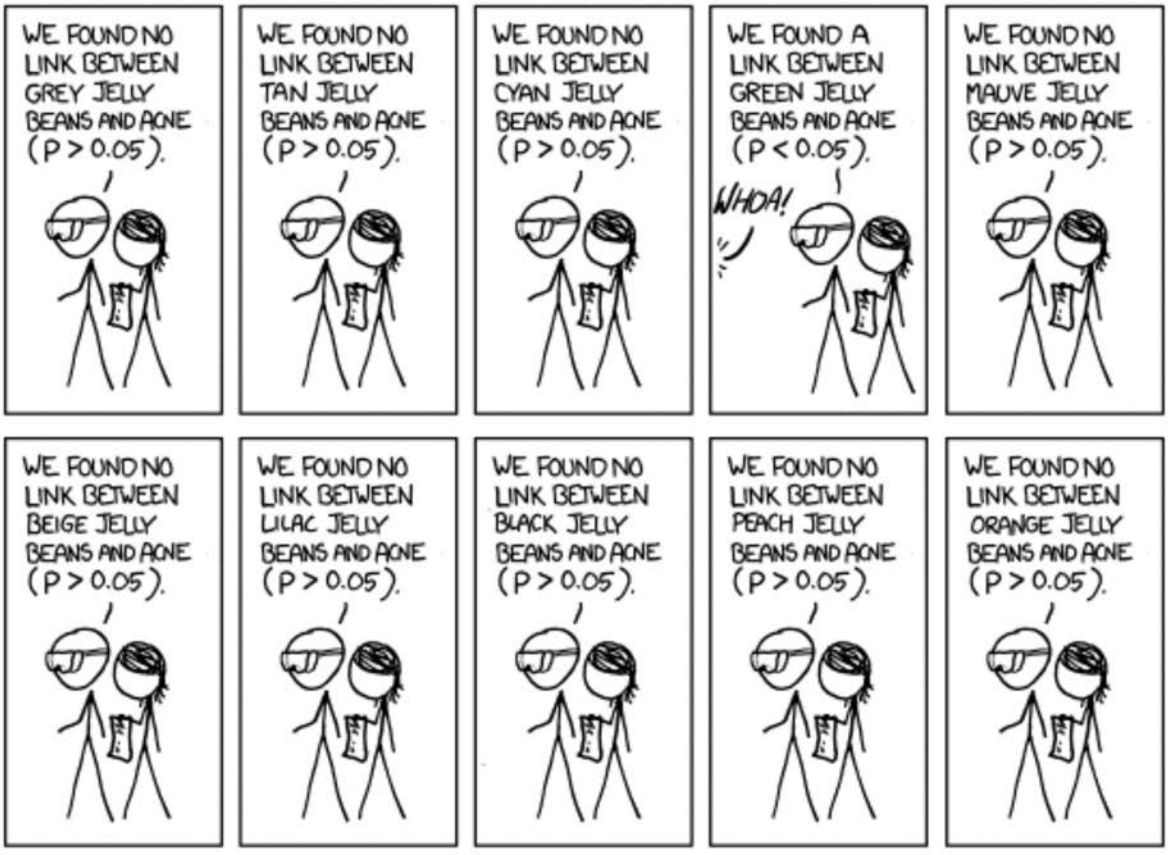
DATA ANALYSIS: A (NOT SO) SILENT REVOLUTION

Udo von Toussaint – Max-Planck-Institute for Plasmaphysics – 85748 Garching, Germany – DOI: https://doi.org/10.1051/epn/2017101

Good statistical practice is an essential component of good science. Unfortunately, in too many cases the analysis of experiments follows recipes taught already in introductory classes, sometimes even accompanied by misunderstandings about properties of statistical tests. One of these methods, significance tests based on p-values, was *the* widely accepted standard. So the warning by the American Statistical Association (ASA) against the use of p-values was thus a huge surprise to many. What were the reasons underlying this unique step?



Statistical reasoning...

It may have escaped the attention of most physicists that the recent statement [1] of the largest statistical organisation, the American Statistical Association (ASA), on the proper use of p-values sent shock waves through the community of statisticians. In a comment published in Science [2] the seriousness of the statement was even compared to a confusion of weight and mass in physics: *'Imagine the American Physical Society convening a panel of experts to issue a missive to the scientific community on the difference between weight and mass. And imagine that the impetus for such a message was a recognition that engineers and builders had been confusing these concepts for decades, making bridges, buildings, and other components of our physical infrastructure much weaker than previously suspected*'. However, even more surprising to many people was the large number of follow-up comments to the ASA recommendation, most of which supported the view that deficiencies of the concept of p-values have long been known and that the common practice with respect to hypothesis testing needs to be altered (see, *e.g.*, supplemental material in [1]). This may come as a surprise to many scientists who – almost by default – have been asked to provide p-values for suggested hypotheses.

To understand why the standard approach for assessing the validity of a new hypothesis has fallen out of favour it may be useful to recollect the concept of the p-value approach.

Testing a hypothesis using the p-value approach

The standard statistical approach to test a precise hypothesis follows the subsequent scheme:

1. Formulate a hypothesis H_0 (the 'null hypothesis')
2. Choose a test variable x of which the probability distribution (*i.e.*, the probability density function, pdf) $p(x|\ H_0)$ is known for the hypothesis H_0.
3. Choose an interval $[x_1,x_2]$ such that there is the probability α that x falls outside the interval: $P(x_1 \le x \le x_2) = 1 - \alpha$, with α commonly chosen as $\alpha = 0.05$
4. Perform an experiment measuring, *e.g.*, the data value x_m
5. Decide depending on the value of x_m: if $x_1 \le x_m \le x_2$, then accept the hypothesis H_0

else H_0 is rejected with significance level α.

The conventional procedure is illustrated in Fig.1. If the probability to observe x_m or values even more extreme is low (*i.e.*, below α) under the hypothesis H_0 then it is concluded that the hypothesis is unlikely. The underlying reasoning of the p-value approach is analogous to the familiar *proof by contradiction* [3]: One first assumes a hypothesis to be true but if a consequence of this hypothesis turns out to be false then the hypothesis can be rejected. In the p-value approach there are two critical modifications: a) the 'false' consequence is replaced by 'unlikely' consequence and b) the probability of any specific outcome $x = x_k$ is typically very small, such that every result would be significant. For that reason the integrated probability for the observation of x_m or even more extreme values is used ('tail probability'). Thus the p-value also depends on the probability of data which have not been observed, *i.e.*, in the present example all values of x with $|x| > x_m$.

So, given these apparently small differences from the logically sound approach of proof by contradiction, why then has it fallen out of favour, even to the extent that the comment of the ASA [1] states in No. 5: *'A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.'*?

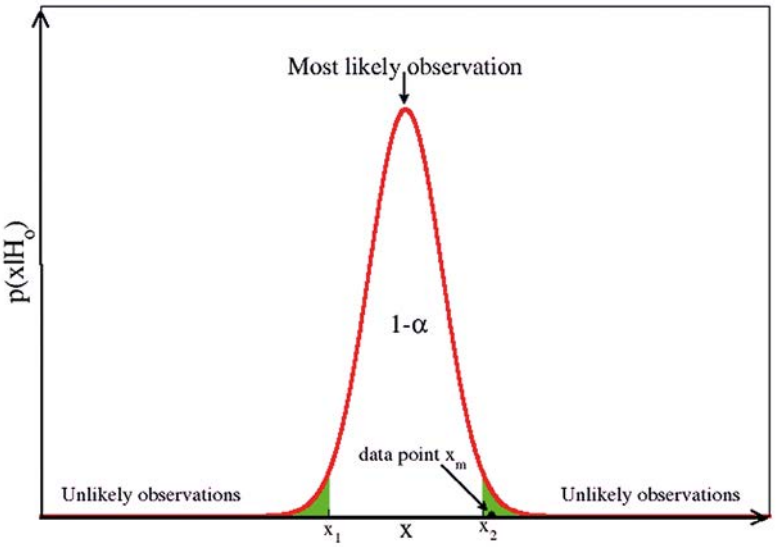
This can be better understood from a Bayesian perspective.

Bayesian model testing

Bayesian model testing is always based on the comparison of at least two different proposed models. There is no counterpart to the conventional significance tests which evaluate a model or hypothesis based on only a single model. The relevant probabilities are P(H|*data*) and P(*data*|H).

The posterior probability P(H|*data*) denotes the probability of the hypothesis H in the light of an observed data set data. More familiar is the probability P(*data*|H), the likelihood which describes the probability of the observables under the condition that the hypothesis H is true. It is important to note that these two probabilities are in general different. Consider, *e.g.*, P(red liquid| blood) *vs.* P(blood| red liquid). The probability of blood being a red liquid is close to one. However, the probability of a red liquid to be actually blood is much lower. A consequence of this is that maximum-likelihood estimates of parameters are only under very specific circumstances the most likely parameters given the data.

The Bayesian approach to model comparison simply computes the ratio of the probabilities for each model, which is given by the ratio of the likelihoods of the data



▼ **FIG. 1:** Hypothesis test scheme. Indicated is the probability density p(x|H) of the observation of x as function of x. The area of the shaded regions sums up to α. The data point at x_m would result in a rejection of the null-hypothesis with significance level α because the point is outside of the interval $[x_1,x_2]$.

under the models considered and the prior probabilities of the respective models, $P(H_i)$ and $P(H_j)$

$$\frac{P(H_i \mid Data)}{P(H_j \mid Data)} = \frac{P(Data \mid H_i)}{P(Data \mid H_j)} \times \frac{P(H_i)}{P(H_j)}$$

which can be derived in a straightforward manner from the elementary sum- and product rules of probability theory. This approach is also undisputed by the conventional statistics community if the prior probabilities are available. Taking this point of view it becomes evident where the problems of the p-value approach for statisticians are rooted:

Firstly: The significance test yields only P(*data*|H). However, the researcher is in practice interested in the probability of the hypothesis, given the data, P(H|*data*). Thus, in practice the significance levels are often incorrectly considered as probabilities of the hypothesis – despite all the warnings to the contrary of statistics experts. 'The null-hypothesis is rejected at 5% significance' is taken to mean the same as 'The null-hypothesis has only 5% probability to be true' [3].

Secondly, the p-value is based on the likelihood term only and neglects the prior contribution. This implies that the p-value measures whether an observed result can be attributed to chance under the null-hypothesis. However, the probability that the hypothesis is correct depends on how plausible the hypothesis was in the first place. If the a-priori probability for a hypothesis is 5% then a measured p-value of 0.05 (*i.e.*, a 'statistically significant' result) increases the probability of that hypothesis to be true by only 6%! [4]. If the a-priori probability of the hypothesis is larger, *e.g.*, 50% then the same result already implies a posterior-probability for the hypothesis of 71%. Thus, significance levels (p-values) can be a highly misleading measure of the evidence provided by the data against a null hypothesis [5,6,7].

This leaves us with the question: why was the p-value approach (apparently) performing not too badly in many cases? A possible explanation is the following: Hypothesis tests have often been performed when doubts about the null hypothesis were growing and reasonable alternatives became available. In that case the (implicit) prior ratio is approximately 1 and the likelihood of the alternative is larger than that of the original hypothesis, thus resulting in a similar outcome as the Bayesian approach (without the possibility of a quantitative measure of the model probabilities).

The third drawback – at least from a Bayesian perspective but also in practice – is the violation of the likelihood principle by the significance test [8,9]. *The likelihood principle* is implied by the generally accepted *sufficiency principle* [10] conditionally on the acceptance of a second principle, the *conditionality principle*: If two experiments on the parameter x, E_1 and E_2, are available and if one of these two experiments is selected with probability 0.5, the resulting inference on x should depend only on the selected measurement. This principle seems difficult to reject [11]. The violation of the likelihood principle introduces a dependence of the significance test result on unobserved data or stopping rules, which is criticized by Bayesian proponents [12, 13] because an evaluation based on the measured data only is not possible. A nice toy example of the induced problems has been given in [14].

The recommendations of the American Statistical Association conclude that in the view of *'the prevalent misuses and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches[...] No single index should substitute for scientific reasoning'*. This is now almost identical to the opinion of Fisher [15] – who introduced the p-value in the 1920s – already 90 years ago. He stated that *'the responsibility of forming correct scientific conclusions cannot be replaced by automated acceptance procedure'*. ASA's statement on p-values should thus be considered as an overdue start towards a more domain-knowledge-inspired and less entrenched approach to data analysis. ■

About the Author

Udo von Toussaint is head of the PMMF-group at the Max-Planck-Institute for Plasmaphysics. He received his PhD in physics from the University in Bayreuth. His research interests are Bayesian inference, inverse problems and machine learning.

References

[1] R.L. Wasserstein, N.A. Lazar, *Am. Statistician* **70:2**, 129 (2016) DOI:10.1080/00031305.2016.1154108

[2] S.N. Goodman, *Science* **352:6290**, 1180 (2016)

[3] G. D'Agostini, *Bayesian Reasoning in Data Analysis*, World Scientific Publishing, Singapore (2003)

[4] R. Nuzzo, *Nature* **506**, 150 (2014)

[5] F. Davidoff, *Ann. Intern. Med.* **130**, 1019 (1999)

[6] S.N. Goodman, *Ann. Intern. Med.* **130**, 995 (1999)

[7] S.N. Goodman, *Ann. Intern. Med.* **130**, 1005 (1999)

[8] A. Birnbaum, *J. Am. Stat. Assoc.* **57**, 269 (1962)

[9] J.O. Berger and R.L. Wolpert, *The Likelihood Principle* (IMS, Hayward, CA), 2nd ed. (1988)

[10] V.S. Huzurbazar, *Sufficient Statistics* (Marcel Dekker, New York) (1976)

[11] C.P. Robert, *The Bayesian Choice* (Springer, Berlin) (1994)

[12] J.O. Berger and T. Sellke, *J. Am. Stat. Assoc.* **82**, 112 (1987)

[13] T.J. Loredo in *Statistical Challenges in Modern Astronomy*, edited by E. Feigelson and G. Babu (Springer, New York), p. 275 (1992)

[14] J.O. Berger and D.A. Berry, *Am. Scientist* **76**, 159 (1988)

[15] R.A. Fisher, *Statistical Methods and Scientific Inference* (Hafner, New York) (1956)