POINTS OF SIGNIFICANCE

P values and the search for significance

Little *P* value
What are you trying to say
Of significance?

—Steve Ziliak

The significance of experimental results is often assessed using *P* values and estimates of effect size. However, the interpretation of these assessment tools can be invalidated by selection bias when testing multiple hypotheses, fitting multiple models or even informally selecting results that seem interesting after observing the data. Our goal this month will be to identify some circumstances that can give rise to such questionable practices—broadly termed '*P* value hacking' and 'data dredging'. In addition, statistically significant results may not translate into biologically meaningful conclusions—with large sample sizes or small variability, even tiny effects can be statistically significant.

We have previously seen how to correctly interpret *P* values in the context of high-throughput 'omics' experiments in which the multiple testing is explicit[1]. We discussed the number of false discoveries that can be expected when a fixed *P* value is used to reject the null hypothesis. Here, to illustrate how *P* values can lead us astray, we reverse that process and instead ask: what is the smallest *P* value we can expect if the null hypothesis is true but we have done many tests, either explicitly or implicitly?

Consider a study in which 10 physiological variables are measured in 100 individuals to determine whether any of the variables are predictive of systolic blood pressure (SBP). Suppose that none of the variables are actually predictive in the population and that they are all independent. If we use simple linear regression[2] and focus on one of the variables as a predictor, a test of association will yield $P < 0.05$ in 5% of samples (**Fig. 1a**). However, if we test each of our predictors, there is now a 40% chance that we'll find $P < 0.05$ for at least one. How does this arise?

When we search for the most significant result, we do not have a fixed null hypothesis. It's entirely possible that a different predictor would be identified as most significant in the next repetition of the

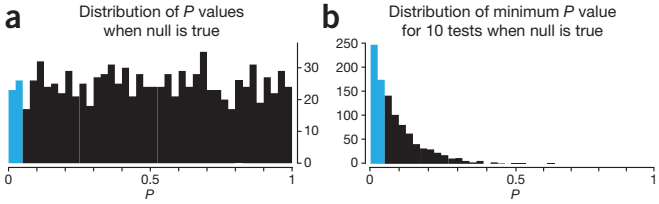

Figure 1 | *P* values are random variables. In assessing statistical significance, we rely on their distribution when the null hypothesis, H_0, is true. (**a**) Simulated *P* values from 1,000 statistical tests when H_0 is true. The distribution is uniform and, on average, 5% of $P < 0.05$ (blue). (**b**) The distribution of the minimum *P* value across 1,000 simulations of 10 tests when H_0 is true. Now, on average, 40% of $P < 0.05$ (blue). Note the difference in *y*-axis scale compared to **a**.

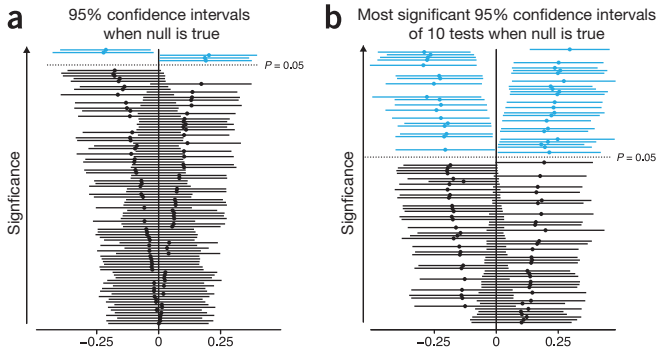

Figure 2 | Merely reporting 95% confidence intervals does not address selection bias. (**a**) 95% confidence intervals for 100 one-sample *t*-tests with samples of size $n = 100$, mean zero and s.d. = 1. Intervals are vertically sorted in increasing order of statistical significance. (**b**) 100 instances of the 95% confidence interval corresponding to the most significant result from a set of 10 one-sample *t*-tests of the kind performed in **a**.

experiment. In reporting the most significant *P* value, we are actually considering the distribution of the minimum of 10 random uniform distributions (**Fig. 1b**). This distribution is readily computed and has density $k(1-x)^{k-1}$ for k independent tests. Using $k = 10$, the probability of observing $P < 0.05$ is $1 - (1 - 0.05)^{10} = 0.40$ (**Fig. 1b**).

Reporting a statistically significant result as if this were the only test performed is an example of selection bias and leads to inflated claims of statistical significance. It's important to realize that it does not matter whether or not the tests were actually performed—any choice of results based on the outcome, rather than on prespecified hypotheses, will lead to selection bias.

We have previously seen that multiplicity adjustment is one way to fix selection bias for *P* values[1]. Using the adjusted *P* values from a family-wise error rate or false discovery rate guards against overinterpreting the *P* values when multiple testing has been done. However, it is less clear how to do this when an interesting effect has been detected after exploration of the data. For example, if we plotted SBP against each of our 10 predictors and felt that 1 predictor might have a quadratic relationship with SBP, should we adjust for 10 comparisons (the 10 plots), 20 comparisons (linear or quadratic effects) or more (to account for nonlinear relationships)? The more models we consider, the greater the danger of overfitting the data and producing false positives.

A common suggestion for supplementing *P* values is to report the confidence interval for the effect. Does this assist with selection bias? **Figure 2** shows the confidence intervals corresponding to the testing scenarios in **Figure 1**. When we perform 100 single hypothesis tests when the null is true, only 5% of the confidence intervals do not cover 0 (**Fig. 2a**). This picture looks very different if we consider only the most significant confidence interval from among 10 tests (**Fig. 2b**). On average, 40% of the confidence intervals do not cover 0, which should not be surprising, as we've already shown that this is the fraction of the underlying *P* values that are less than 0.05 (**Fig. 1b**). Selection bias is not addressed or corrected merely by reporting both the *P* value and the confidence interval.

Another common analysis in which *P* values can easily be misinterpreted is the selection of a prediction model for multiple regression or classification. To show how this can occur, we performed 1,000 simulations of our 10 physiological variables that were, as before, random and independent of each other and of SBP. We then applied forward selection to identify variables that

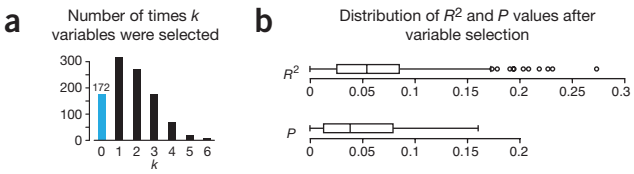

Figure 3 | Variable selection during model building greatly inflates statistical significance. (**a**) Number of times that 0 (the correct number) to 6 of predictors were selected as explanatory from 1,000 simulations. (**b**) Distribution of R^2 (top) and P values of the F-test (bottom) for the 828 cases from **a** in which the incorrect number ($k > 0$) of predictors was selected.

statistically predicted SBP. In this selection process, we start with no variables in the model and iteratively add the variables that provide the most statistically significant improvement, repeating this until no further variables add to the explanatory power of the model.

First, if we fit all 10 variables simultaneously and test at $P \leq 0.05$, we reject the null hypothesis of no association between the predictors and SBP only 5% of the time, as expected. However, using forward stepwise variable selection, we correctly identify 0 variables as predictive in only 172 out of 1,000 simulations (**Fig. 3a**). We reject the null hypothesis 82.8% of the time and observe compellingly low P values (**Fig. 3b**). Our results have a very high false discovery rate, even though there are 100 observations with only 10 predictors. Stepwise regression, which has even greater flexibility in choice of predictors, can boost the false discovery rate even more.

Although there is some recent work on inference after model selection, these are approximations that work only under limited conditions. The only universally acceptable method for validating a model and assessing its goodness of fit after model selection is use of an independent test sample[3].

So far, we have discussed only the simplest case, in which our set of putative predictors are independent. Dependence among the predictors complicates matters—if one of the predictors is statistically significant by chance, then other correlated predictors are also more likely to be statistically significant, which may appear to add weight to the significant results. For example, we might have several correlated metabolites as predictors. When one is selected, others may also be pulled into the model as predictors, creating a readily interpretable (yet wrong) biological explanation.

Another issue in the search for significance is overinterpretation of the relationship between statistical and biological significance. For example, suppose we have found a drug that can lower SBP on average by 10% of the population standard deviation, or about 2 mmHg. This is unlikely to be a medically relevant reduction. The power to detect such a small change is only 9% if the sample size is 10, but rises to 93.5% if the sample size is 1,000 (using a one-sided paired t-test). If the sample size is large enough, a study may correctly identify that there is a non-zero effect even if it is very small. However, to understand the biological relevance of the effect, we need an estimate of the effect size, such as a confidence interval. In the above example, computing a 95% confidence interval of 2 mm Hg ± 1 mm Hg would allow us to identify the lack of biological relevance. In contrast, merely stating that a significant reduction was found would obscure the fact that the result, although statistically significant, is not likely to be biologically relevant because the SBP reduction reported in the confidence interval is so small.

Recently, the American Statistical Association issued a statement on the appropriate use of P values and other inferential statistical methods, calling for caution in searching for significance[4]. The report warned against confounding relevance with statistical significance and effect sizes, inadequately exploring the data, not considering relevant covariates and overfitting—all practices that can lead to misuse and squandering of a data resource.

During statistical analysis, we must carefully distinguish between using data to confirm inferences and using data to generate hypotheses. In confirmatory use, P values and confidence intervals can be computed and interpreted as taught in basic statistics courses. In exploratory use, P values can be interpreted as measures of statistical significance only if appropriately adjusted for multiple testing or selection; confidence intervals also need to be adjusted for multiple testing. There are no simple and well-accepted means of doing this adjustment except in the case of explicit multiple testing. Our next column will discuss some suggested rules of thumb for interpreting and adjusting P values that combine frequentist methods with Bayesian paradigms.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

1. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 355–356 (2014).
2. Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 999–1000 (2015).
3. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 703–704 (2016).
4. Wasserstein, R.L. & Lazar, N.A. *Am. Stat.* **70**, 129–133 (2016).

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.