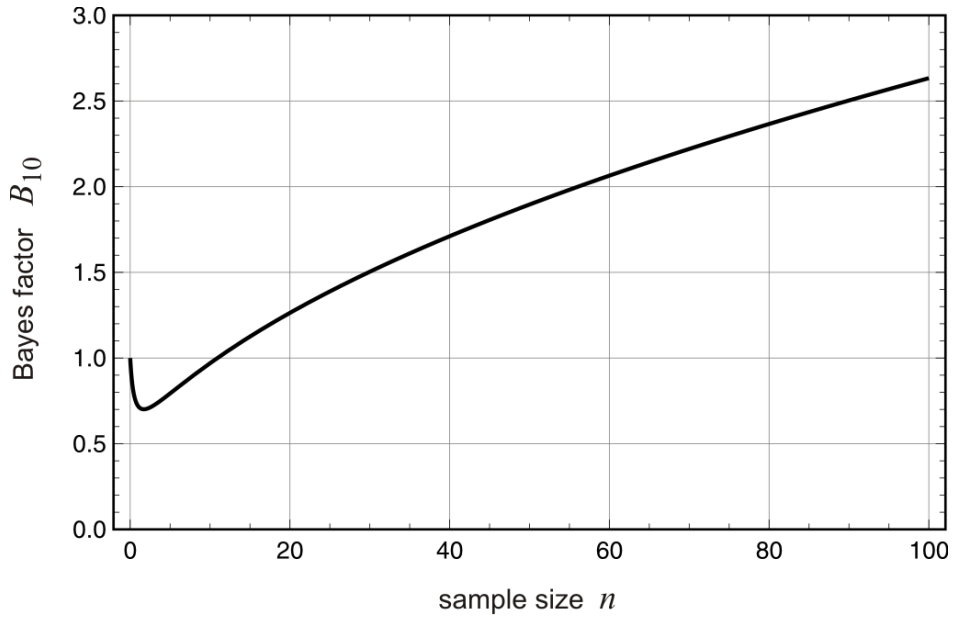

Figure 2: **Bayes factor B_{01} as a function of sample size**. The Bayes factor B_{01} defined in equation (2) is a measure for the preference of the null hypothesis H_0 over the alternative H_1. Here it is plotted against the sample size *n*. It increases asymptotically with $n^{½}$ implying more and more preference for the null hypothesis H_0. Parameter choice: $\sigma^2=s^2=¼$ and $z(x)=1.645$.

In order to visualize Lindley's paradox we use two functions, an acceptance factor

$$\alpha = \sqrt{\frac{\theta_0 \cdot (1-\theta_0)}{n}} \cdot z(x) \tag{1}$$

for the frequentist approach and an approximation for the Bayes factor[2]

$$B_{01} = \sqrt{1+\frac{n \cdot \sigma^2}{s^2}} \cdot \exp\left(-\frac{n \cdot z(x)^2}{2(n+s^2/\sigma^2)} \right) \tag{2}$$

where σ^2 is the variance of the prior of H_1 and s^2 the variance of the sample. In order to compare the two expressions we assume a constant *z*-score and take the value obtained for the acceptance limit of the frequentist approach: $z(\gamma) = z(0.05) = 1.645$. All variances are chosen to be $\sigma^2 = s^2 = \theta_0(1-\theta_0) = ¼$, and *n* is the variable under consideration. The corresponding curves are shown in figures 1 and 2. The high sample size limits are seen straightforwardly from the two equations: The acceptance factor α decreases as $1/\sqrt{n}$ and approaches the asymptotic value $\alpha = 0$ more and more closely the larger the sample size *n* is. Any kind of small fluctuation will eventually lead to a rejection of the null hypothesis. In the Bayesian approach the exponential term of the factor B_{01} approaches a constant, $\exp(-z(x)^2/2)$, and in total the Bayesian factor grows with $\sqrt{n}$. No matter how improbable the null hypothesis H_0 actually was, there will be a population size *n* above which the data will favor it. Coming back to our initial claim this result is grist to the mill of the skeptic layman: In order to get a desired result you need only choose the right method! Although one has to admit that a sharp null hypothesis, $\theta = \theta_0$, together with a huge sample size, $n > 10^8$, is not the kind of problem that is often encountered by a statistician – except perhaps, in the analysis of extrasensory perception – the situation is, nevertheless, highly unsatisfactory.

Is there a way out of the dilemma? Jan Sprenger presents José Bernardo's Bayesian reference criterion (BRC) as a possibility for recovery[7]. The remedy comes from shifting the focus from evaluating the 'truth' of H_0 to considering its predictive value and its prediction-based utilities. Similar as in the frequentist mood,[8] the reproduction of previously observed regularities is understood as the main motivation for significance tests in Bernardo's model, which in essence is built upon the Bayesian concept combined with decision theory. In order to evaluate the predictive score *q* of a hypothesis *H* expressed in form of a parameter value θ given some data *y* the function

$$q(\theta, y) = a \log P(y \mid \theta) + b(y)$$

is applied, where *a* is a scaling factor and $b(y)$ is a function that does not depend on the parameter θ. For prediction only the first part is relevant because $b(y)$ cannot be adjusted by varying θ. The basis for a decision is the utility function *U* that depends on the predictive score *q* and the cost *C* for selecting a particular hypothesis *H*. For a null hypothesis H_0, which is assumed to be simpler, more informative, and less prone to overfitting, the costs should be smaller than for a specific hypothesis H_1 and we set $C_1 > C_0$. Bernardo defines the utilities of accepting H_0 or H_1 by

$$U(H_0,\theta) = \int dy \big(a \log P(y \mid \theta_0) + b(y)\big) P(y \mid \theta) - C_0$$

$$U(H_1,\theta) = \int dy \big(a \log P(y \mid \theta) + b(y)\big) P(y \mid \theta) - C_1$$

It is important to see two features that make the BRC method essentially different from the conventional Bayesian approach: (i) The utility of accepting H_0 is evaluated against the true parameter θ and not against the point null hypothesis θ_0, and this introduces an essential asymmetry between H_1 and H_0, and (ii) the alternative H_1 is not represented by the probabilistic average in form of the posterior mean but by its best element that is, of course, unknown.

The Bayesian reference criterion (BRC) can be cast in a brief lemma: Data x, which have been generated from the probability model $P(x|\theta)$ with $\theta \in \Theta$,[g] are incompatible with the null hypothesis H_0, $\theta = \theta_0$, if and only if the expected intrinsic discrepancy

$$d = \int_{\theta \in \Theta} d\theta \; P(\theta \mid x) \left(\int dy \log \frac{P(y \mid \theta)}{P(y \mid \theta_0)} \; P(y \mid \theta_0) \right) > \; d_0(a, C_1 - C_0)$$

exceeds a threshold value d_0.[2,7] Simple choices of d_0 are $\ln 10 \approx 2.3$ for mild, $\ln 100 \approx 4.6$ for moderate, and $\ln 1000 \approx 6.9$ for strong evidence against the null hypothesis. It is possible to compare the threshold d_0 with the frequentist choice of the threshold level, $\gamma = 0.05$, which corresponds to $d_0 \approx \ln 11$ or mild evidence against H_0 and which can explain the often false rejections reported in the literature. With the BRC a firm connection is made between hypothesis testing and decision theory. The new approach at the same time extends Bayesian reasoning and puts the frequentist procedures on a firm ground.

Finally, we apply José Bernardo's method to Lindley's paradox and, in particular, to the analysis of the extrasensory capacity case.[5,7] His calculation of the expected intrinsic discrepancy yields $d = \ln 1400 = 7.24$ and the null hypothesis is strongly rejected as it was in the frequentist case. The results can be easily verified by visualizing the posterior distribution of the ratio θ, which is a normal distribution $\mathcal{N}$ with mean $\mu_\theta = 0.50018$ and standard deviation $\sigma_\theta = 0.000049$. Accordingly, the null hypothesis is 3.614 standard deviations away from the mean as already expressed by the z-score. Does this also mean that spooky extrasensory influence is present in the experiment? Definitely not! Despite being significant such a small deviation from the unbiased value $\theta_0 = ½$ can have a great variety of other causes: Most likely the random number generator was not accurate enough for such large samples.[2] In view of the (infinitely) sharp hull hypothesis, $\theta = \theta_0$, the very large sample size amplifies all kinds of otherwise unnoticeable and sometimes spurious effects.

In this essay we made an attempt to identify problems in statistics and statistical inference that might contribute to the bad reputation of statistics in the public. At first we showed with reference to Lindley's paradox that there are fundamental differences between the

[g] The entire sample space of possible alternative hypotheses is denoted by Θ.

frequentist approach and the subjective Bayesian approach that cannot be reconciled by more careful analysis. This becomes particularly clear through the consideration of sharp null hypotheses, e.g. $\theta = \theta_0$, at very large sample sizes: The frequentists will always reject the null hypothesis whereas the Bayesians will always confirm it. The fact that fixed significance levels used in the frequentists' analysis of samples with different sizes are problematic, was already well known at the time of Lindley's publication. It is clearly reflected by a comment of Maurice Bartlett.[9] One way out of the dilemma is José Bernardo's objective Bayesian approach called Baesian reference criterion (BRC), which merges the Bayesian method and the frequentist goal of precise reproduction of data by introducing rigorous methods from decision theory. In the special case of the extrasensory experiment discussed here[5] the BCR analysis provides essentially the same result as the frequentist significance analysis.

Now a different issue comes into play: Rejection of the null hypothesis at large sample sizes does not imply that the conjectured alternative is true, because at extremely high resolution all kinds of effects become important that are spurious and play no role for smaller samples. An important take-home lesson is: Increasing sample sizes alone in no universal remedy for bad quality results unless a careful analysis of possible artifacts at the higher resolution is undertaken. Misinterpretations like evidence for extrasensory power in the quoted example will definitely reflect discredit on statistics. The literature on significance tests and frequent pitfalls in their application is enormous and there is no need at present to fall into the common traps of testing hypotheses. If the majority of researchers were more rigorous in using statistical tools and more careful in the interpretations, the reputation of statistics would definitely be improved.

---

[1] Murray, C. How to accuse the other guy of lying with statistics. Statistical Science, 2005, 20, 239-241.

[2] Sprenger, J. Testing a precise null hypothesis: The case of Lindley's paradox. Philosophy of Science, 2013, 80, 733-744.

[3] Lindley, D.V. A statistical paradox. Biometrika, 1957, 44, 187-192.

[4] Jeffreys, H. Theory of probability. Third ed. Oxford University Press: Oxford, UK, 1961.

[5] Jahn, R.G., Dunne, B.J., Nelson, R.D. Engineering anomalies research. J. of Scientific Exploration 1, 21-50, 1987.

[6] Jefferys, W.H. Bayesian analysis of random event generator data. J. of Scientific Exploration 4, 153-168, 1990.

[7] Bernardo, J.M. Integrated objective Bayesian estimation and hypothesis testing. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith., A.F.M., West. M., Eds. Bayesian Statistics 9. Oxford Unversity Press, Oxford UK, 2012, pp. 1-68.

[8] Schmidt, F.L., Hunter, J.E. Eight but false objections to the discontinuation of significance testing in the analysis of research data. In: Harlow, L.L., Mulaik, S.A., Steiger, J.H., Eds. What if there were no significance tests? Lawrence Erlbaum Associates, Mawah, NJ, 1997, pp. 37-64.

[9] Bartlett, M.S. A comment on D.V. Lindley's statistical paradox. Biometrika, 1957, 44, 533-534.