POINTS OF SIGNIFICANCE

# Interpreting *P* values

A *P* value measures a sample's compatibility with a hypothesis, not the truth of the hypothesis.

Although *P* values are convenient and popular summaries of experimental results, we can be led astray if we consider them as our only metric[1]. Even in the ideal case of a rigorously designed randomized study fit to a predetermined model, *P* values still need to be supplemented with other information to avoid misinterpretation.

A *P* value is a probability statement about the observed sample in the context of a hypothesis, not about the hypotheses being tested. For example, suppose we wish to know whether disease affects the level of a biomarker. The *P* value of a comparison of the mean biomarker levels in healthy versus diseased samples would be the probability that a difference in means at least as large as the one observed can be generated from random samples if the disease does not affect the mean biomarker level. It is not the probability of the biomarker-level means in the two samples being equal—they either are or are not equal.

However, this relationship between *P* values and inference about hypotheses is a critical point—interpretation of statistical analysis depends on it. It is one of the key themes in the American Statistical Association's statement on statistical significance and *P* values[2], published to mitigate widespread misuse and misinterpretation of *P* values. This relationship is discussed in some of the 18 short commentaries that accompany the statement, from which three main ideas for using, interpreting and reporting *P* values emerge: the use of more stringent *P* value cutoffs supported by Bayesian analysis, use of the observed *P* value to estimate false discovery rate (FDR), and the combination of *P* values and effect sizes to create more informative confidence intervals. The first two of these ideas are currently most useful as guidelines for assessing how strongly the data support null versus alternative hypotheses, whereas the third could be used to assess how strongly the data support parameter values in the confidence interval. However, like *P* values, these methods will be biased toward the alternative hypothesis when used with a *P* value selected from the most significant of multiple tests or models[1].

To illustrate these three ideas, let's expand on the biomarker example above with the null hypothesis that disease does not influence the biomarker level. For samples, we'll use $n = 10$ individuals, randomly chosen from each of the unaffected and affected populations, assumed to be normally distributed with $\sigma^2 = 1$. At this sample size, a two-sample *t*-test has 80% power to reject the null at significance $\alpha = 0.05$ when the effect size is 1.32 (**Fig. 1a**). Suppose that we observe a difference in sample means of 1.2 and that our samples have a pooled s.d. of $s_p = 1.1$. These give us $t = 1.2/(s_p\sqrt{(2/n)}) = 2.44$ with d.f. $= 2(n - 1) = 18$ and a *P* value of 0.025.

Once a *P* value has been computed, it is useful to assess the strength of evidence of the truth or falsehood of the null hypothesis. Here we can look to Bayesian analysis for ways to make this connection[3], where decisions about statistical significance can be based on the Bayes factor, *B*, which is the ratio of average likelihoods under the alternative and null hypotheses. However, using Bayesian analysis adds an element of subjectivity because it requires the specification of a prior distribution for the model parameters under both hypotheses.

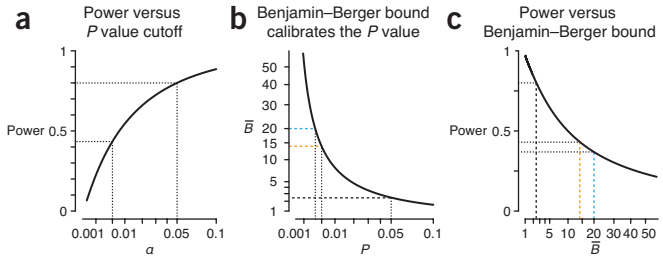


**Figure 1** | Using a Bayesian heuristic to interpret the *P* value. (**a**) Power drops at more stringent *P* value cutoffs $\alpha$. The curve is based on a two-sample *t*-test with $n = 10$ and an effect size of 1.32. (**b**) The Benjamin and Berger bound calibrates the *P* value to probability statements about the hypothesis. At $P = 0.05$, the bound suggests that our alternative hypothesis is at most 2.5 times more likely than the null (black dashed line). Also shown are the conventional Bayesian $\bar{B} = 20$ (blue dashed line; $P = 0.0032$) cutoff and $\bar{B} = 14$ (orange dashed line; $P = 0.005$), suggested by Johnson in ref. 2. (**c**) Use of the more stringent Benjamin and Berger bounds in **b** reduces the power of the test, because now testing is performed at $\alpha < 0.05$. For $\bar{B} = 14$ (orange dashed line; $\alpha = 0.005$), the power is only 43%. The blue and orange dashed lines show the same bounds as in **b**. In all panels, black dotted lines are present to help the reader locate values discussed in the text.

Benjamin and Berger, in their discussion in ref. 2, note that the *P* value can be used to compute an upper bound for the Bayes factor, $\bar{B}$. The bound does not require the specification of a prior and holds for many reasonable choices of priors. For example, $\bar{B} = 10$ means that the alternative hypothesis is at most ten times more likely to be true than the null.

Because it quantifies the extent to which the alternative hypothesis is more likely, the Bayes factor can be used for significance testing. Decision boundaries for the Bayes factor are less prescriptive than those for *P* values, with descriptors such as "anectodal," "substantial," "strong" and "decisive" often used for cutoff values. The exact terms and corresponding values vary across the literature, and their interpretation requires active consideration on the part of the researcher[4]. A Bayes factor of 20 or more is generally considered to be strong evidence for the alternative hypothesis.

The Benjamin and Berger bound is given by $\bar{B} \le -1/(e\,P\ln(P))$ for a given *P* value[5] (**Fig. 1b**). For example, when we reject the null at $P < \alpha = 0.05$, we do so when the alternative hypothesis is at most $\bar{B} \le 2.5$ times more likely than the null! This significance boundary is considered by many Bayesians to be extremely weak to nonexistent evidence against the null hypothesis.

For our biomarker example, we found $P = 0.025$ and thus conclude that the alternative hypothesis that disease affects the biomarker level is at most $\bar{B} \le 3.9$ times more likely than the null. If we insist on $\bar{B} > 20$, which corresponds to 'strong' evidence for the alternative, we need $P < 0.0032$ (**Fig. 1b**). Johnson, in a discussion in ref. 2, suggests testing at $P < \alpha = 0.005$ ($\bar{B} > 14$) for statistical significance (**Fig. 1b**). Notice that the computation for $\bar{B}$ does not use the power of the test. If we compute power using the same effect size of 1.32 but reject the null at $\alpha < 0.005$ ($\bar{B} > 14$), the power is only 43% (**Fig. 1c**). To achieve 80% power at this cutoff, we would need a sample size of $n = 18$.

Altman (this author), in a discussion in ref. 2, proposes to supplement *P* values with an estimate of the FDR by using plug-in values to account for both the power of the test and the prior evidence in

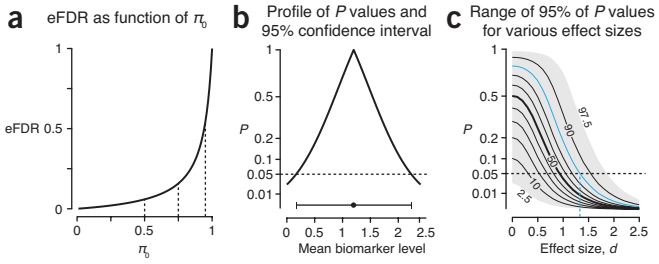


**Figure 2** | Interpretation of the *P* value with heuristics based on the false discovery rate (FDR) and by examination of *P* values across a range of hypotheses. (**a**) The relationship between the estimated FDR (eFDR) and the proportion of tests expected to be null, $\pi_0$, when testing at $\alpha = 0.05$. Dashed lines indicate Altman's proposals[2] for $\pi_0$. (**b**) The profile of *P* values for our biomarker example ($n = 10$, $s_p = 1.1$). The dashed line at $P = 0.05$ cuts the curve at the boundaries of the 95% confidence interval (0.17, 2.23), shown as an error bar. (**c**) *P* value percentiles (shown by contour lines) and 95% range (gray shading) expected from a two-sample *t*-test as effect size is increased. At each effect size *d*, data were simulated from 100,000 normally distributed samples ($n = 10$ per sample) with means 0 and *d*, respectively, and $\sigma^2 = 1$. The fraction of *P* values smaller than $\alpha$ is the power of the test—for example, 80% (blue contour) are smaller than 0.05 for $d = 1.32$ (blue dashed line). When $d = 0$, *P* values are randomly uniformly distributed.

favor of the null hypothesis. In high-throughput multiple-testing problems, the FDR is the expected proportion of the rejected null hypotheses that consists of false rejections. If some proportion $\pi_0$ of the tests are truly null and we reject at $P < \alpha$, we expect $\alpha\pi_0$ of the tests to be false rejections. Given that $1 - \pi_0$ of the tests are non-null, then with power $\beta$ we reject $\beta(1 - \pi_0)$ of these tests. So, a reasonable estimate of the FDR is the ratio of expected false rejections to all expected rejections, $\text{eFDR} = \alpha\pi_0/(\alpha\pi_0 + \beta(1 - \pi_0))$.

For low-throughput testing, Altman uses the heuristic that $\pi_0$ is the probability that the null hypothesis is true as based on prior evidence. She suggests using $\pi_0 = 0.5$ or 0.75 for the primary hypotheses or secondary hypotheses of a research proposal, respectively, and $\pi_0 = 0.95$ for hypotheses formulated after exploration of the data (*post hoc* tests) (**Fig. 2a**). In the high-throughput scenario, $\pi_0$ can be estimated from the data, but for low-throughput experiments Altman uses the Bayesian argument that $\pi_0$ should be based on the prior odds that the investigator would be willing to put on the truth of the null hypothesis. She then replaces $\alpha$ with the observed *P* value, and $\beta$ with the planned power of the study.

For our example, using $P = 0.025$ and 80% power gives eFDR = 0.03, 0.09 and 0.38 for primary, secondary and *post hoc* tests, respectively (**Fig. 2a**). In other words, for a primary hypothesis in our study, we estimate that only 3% of the tests where we reject the null at this level of *P* are actually false discoveries, but if we tested only after exploring the data, we would expect 38% of the discoveries to be false.

Altman's 'rule-of-thumb' values for $\pi_0$ are arbitrary. A simple way to avoid this is to determine the value of $\pi_0$ required to achieve a given eFDR. For example, to achieve eFDR = 0.05 for our example with 80% power, we require $\pi_0 \leq 0.62$, which is fairly strong prior evidence for the alternative hypothesis. For our biomarker example, this might be reasonable if studies in other labs or biological arguments suggest that this biomarker is associated with disease status, but it is unreasonable if multiple models were fitted or if this is the most significant of multiple biomarkers tested with little biological guidance.

Many investigators and journals advocate supplementing *P* values with confidence intervals, which provide a range of effect sizes compatible with the observations. Mayo, in a discussion in ref. 2, suggests considering the *P* value for a range of hypotheses. We demonstrate this approach in **Figure 2b**, which shows the *P* values of other levels of the biomarker in comparison to one that is observed. The 95% confidence interval, which is (0.17, 2.23) for this example, is the range of levels that are not significantly different at $\alpha = 0.05$ from the observed level of 1.2.

As a final comment, we stress that *P* values are random variables—that is, random draws of data will yield a distribution for the *P* value[1]. When the data are continuous and the null hypothesis is true, the *P* value is uniformly distributed on (0,1), with a mean of 0.5 and s.d. of $1/\sqrt{12} \approx 0.29$ (ref. 1). This means that the *P* value is very variable from sample to sample, and this variability is not a function of the sample size or the power of the study. When the alternative hypothesis is true, the variability decreases as the power increases, but the *P* value is still random. We show this in **Figure 2c**, in which we simulate 100,000 sample pairs for each mean biomarker level.

*P* values can provide a useful assessment of whether data observed in an experiment are compatible with a null hypothesis. However, the proper use of *P* values requires that they be properly computed (with appropriate attention to the sampling design), reported only for analyses for which the analysis pipeline was specified ahead of time, and appropriately adjusted for multiple testing when present. Interpretation of *P* values can be greatly assisted by accompanying heuristics, such as those based on the Bayes factor or the FDR, which translate the *P* value into a more intuitive quantity. Finally, variability of the *P* value from different samples points to the need to bring many sources of evidence to the table before drawing scientific conclusions.

**COMPETING FINANCIAL INTERESTS**
The authors declare no competing financial interests.

**Naomi Altman & Martin Krzywinski**

1. Altman, N. & Krzywinski, M. *Nat. Methods* **14**, 3–4 (2017).
2. Wasserstein, R. & Lazar, N.A. *Am. Stat.* **70**, 129–133 (2016).
3. López Puga, J., Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 277–278 (2015).
4. Jarosz, A.F. & Wiley, J.J. *J. Probl. Solving* **7**, 2 (2014).
5. Selke, T. *et al. Am. Stat.* **55**, 62–71 (2001).

Naomi Altman is a Professor of Statistics at The Pennsylvania State University. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.