POINTS OF SIGNIFICANCE

Importance of being uncertain

Statistics does not tell us whether we are right. It tells us the chances of being wrong.

When an experiment is reproduced we almost never obtain exactly the same results. Instead, repeated measurements span a range of values because of biological variability and precision limits of measuring equipment. But if results are different each time, how do we determine whether a measurement is compatible with our hypothesis? In "the great tragedy of Science—the slaying of a beautiful hypothesis by an ugly fact"[1], how is 'ugliness' measured?

Statistics helps us answer this question. It gives us a way to quantitatively model the role of chance in our experiments and to represent data not as precise measurements but as estimates with error. It also tells us how error in input values propagates through calculations. The practical application of this theoretical framework is to associate uncertainty to the outcome of experiments and to assign confidence levels to statements that generalize beyond observations.

Although many fundamental concepts in statistics can be understood intuitively, as natural pattern-seekers we must recognize the limits of our intuition when thinking about chance and probability. The Monty Hall problem is a classic example of how the wrong answer can appear far too quickly and too credibly before our eyes. A contestant is given a choice of three doors, only one leading to a prize. After selecting a door (e.g., door 1), the host opens one of the other two doors that does not lead to a prize (e.g., door 2) and gives the contestant the option to switch their pick of doors (e.g., door 3). The vexing question is whether it is in the contestant's best interest to switch. The answer is yes, but you would be in good company if you thought otherwise. When a solution was published in *Parade* magazine, thousands of readers (many with PhDs) wrote in that the answer was wrong[2]. Comments varied from "You made a mistake, but look at the positive side. If all those PhDs were wrong, the country would be in some very serious trouble" to "I must admit I doubted you until my fifth grade math class proved you right"[2].

The Points of Significance column will help you move beyond an intuitive understanding of fundamental statistics relevant to your work. Its aim will be to address the observation that "approximately half the articles published in medical journals that use statistical methods use them incorrectly"[3]. Our presentation will be practical and cogent, with focus on foundational concepts, practical tips and common misconceptions[4]. A spreadsheet will often accompany each column to demonstrate the calculations (**Supplementary Table 1**). We will not exhaust you with mathematics.

Statistics can be broadly divided into two categories: descriptive and inferential. The first summarizes the main features of a data set with measures such as the mean and standard deviation (s.d.). The second generalizes from observed data to the world at large. Underpinning both are the concepts of sampling and estimation, which address the process of collecting data and quantifying the uncertainty in these generalizations.

To discuss sampling, we need to introduce the concept of a population, which is the set of entities about which we make inferences. The frequency histogram of all possible values of an experimental variable is called the population distribution (**Fig. 1a**). We are typically interested in inferring the mean (μ) and the s.d. (σ) of a population, two measures that characterize its location and spread (**Fig. 1b**). The mean is calculated as the arithmetic average of values and can be unduly influenced by extreme values. The median is a more robust measure

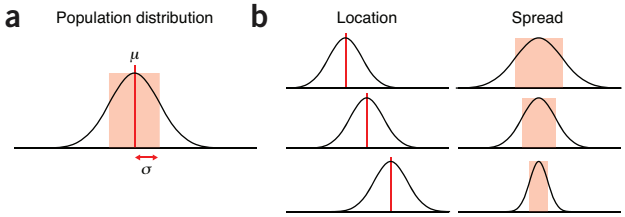

Figure 1 | The mean and s.d. are commonly used to characterize the location and spread of a distribution. When referring to a population, these measures are denoted by the symbols μ and σ.

of location and more suitable for distributions that are skewed or otherwise irregularly shaped. The s.d. is calculated based on the square of the distance of each value from the mean. It often appears as the variance (σ^2) because its properties are mathematically easier to formulate. The s.d. is not an intuitive measure, and rules of thumb help us in its interpretation. For example, for a normal distribution, 39%, 68%, 95% and 99.7% of values fall within $\pm 0.5\sigma$, $\pm 1\sigma$, $\pm 2\sigma$ and $\pm 3\sigma$. These cutoffs do not apply to populations that are not approximately normal, whose spread is easier to interpret using the interquartile range.

Fiscal and practical constraints limit our access to the population: we cannot directly measure its mean (μ) and s.d. (σ). The best we can do is estimate them using our collected data through the process of sampling (**Fig. 2**). Even if the population is limited to a narrow range of values, such as between 0 and 30 (**Fig. 2a**), the

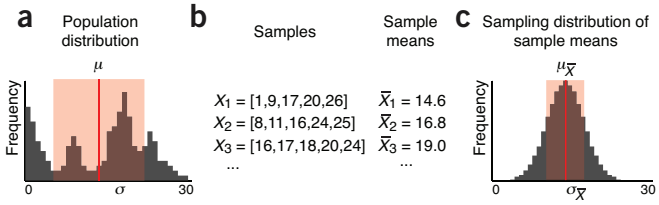

Figure 2 | Population parameters are estimated by sampling. (**a**) Frequency histogram of the values in a population. (**b**) Three representative samples taken from the population in **a**, with their sample means. (**c**) Frequency histogram of means of all possible samples of size $n = 5$ taken from the population in **a**.

random nature of sampling will impart uncertainty to our estimate of its shape. Samples are sets of data drawn from the population (**Fig. 2b**), characterized by the number of data points n, usually denoted by X and indexed by a numerical subscript (X_1). Larger samples approximate the population better.

To maintain validity, the sample must be representative of the population. One way of achieving this is with a simple random sample, where all values in the population have an equal chance of being selected at each stage of the sampling process. Representative does not mean that the sample is a miniature replica of the population. In general, a sample will not resemble the population unless n is very

large. When constructing a sample, it is not always obvious whether it is free from bias. For example, surveys sample only individuals who agreed to participate and do not capture information about those who refused. These two groups may be meaningfully different.

Samples are our windows to the population, and their statistics are used to estimate those of the population. The sample mean and s.d. are denoted by $\bar{X}$ and s. The distinction between sample and population variables is emphasized by the use of Roman letters for samples and Greek letters for population (s versus σ).

Sample parameters such as $\bar{X}$ have their own distribution, called the sampling distribution (**Fig. 2c**), which is constructed by considering all possible samples of a given size. Sample distribution parameters are marked with a subscript of the associated sample variable (for example, $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ are the mean and s.d. of the sample means of all samples). Just like the population, the sampling distribution is not directly measurable because we do not have access to all possible samples. However, it turns out to be an extremely useful concept in the process of estimating population statistics.

Notice that the distribution of sample means in **Figure 2c** looks quite different than the population in **Figure 2a**. In fact, it appears similar in shape to a normal distribution. Also notice that its spread, $\sigma_{\bar{X}}$, is quite a bit smaller than that of the population, σ. Despite these differences, the population and sampling distributions are intimately related. This relationship is captured by one of the most important and fundamental statements in statistics, the central limit theorem (CLT).

The CLT tells us that the distribution of sample means (**Fig. 2c**) will become increasingly close to a normal distribution as the sample size increases, regardless of the shape of the population distribution

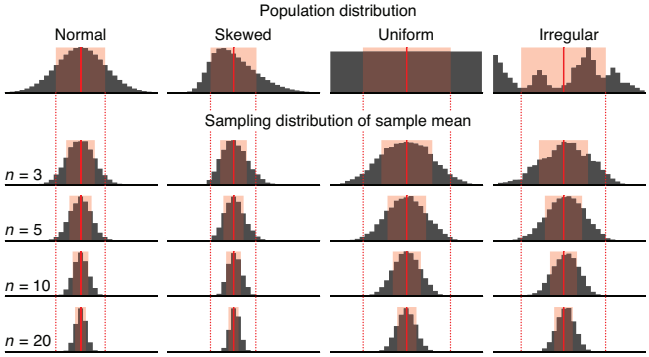

Figure 3 | The distribution of sample means from most distributions will be approximately normally distributed. Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in **Figure 1**.

(**Fig. 2a**) as long as the frequency of extreme values drops off quickly. The CLT also relates population and sample distribution parameters by $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. The terms in the second relationship are often confused: $\sigma_{\bar{X}}$ is the spread of sample means, and σ is the spread of the underlying population. As we increase n, $\sigma_{\bar{X}}$ will decrease (our samples will have more similar means) but σ will not change (sampling has no effect on the population). The measured spread of sample means is also known as the standard error of the mean (s.e.m., $\mathrm{SE}_{\bar{X}}$) and is used to estimate $\sigma_{\bar{X}}$.

A demonstration of the CLT for different population distributions (**Fig. 3**) qualitatively shows the increase in precision of our estimate of the population mean with increase in sample

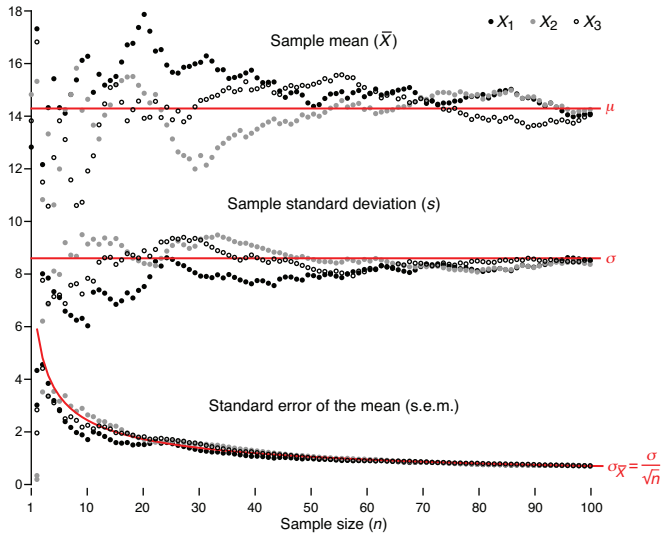

Figure 4 | The mean ($\bar{X}$), s.d. (s) and s.e.m. of three samples of increasing size drawn from the distribution in **Figure 2a**. As n is increased, $\bar{X}$ and s more closely approximate μ and σ. The s.e.m. ($s/\sqrt{n}$) is an estimate of $\sigma_{\bar{X}}$ and measures how well the sample mean approximates the population mean.

size. Notice that it is still possible for a sample mean to fall far from the population mean, especially for small n. For example, in ten iterations of drawing 10,000 samples of size $n = 3$ from the irregular distribution, the number of times the sample mean fell outside $\mu \pm \sigma$ (indicated by vertical dotted lines in **Fig. 3**) ranged from 7.6% to 8.6%. Thus, use caution when interpreting means of small samples.

Always keep in mind that your measurements are estimates, which you should not endow with "an aura of exactitude and finality"[5]. The omnipresence of variability will ensure that each sample will be different. Moreover, as a consequence of the $1/\sqrt{n}$ proportionality factor in the CLT, the precision increase of a sample's estimate of the population is much slower than the rate of data collection. In **Figure 4** we illustrate this variability and convergence for three samples drawn from the distribution in **Figure 2a**, as their size is progressively increased from $n = 1$ to $n = 100$. Be mindful of both effects and their role in diminishing the impact of additional measurements: to double your precision, you must collect four times more data.

Next month we will continue with the theme of estimation and discuss how uncertainty can be bounded with confidence intervals and visualized with error bars.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2613).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Huxley, T.H. in *Collected Essays* **8**, 229 (Macmillan, 1894).
2. vos Savant, M. Game show problem. http://marilynvossavant.com/game-show-problem (accessed 29 July 2013).
3. Glantz, S.A. *Circulation* **61**, 1–7 (1980).
4. Huck, S.W. *Statistical Misconceptions* (Routledge, 2009).
5. Ableson, R.P. *Statistics as Principled Argument* 27 (Psychology Press, 1995).

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

POINTS OF SIGNIFICANCE

Error bars

The meaning of error bars is often misinterpreted, as is the statistical significance of their overlap.

Last month in Points of Significance, we showed how samples are used to estimate population statistics. We emphasized that, because of chance, our estimates had an uncertainty. This month we focus on how uncertainty is represented in scientific publications and reveal several ways in which it is frequently misinterpreted.

The uncertainty in estimates is customarily represented using error bars. Although most researchers have seen and used error bars, misconceptions persist about how error bars relate to statistical significance. When asked to estimate the required separation between two points with error bars for a difference at significance $P = 0.05$, only 22% of respondents were within a factor of 2 (ref. 1). In light of the fact that error bars are meant to help us assess the significance of the difference between two values, this observation is disheartening and worrisome.

Here we illustrate error bar differences with examples based on a simplified situation in which the values are means of independent (unrelated) samples of the same size and drawn from normal populations with the same spread. We calculate the significance of the difference in the sample means using the two-sample *t*-test and report it as the familiar *P* value. Although reporting the exact *P* value is preferred, conventionally, significance is often assessed at a $P = 0.05$ threshold. We will discuss *P* values and the *t*-test in more detail in a subsequent column.

The importance of distinguishing the error bar type is illustrated in **Figure 1**, in which the three common types of error bars—standard deviation (s.d.), standard error of the mean (s.e.m.) and confidence interval (CI)—show the spread in values of two samples of size $n = 10$ together with the *P* value of the difference in sample means. In **Figure 1a**, we simulated the samples so that each error bar type has the same length, chosen to make them exactly abut. Although these three data pairs and their error bars are visually identical, each represents a different data scenario with a different *P* value. In **Figure 1b**, we fixed the *P* value to $P = 0.05$ and show the length of each type of bar for this level of significance. In this latter scenario, each of the three pairs of points represents the same pair of samples, but the bars have different lengths because they indicate different statistical properties of the same data. And because each bar is a different length, you are likely to interpret each one quite differently. In general, a gap between bars

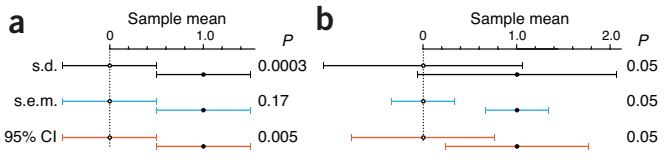

Figure 1 | Error bar width and interpretation of spacing depends on the error bar type. (**a**,**b**) Example graphs are based on sample means of 0 and 1 ($n = 10$). (**a**) When bars are scaled to the same size and abut, *P* values span a wide range. When s.e.m. bars touch, *P* is large ($P = 0.17$). (**b**) Bar size and relative position vary greatly at the conventional *P* value significance cutoff of 0.05, at which bars may overlap or have a gap.

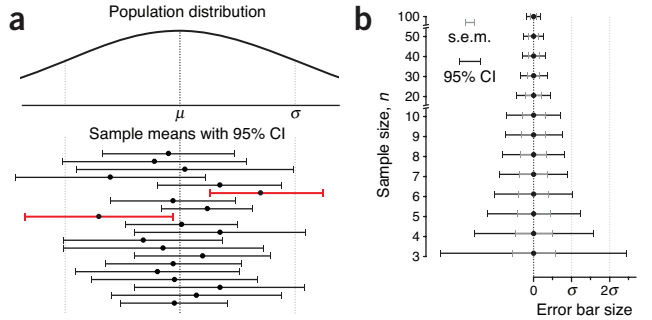

Figure 2 | The size and position of confidence intervals depend on the sample. On average, CI% of intervals are expected to span the mean—about 19 in 20 times for 95% CI. (**a**) Means and 95% CIs of 20 samples ($n = 10$) drawn from a normal population with mean μ and s.d. σ. By chance, two of the intervals (red) do not capture the mean. (**b**) Relationship between s.e.m. and 95% CI error bars with increasing *n*.

does not ensure significance, nor does overlap rule it out—it depends on the type of bar. Chances are you were surprised to learn this unintuitive result.

The first step in avoiding misinterpretation is to be clear about which measure of uncertainty is being represented by the error bar. In 2012, error bars appeared in *Nature Methods* in about two-thirds of the figure panels in which they could be expected (scatter and bar plots). The type of error bars was nearly evenly split between s.d. and s.e.m. bars (45% versus 49%, respectively). In 5% of cases the error bar type was not specified in the legend. Only one figure[2] used bars based on the 95% CI. CIs are a more intuitive measure of uncertainty and are popular in the medical literature.

Error bars based on s.d. inform us about the spread of the population and are therefore useful as predictors of the range of new samples. They can also be used to draw attention to very large or small population spreads. Because s.d. bars only indirectly support visual assessment of differences in values, if you use them, be ready to help your reader understand that the s.d. bars reflect the variation of the data and not the error in your measurement. What should a reader conclude from the very large and overlapping s.d. error bars for $P = 0.05$ in **Figure 1b**? That although the means differ, and this can be detected with a sufficiently large sample size, there is considerable overlap in the data from the two populations.

Unlike s.d. bars, error bars based on the s.e.m. reflect the uncertainty in the mean and its dependency on the sample size, *n* (s.e.m. = s.d./$\sqrt{n}$). Intuitively, s.e.m. bars shrink as we perform more measurements. Unfortunately, the commonly held view that "if the s.e.m. bars do not overlap, the difference between the values is statistically significant" is incorrect. For example, when $n = 10$ and s.e.m. bars just touch, $P = 0.17$ (**Fig. 1a**). Conversely, to reach $P = 0.05$, s.e.m. bars for these data need to be about 0.86 arm lengths apart (**Fig. 1b**). We cannot overstate the importance of recognizing the difference between s.d. and s.e.m.

The third type of error bar you are likely to encounter is that based on the CI. This is an interval estimate that indicates the reliability of a measurement[3]. When scaled to a specific confidence level (CI%)—the 95% CI being common—the bar captures the population mean CI% of the time (**Fig. 2a**). The size of the s.e.m. is compared to the 95% CI in **Figure 2b**. The two are related by the *t*-statistic, and in large samples the s.e.m. bar can be interpreted as a CI with a confidence level of 67%. The size of the CI depends on *n*; two useful approximations for the CI are 95% CI ≈ 4 × s.e.m ($n = 3$) and 95% CI ≈ 2 × s.e.m. ($n > 15$).

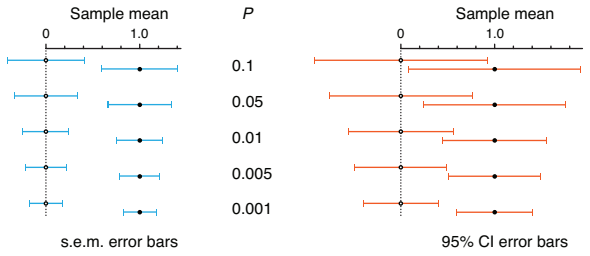

Figure 3 | Size and position of s.e.m. and 95% CI error bars for common *P* values. Examples are based on sample means of 0 and 1 ($n = 10$).

A common misconception about CIs is an expectation that a CI captures the mean of a second sample drawn from the same population with a CI% chance. Because CI position and size vary with each sample, this chance is actually lower.

This variety in bars can be overwhelming, and visually relating their relative position to a measure of significance is challenging. We provide a reference of error bar spacing for common *P* values in **Figure 3**. Notice that $P = 0.05$ is not reached until s.e.m. bars are separated by about 1 s.e.m, whereas 95% CI bars are more generous and can overlap by as much as 50% and still indicate a significant difference. If 95% CI bars just touch, the result is highly significant ($P = 0.005$). All the figures can be reproduced using the spreadsheet available in **Supplementary Table 1**, with which you can explore the relationship between error bar size, gap and *P* value.

Be wary of error bars for small sample sizes—they are not robust, as illustrated by the sharp decrease in size of CI bars in that regime (**Fig. 2b**). In these cases (e.g., $n = 3$), it is better to show individual data values. Furthermore, when dealing with samples that are related (e.g., paired, such as before and after treatment), other types of error bars are needed, which we will discuss in a future column.

It would seem, therefore, that none of the error bar types is intuitive. An alternative is to select a value of CI% for which the bars touch at a desired *P* value (e.g., 83% CI bars touch at $P = 0.05$). Unfortunately, owing to the weight of existing convention, all three types of bars will continue to be used. With our tips, we hope you'll be more confident in interpreting them.

Martin Krzywinski & Naomi Altman

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2659).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Belia, S.F., Fidler, F., Williams, J. & Cumming, G. *Psychol. Methods* **10**, 389–396 (2005).
2. Frøkjær-Jensen, C., Davis, M.W., Ailion, M. & Jorgensen, E.M. *Nat. Methods* **9**, 117–118 (2012).
3. Cumming, G., Fidler, F. & Vaux, D.L. *J. Cell. Biol.* **177**, 7–11 (2007).

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

POINTS OF SIGNIFICANCE

Significance, *P* values and *t*-tests

The *P* value reported by tests is a probabilistic significance, not a biological one.

Bench scientists often perform statistical tests to determine whether an observation is statistically significant. Many tests report the *P* value to measure the strength of the evidence that a result is not just a likely chance occurrence. To make informed judgments about the observations in a biological context, we must understand what the *P* value is telling us and how to interpret it. This month we will develop the concept of statistical significance and tests by introducing the one-sample *t*-test.

To help you understand how statistical testing works, consider the experimental scenario depicted in **Figure 1** of measuring protein expression level in a cell line with a western blot. Suppose we measure an expression value of $x = 12$ and have good reason to believe (for example, from past measurements) that the reference level is $\mu = 10$ (**Fig. 1a**). What can we say about whether this difference is due to random chance? Statistical testing can answer this question. But first, we need to mathematically frame our intuitive understanding of the biological and technical factors that disperse our measurements across a range of values.

We begin with the assumption that the random fluctuations in the experiment can be characterized by a distribution (**Fig. 1b**). This distribution is called the null distribution, and it embodies the null hypothesis (H_0) that our observation is a sample from the pool of all possible instances of measuring the reference. We can think of constructing this distribution by making a large number of independent measurements of a protein whose mean expression is known to equal the reference value. This distribution represents the probability of observing a given expression level for a protein that is being expressed at the reference level. The mean of this distribution, μ, is the reference expression, and its spread is determined by reproducibility factors inherent to our experiment. The purpose of a statistical test is to locate our observation on this distribution to identify the extent to which it is an outlier.

Statistics quantifies the outlier status of an observation by the probability of sampling another observation from the null distribution that is as far or farther away from μ. In our example, this corresponds to measuring an expression value further from the reference than x. This probability is the *P* value, which is the output of common statistical tests. It is calculated from the area under the distribution curve in the shaded regions (**Fig. 1c**). In some situations we may care only if x is too big (or too small), in which case we would compute the area of only the dark (light) shaded region of **Figure 1c**.

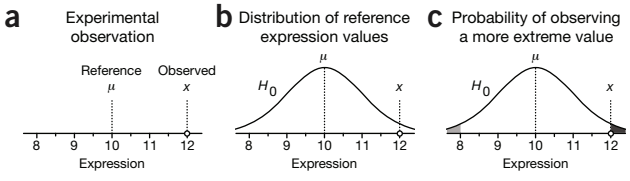

Figure 1 | The mechanism of statistical testing. (**a**–**c**) The significance of the difference between observed (x) and reference (μ) values (**a**) is calculated by assuming that observations are sampled from a distribution H_0 with mean μ (**b**). The statistical significance of the observation x is the probability of sampling a value from the distribution that is at least as far from the reference, given by the shaded areas under the distribution curve (**c**). This is the *P* value.

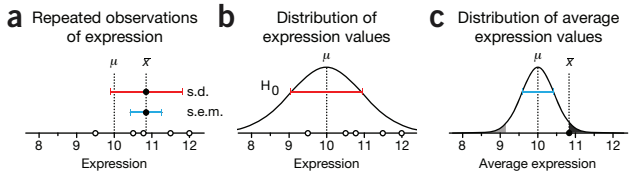

Figure 2 | Repeated independent observations are used to estimate the s.d. of the null distribution and derive a more robust *P* value. (**a**) A sample of $n = 5$ observations is taken and characterized by the mean $\bar{x}$, with error bars showing s.d. (s_x) and s.e.m. ($s_x/\sqrt{n}$). (**b**) The null distribution is assumed to be normal, and its s.d. is estimated by s_x. As in **Figure 1b**, the population mean is assumed to be μ. (**c**) The average expression is located on the sampling distribution of sample means, whose spread is estimated by the s.e.m. and whose mean is also μ. The *P* value of $\bar{x}$ is the shaded area under this curve.

Unfortunately, the *P* value is often misinterpreted as the probability that the null hypothesis (H_0) is true. This mistake is called the 'prosecutor's fallacy', which appeals to our intuition and was so coined because of its frequent use in courtroom arguments. In the process of calculating the *P* value, we assumed that H_0 was true and that x was drawn from H_0. Thus, a small *P* value (for example, $P = 0.05$) merely tells us that an improbable event has occurred in the context of this assumption. The degree of improbability is evidence against H_0 and supports the alternative hypothesis that the sample actually comes from a population whose mean is different than μ. Statistical significance suggests but does not imply biological significance.

At this point you may ask how we arrive at our assumptions about the null distribution in **Figure 1b**. After all, in order to calculate *P*, we need to know its precise shape. Because experimentally determining it is not practical, we need to make an informed guess. For the purposes of this column, we will assume that it is normal. We will discuss robustness of tests to this assumption of normality in another column. To complete our model of H_0, we still need to estimate its spread. To do this we return to the concept of sampling.

To estimate the spread of H_0, we repeat the measurement of our protein's expression. For example, we might make four additional independent measurements to make up a sample with $n = 5$ (**Fig. 2a**). We use the mean of expression values ($\bar{x} = 10.85$) as a measure of our protein's expression. Next, we make the key assumption that the s.d. of our sample ($s_x = 0.96$) is a suitable estimate of the s.d. of the null distribution (**Fig. 2b**). In other words, regardless of whether the sample mean is representative of the null distribution, we assume that its spread is. This assumption of equal variances is common, and we will be returning to it in future columns.

From our discussion about sampling[1], we know that given that H_0 is normal, the sampling distribution of means will also be normal, and we can use $s_x/\sqrt{n}$ to estimate its s.d. (**Fig. 2c**). We localize the mean expression on this distribution to calculate the *P* value, analogously to what was done with the single value in **Figure 1c**. To avoid the nuisance of dealing with a sampling distribution of means for each combination of population parameters, we can transform

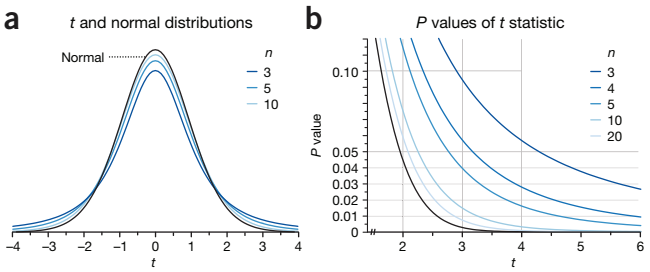

Figure 3 | The *t* and normal distributions. (**a**) The *t* distribution has higher tails that take into account that most samples will underestimate the variability in a population. The distribution is used to evaluate the significance of a *t* statistic derived from a sample of size *n* and is characterized by the degrees of freedom, d.f. = *n* – 1. (**b**) When *n* is small, *P* values derived from the *t* distribution vary greatly as *n* changes.

the mean $\bar{x}$ to a value determined by the difference of the sample and population means $D = \bar{x} - \mu$ divided by the s.e.m. ($s_x/\sqrt{n}$). This is called the test statistic.

It turns out, however, that the shape of this sampling distribution is close to, but not exactly, normal. The extent to which it departs from normal is known and given by the Student's *t* distribution (**Fig. 3a**), first described by William Gosset, who published under the pseudonym 'Student' (to avoid difficulties with his employer, Guinness) in his work on optimizing barley yields. The test statistic described above is compared to this distribution and is thus called the *t* statistic. The test illustrated in **Figure 2** is called the one-sample *t*-test.

This departure in distribution shape is due to the fact that for most samples, the sample variance, s_x^2, is an underestimate of the variance of the null distribution. The distribution of sample variances turns out to be skewed. The asymmetry is more evident for small *n*, where it is more likely that we observe a variance smaller than that of the population. The *t* distribution accounts for this underestimation by having higher tails than the normal distribution (**Fig. 3a**). As *n* grows, the *t* distribution looks very much like the normal, reflecting that the sample's variance becomes a more accurate estimate.

As a result, if we do not correct for this—if we use the normal distribution in the calculation depicted in **Figure 2c**—we will be using a distribution that is too narrow and will overestimate the significance of our finding. For example, using the $n = 5$ sample in **Figure 2b** for which $t = 1.98$, the *t* distribution gives us $P = 0.119$. Without the correction built into this distribution, we would underestimate *P* using the normal distribution as $P = 0.048$ (**Fig. 3b**). When *n* is large, the required correction is smaller: the same $t = 1.98$ for $n = 50$ gives $P = 0.054$, which is now much closer to the value obtained from the normal distribution.

The relationship between *t* and *P* is shown in **Figure 3b** and can be used to express *P* as a function of the quantities on which *t* depends (D, s_x, n). For example, if our sample in **Figure 2b** had a size of at least $n = 8$, the observed expression difference $D = 0.85$ would be significant at $P < 0.05$, assuming we still measured $s_x = 0.96$ ($t = 2.50$, $P = 0.041$). A more general type of calculation can identify conditions for which a test can reliably detect whether a sample comes from a distribution with a different mean. This speaks to the test's power, which we will discuss in the next column.

Another way of thinking about reaching significance is to consider what population means would yield $P < 0.05$. For our example, these would be $\mu < 9.66$ and $\mu > 12.04$ and define the range of standard expression values (9.66–12.04) that are compatible with our sample. In other words, if the null distribution had a mean within this interval, we would not be able to reject H_0 at $P = 0.05$ on the basis of our sample. This is the 95% confidence interval introduced last month, given by $\mu = \bar{x} \pm t^* \times$ s.e.m. (a rearranged form of the one-sample *t*-test equation), where t^* is the critical value of the *t* statistic for a given *n* and *P*. In our example, $n = 5$, $P = 0.05$ and $t^* = 2.78$. We encourage readers to explore these concepts for themselves using the interactive graphs in **Supplementary Table 1**.

The one-sample *t*-test is used to determine whether our samples could come from a distribution with a given mean (for example, to compare the sample mean to a putative fixed value μ) and for constructing confidence intervals for the mean. It appears in many contexts, such as measuring protein expression, the quantity of drug delivered by a medication or the weight of cereal in your cereal box. The concepts underlying this test are an important foundation for future columns in which we will discuss the comparisons across samples that are ubiquitous in the scientific literature.

Martin Krzywinski & Naomi Altman

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2698).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 809–810 (2013).

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

POINTS OF SIGNIFICANCE

Power and sample size

The ability to detect experimental effects is undermined in studies that lack power.

Statistical testing provides a paradigm for deciding whether the data are or are not typical of the values expected when the hypothesis is true. Because our objective is usually to detect a departure from the null hypothesis, it is useful to define an alternative hypothesis that expresses the distribution of observations when the null is false. The difference between the distributions captures the experimental effect, and the probability of detecting the effect is the statistical power.

Statistical power is critically relevant but often overlooked. When power is low, important effects may not be detected, and in experiments with many conditions and outcomes, such as 'omics' studies, a large percentage of the significant results may be wrong. **Figure 1** illustrates this by showing the proportion of inference outcomes in two sets of experiments. In the first set, we optimistically assume that hypotheses have been screened, and 50% have a chance for an effect (**Fig. 1a**). If they are tested at a power of 0.2, identified as the median in a recent review of neuroscience literature[1], then 80% of true positive results will be missed, and 20% of positive results will be wrong (positive predictive value, PPV = 0.80), assuming testing was done at the 5% level (**Fig. 1b**).

In experiments with multiple outcomes (e.g., gene expression studies), it is not unusual for fewer than 10% of the outcomes to have an a priori chance of an effect. If 90% of hypotheses are null (**Fig. 1a**), the situation at a 0.2 power level is bleak—over two-thirds of the positive results are wrong (PPV = 0.31; **Fig. 1b**). Even at the conventionally acceptable minimum power of 0.8, more than one-third of positive results are wrong (PPV = 0.64) because although we detect a greater fraction of the true effects (8 out of 10), we declare a larger absolute number of false positives (4.5 out of 90 nulls).

Fiscal constraints on experimental design, together with a commonplace lack of statistical rigor, contribute to many underpowered studies with spurious reports of both false positive and false negative effects. The consequences of low power are particularly dire in the search for high-impact

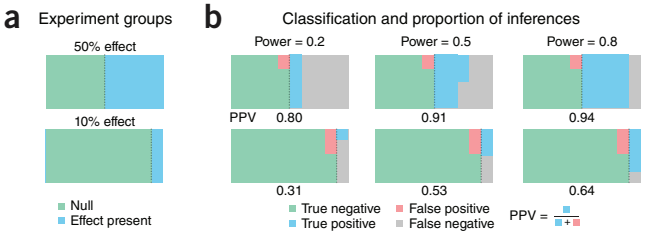

Figure 1 | When unlikely hypotheses are tested, most positive results of underpowered studies can be wrong. (**a**) Two sets of experiments in which 50% and 10% of hypotheses correspond to a real effect (blue), with the rest being null (green). (**b**) Proportion of each inference type within the null and effect groups encoded by areas of colored regions, assuming 5% of nulls are rejected as false positives. The fraction of positive results that are correct is the positive predictive value, PPV, which decreases with a lower effect chance.

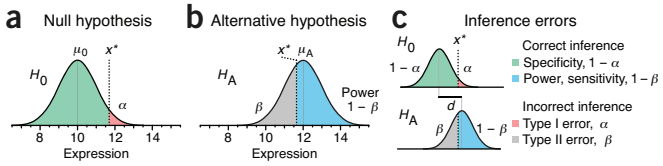

Figure 2 | Inference errors and statistical power. (**a**) Observations are assumed to be from the null distribution (H_0) with mean μ_0. We reject H_0 for values larger than x^* with an error rate α (red area). (**b**) The alternative hypothesis (H_A) is the competing scenario with a different mean μ_A. Values sampled from H_A smaller than x^* do not trigger rejection of H_0 and occur at a rate β. Power (sensitivity) is $1 - \beta$ (blue area). (**c**) Relationship of inference errors to x^*. The color key is same as in **Figure 1**.

results, when the researcher may be willing to pursue low-likelihood hypotheses for a groundbreaking discovery (**Fig. 1**). One analysis of the medical research literature found that only 36% of the experiments examined that had negative results could detect a 50% relative difference at least 80% of the time[2]. More recent reviews of the literature[1,3] also report that most studies are underpowered. Reduced power and an increased number of false negatives is particularly common in omics studies, which test at very small significance levels to reduce the large number of false positives.

Studies with inadequate power are a waste of research resources and arguably unethical when subjects are exposed to potentially harmful or inferior experimental conditions. Addressing this shortcoming is a priority—the Nature Publishing Group checklist for statistics and methods (http://www.nature.com/authors/policies/checklist.pdf) includes as the first question: "How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?" Here we discuss inference errors and power to help you answer this question. We'll focus on how the sensitivity and specificity of an experiment can be balanced (and kept high) and how increasing sample size can help achieve sufficient power.

Let's use the example from last month of measuring a protein's expression level x against an assumed reference level μ_0. We developed the idea of a null distribution, H_0, and said that x was statistically significantly larger than the reference if it exceeded some critical value x^* (**Fig. 2a**). If such a value is observed, we reject H_0 as the candidate model.

Because H_0 extends beyond x^*, it is possible to falsely reject H_0, with a probability of α (**Fig. 2a**). This is a type I error and corresponds to a false positive—that is, inferring an effect when there is actually none. In good experimental design, α is controlled and set low, traditionally at $\alpha = 0.05$, to maintain a high specificity $(1 - \alpha)$, which is the chance of a true negative—that is, correctly inferring that no effect exists.

Let's suppose that $x > x^*$, leading us to reject H_0. We may have found something interesting. If x is not drawn from H_0, what distribution does it come from? We can postulate an alternative hypothesis that characterizes an alternative distribution, H_A, for the observation. For example, if we expect expression values to be larger by 20%, H_A would have the same shape as H_0 but a mean of $\mu_A = 12$ instead of $\mu_0 = 10$ (**Fig. 2b**). Intuitively, if both of these distributions have similar means, we anticipate that it will be more difficult to reliably distinguish between them. This difference between the distributions is typically expressed by the difference in their means, in units of their s.d., σ. This measure, given by

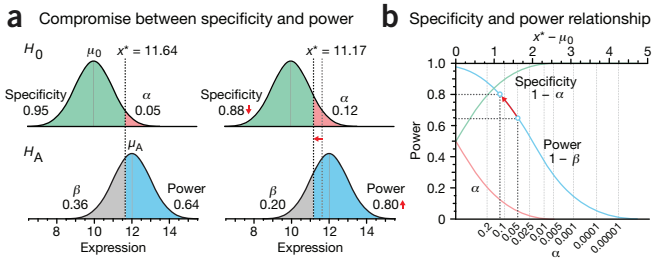

Figure 3 | Decreasing specificity increases power. H_0 and H_A are assumed normal with $\sigma = 1$. (**a**) Lowering specificity decreases the H_0 rejection cutoff x^*, capturing a greater fraction of H_A beyond x^*, and increases the power from 0.64 to 0.80. (**b**) The relationship between specificity and power as a function of x^*. The open circles correspond to the scenarios in **a**.

$d = (\mu_A - \mu_0)/\sigma$, is called the effect size. Sometimes effect size is combined with sample size as the noncentrality parameter, $d\sqrt{n}$.

In the context of these distributions, power (sensitivity) is defined as the chance of appropriately rejecting H_0 if the data are drawn from H_A. It is calculated from the area of H_A in the H_0 rejection region (**Fig. 2b**). Power is related by $1 - \beta$ to the type II error rate, β, which is the chance of a false negative (not rejecting H_0 when data are drawn from H_A).

A test should ideally be both sensitive (low false positive rate, α) and specific (low false negative rate, β). The α and β rates are inversely related: decreasing α increases β and reduces power (**Fig. 2c**). Typically, $\alpha < \beta$ because the consequences of false positive inference (in an extreme case, a retracted paper) are more serious than those of false negative inference (a missed opportunity to publish). But the balance between α and β depends on the objectives: if false positives are subject to another round of testing but false negatives are discarded, β should be kept low.

Let's return to our protein expression example and see how the magnitudes of these two errors are related. If we set $\alpha = 0.05$ and assume normal H_0 with $\sigma = 1$, then we reject H_0 when $x > 11.64$ (**Fig. 3a**). The fraction of H_A beyond this cutoff region is the power (0.64). We can increase power by decreasing sensitivity. Increasing α to 0.12 lowers the cutoff to $x > 11.17$, and power is now 0.80. This 25% increase in power has come at a cost: we are now more than twice as likely to make a false positive claim ($\alpha = 0.12$ vs. 0.05).

Figure 3b shows the relationship between α and power for our single expression measurement as a function of the position of

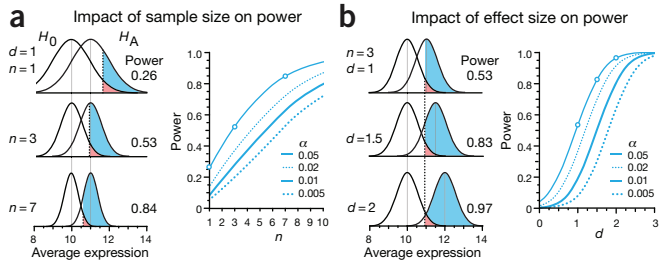

Figure 4 | Impact of sample (n) and effect size (d) on power. H_0 and H_A are assumed normal with $\sigma = 1$. (**a**) Increasing n decreases the spread of the distribution of sample averages in proportion to $1/\sqrt{n}$. Shown are scenarios at $n = 1$, 3 and 7 for $d = 1$ and $\alpha = 0.05$. Right, power as function of n at four different α values for $d = 1$. The circles correspond to the three scenarios. (**b**) Power increases with d, making it easier to detect larger effects. The distributions show effect sizes $d = 1$, 1.5 and 2 for $n = 3$ and $\alpha = 0.05$. Right, power as function of d at four different α values for $n = 3$.

H_0 rejection cutoff, x^*. The S-shape of the power curve reflects the rate of change of the area under H_A beyond x^*. The close coupling between α and power suggests that for $\mu_A = 12$ the highest power we can achieve for $\alpha \leq 0.05$ is 0.64. How can we improve our chance to detect increased expression from H_A (increase power) without compromising α (increasing false positives)?

If the distributions in **Figure 3a** were narrower, their overlap would be reduced, a greater fraction of H_A would lie beyond the x^* cutoff and power would be improved. We can't do much about σ, although we could attempt to lower it by reducing measurement error. A more direct way, however, is to take multiple samples. Now, instead of using single expression values, we formulate null and alternative distributions using the average expression value from a sample $\bar{x}$ that has spread $\sigma/\sqrt{n}$ (ref. 4).

Figure 4a shows the effect of sample size on power using distributions of the sample mean under H_0 and H_A. As n is increased, the H_0 rejection cutoff is decreased in proportion with the s.e.m., reducing the overlap between the distributions. Sample size substantially affects power in our example. If we average seven measurements ($n = 7$), we are able to detect a 10% increase in expression levels ($\mu_A = 11$, $d = 1$) 84% of the time with $\alpha = 0.05$. By varying n we can achieve a desired combination of power and α for a given effect size, d. For example, for $d = 1$, a sample size of $n = 22$ achieves a power of 0.99 for $\alpha = 0.01$.

Another way to increase power is to increase the size of the effect we want to reliably detect. We might be able to induce a larger effect size with a more extreme experimental treatment. As d is increased, so is power because the overlap between the two distributions is decreased (**Fig. 4b**). For example, for $\alpha = 0.05$ and $n = 3$, we can detect $\mu_A = 11$, 11.5 and 12 (10%, 15% and 20% relative increase; $d = 1$, 1.5 and 2) with a power of 0.53, 0.83 and 0.97, respectively. These calculations are idealized because the exact shapes of H_0 and H_A were assumed known. In practice, because we estimate population σ from the samples, power is decreased and we need a slightly larger sample size to achieve the desired power.

Balancing sample size, effect size and power is critical to good study design. We begin by setting the values of type I error (α) and power ($1 - \beta$) to be statistically adequate: traditionally 0.05 and 0.80, respectively. We then determine n on the basis of the smallest effect we wish to measure. If the required sample size is too large, we may need to reassess our objectives or more tightly control the experimental conditions to reduce the variance. Use the interactive graphs in **Supplementary Table 1** to explore power calculations.

When the power is low, only large effects can be detected, and negative results cannot be reliably interpreted. Ensuring that sample sizes are large enough to detect the effects of interest is an essential part of study design.

Martin Krzywinski & Naomi Altman

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2738).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Corrected after print 26 November 2013.

1. Button, K.S. *et al. Nat. Rev. Neurosci.* **14**, 365–376 (2013).
2. Moher, D., Dulberg, C.S. & Wells, G.A. *J. Am. Med. Assoc.* **272**, 122–124 (1994).
3. Breau, R.H., Carnat, T.A. & Gaboury, I. *J. Urol.* **176**, 263–266 (2006).
4. Krzywinski, M.I. & Altman, N. *Nat. Methods* **10**, 809–810 (2013).

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

Erratum: Power and sample size

Martin Krzywinski & Naomi Altman
Nat. Methods **10, 1139–1140 (2013); published online 26 November 2013; corrected after print 26 November 2013**

In the print version of this article initially published, the symbol μ_0 was represented incorrectly in the equation for effect size, $d = (\mu_A - \mu_0)/\sigma$. The error has been corrected in the HTML and PDF versions of the article.

POINTS OF SIGNIFICANCE

Visualizing samples with box plots

Use box plots to illustrate the spread and differences of samples.

Visualization methods enhance our understanding of sample data and help us make comparisons across samples. Box plots are a simple but powerful graphing tool that can be used in place of histograms to address both goals. Whereas histograms require a sample size of at least 30 to be useful, box plots require a sample size of only 5, provide more detail in the tails of the distribution and are more readily compared across three or more samples. Several enhancements to the basic box plot can render it even more informative.

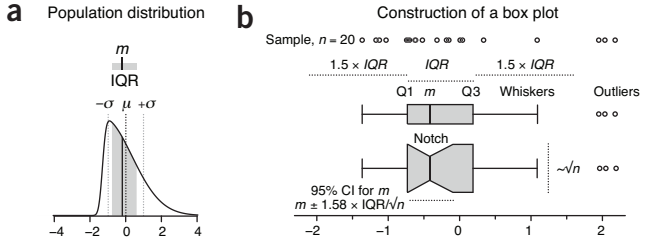

Figure 1 | The construction of a box plot. (**a**) The median (m = –0.19, solid vertical line) and interquartile range (IQR = 1.38, gray shading) are ideal for characterizing asymmetric or irregularly shaped distributions. A skewed normal distribution is shown with mean $\mu = 0$ (dark dotted line) and s.d. $\sigma = 1$ (light dotted lines). (**b**) Box plots for an $n = 20$ sample from **a**. The box bounds the IQR divided by the median, and Tukey-style whiskers extend to a maximum of 1.5 × IQR beyond the box. The box width may be scaled by $\sqrt{n}$, and a notch may be added approximating a 95% confidence interval (CI) for the median. Open circles are sample data points. Dotted lines indicate the lengths or widths of annotated features.

Box plots characterize a sample using the 25th, 50th and 75th percentiles—also known as the lower quartile (Q1), median (m or Q2) and upper quartile (Q3)—and the interquartile range (IQR = Q3 – Q1), which covers the central 50% of the data. Quartiles are insensitive to outliers and preserve information about the center and spread. Consequently, they are preferred over the mean and s.d. for population distributions that are asymmetric or irregularly shaped and for samples with extreme outliers. In such cases these measures may be difficult to intuitively interpret: the mean may be far from the bulk of the data, and conventional rules for interpreting the s.d. will likely not apply.

The core element that gives the box plot its name is a box whose length is the IQR and whose width is arbitrary (**Fig. 1**). A line inside the box shows the median, which is not necessarily central. The plot may be oriented vertically or horizontally—we use here (with one exception) horizontal boxes to maintain consistent orientation with corresponding sample distributions. Whiskers are conventionally extended to the most extreme data point that is no more than 1.5 × IQR from the edge of the box (Tukey style) or all the way to minimum and maximum of the data values (Spear style). The use

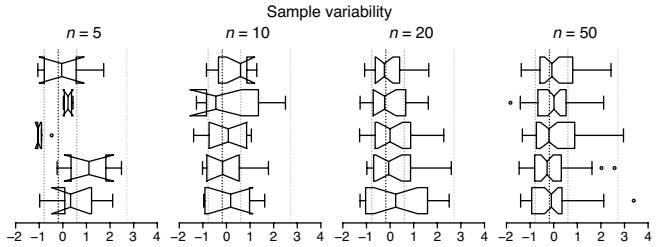

Figure 2 | Box plots reflect sample variability and should be avoided for very small samples ($n < 5$), with notches shown only when they appear within the IQR. Tukey-style box plots for five samples with sample size $n = 5$, 10, 20 and 50 drawn from the distribution in **Figure 1a** are shown; notch width is as in **Figure 1b**. Vertical dotted lines show Q1 (–0.78), median (–0.19), Q3 (0.60) and Q3 + 1.5 × IQR (2.67) values for the distribution.

of quartiles for box plots is a well-established convention: boxes or whiskers should never be used to show the mean, s.d. or s.e.m. As with the division of the box by the median, the whiskers are not necessarily symmetrical (**Fig. 1b**). The 1.5 multiplier corresponds to approximately $\pm 2.7\sigma$ (where σ is s.d.) and 99.3% coverage of the data for a normal distribution. Outliers beyond the whiskers may be individually plotted. Box plot construction requires a sample of at least $n = 5$ (preferably larger), although some software does not check for this. For $n < 5$ we recommend showing the individual data points.

Sample size differences can be assessed by scaling the box plot width in proportion to $\sqrt{n}$ (**Fig. 1b**), the factor by which the precision of the sample's estimate of population statistics improves as sample size is increased.

To assist in judging differences between sample medians, a notch (**Fig. 1b**) can be used to show the 95% confidence interval (CI) for the median, given by $m \pm 1.58 \times \text{IQR}/\sqrt{n}$ (ref. 1). This is an approximation based on the normal distribution and is accurate in large samples for other distributions. If you suspect the population distribution is not close to normal and your sample size is small, avoid interpreting the interval analytically in the way we have described for CI error bars[2]. In general, when notches do not overlap, the medians can be judged to differ significantly, but overlap does not rule out a significant difference. For small samples the notch may span a larger interval than the box (**Fig. 2**).

The exact position of box boundaries will be software dependent. First, there is no universally agreed-upon method to calculate quartile values, which may be based on simple averaging or linear interpolation. Second, some applications, such as R, use hinges instead of quartiles for box boundaries. The lower and upper hinges are the median of the

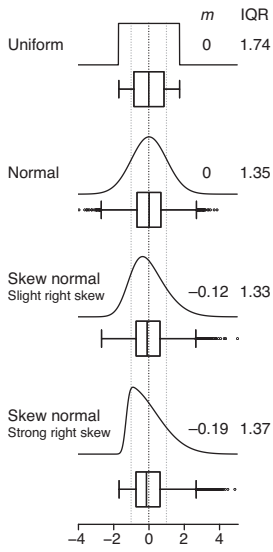

Figure 3 | Quartiles are more intuitive than the mean and s.d. for samples from skewed distributions. Four distributions with the same mean ($\mu = 0$, dark dotted line) and s.d. ($\sigma = 1$, light dotted lines) but significantly different medians (m) and IQRs are shown with corresponding Tukey-style box plots for $n = 10{,}000$ samples.

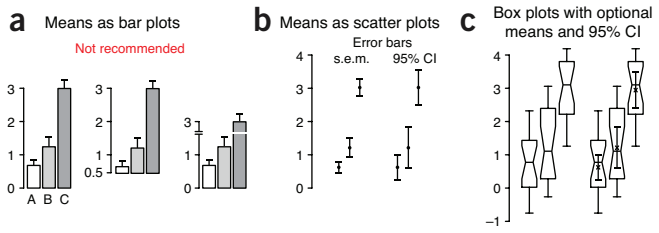

Figure 4 | Box plots are a more communicative way to show sample data. Data are shown for three $n = 20$ samples from normal distributions with s.d. $\sigma = 1$ and mean $\mu = 1$ (A,B) or 3 (C). (**a**) Showing sample mean and s.e.m. using bar plots is not recommended. Note how the change of baseline or cutting the *y* axis affects the comparative heights of the bars. (**b**) When sample size is sufficiently large ($n > 3$), scatter plots with s.e.m. or 95% confidence interval (CI) error bars are suitable for comparing central tendency. (**c**) Box plots may be combined with sample mean and 95% CI error bars to communicate more information about samples in roughly the same amount of space.

lower and upper half of the data, respectively, including the median if it is part of the data. Boxes based on hinges will be slightly different in some circumstances than those based on quartiles.

Aspects of the box plot such as width, whisker position, notch size and outlier display are subject to tuning; it is therefore important to clearly label how your box plot was constructed. Fewer than 20% of box plot figures in 2013 *Nature Methods* papers specified both sample size and whisker type in their legends—we encourage authors to be more specific.

The box plot is based on sample statistics, which are estimates of the corresponding population values. Sample variability will be reflected in the variation of all aspects of the box plot (**Fig. 2**). Modest sample sizes ($n = 5–10$) from the same population can yield very different box plots whose notches are likely to extend beyond the IQR. Even for large samples ($n = 50$), whisker positions can vary greatly. We recommend always indicating the sample size and avoiding notches unless they fall entirely within the IQR.

Although the mean and s.d. can always be calculated for any sample, they do not intuitively communicate the distribution of values (**Fig. 3**). Highly skewed distributions appear in box plot form with a markedly shorter whisker-and-box region and an absence of outliers on the side opposite the skew. Keep in mind that for small sample sizes, which do not necessarily represent the distribution well, these features may appear by chance.

We strongly discourage using bar plots with error bars (**Fig. 4a**), which are best used for counts or proportions[3]. These charts continue to be prevalent (we counted 100 figures that used them in 2013 *Nature Methods* papers, compared to only 20 that used box plots). They typically show only one arm of the error bar, making overlap comparisons difficult. More importantly, the bar itself encourages the perception that the mean is related to its height rather than the position of its top. As a result, the choice of baseline can interfere with assessing relative sizes of means and their error bars. The addition of axis breaks and log scaling makes visual comparisons even more difficult.

The traditional mean-and-error scatter plot with s.e.m. or 95% CI error bars (**Fig. 4b**) can be incorporated into box plots (**Fig. 4c**), thus combining details about the sample with an estimate of the population mean. For small samples, the s.e.m. bar may extend beyond the box. If data are normally distributed, >95% of s.e.m. bars will be within the IQR for $n \geq 14$. For 95% CI bars, the cutoff is $n \geq 28$.

Because they are based on statistics that do not require us to assume anything about the shape of the distribution, box plots robustly provide more information about samples than conventional error bars. We encourage their wider use and direct the reader to http://boxplot.tyerslab.com/ (ref. 4), a convenient online tool to create box plots that implements all the options described here.

Martin Krzywinski & Naomi Altman

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. McGill, R., Tukey, J.W & Larsen, W.A. *Am. Stat.* **32**, 12–16 (1978).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 921–922 (2013).
3. Streit, M. & Gehlenborg, N. *Nat. Methods* **11**, 117 (2014).
4. Spitzer, M. *et al. Nat. Methods* **11**, 121–122 (2014).

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

POINTS OF SIGNIFICANCE

Comparing samples—part I

Robustly comparing pairs of independent or related samples requires different approaches to the *t*-test.

Among the most common types of experiments are comparative studies that contrast outcomes under different conditions such as male versus female, placebo versus drug, or before versus after treatment. The analysis of these experiments calls for methods to quantitatively compare samples to judge whether differences in data support the existence of an effect in the populations they represent. This analysis is straightforward and robust when independent samples are compared; but researchers must often compare related samples, and this requires a different approach. We discuss both situations.

We'll begin with the simple scenario of comparing two conditions. This case is important to understand because it serves as a foundation for more complex designs with multiple simultaneous comparisons. For example, we may wish to contrast several treatments, track the evolution of an effect over time or consider combinations of treatments and subjects (such as different drugs on different genotypes).

We will want to assess the size of observed differences relative to the uncertainty in the samples. By uncertainty, we mean the spread as measured by the s.d., written as σ and s when referring to the population and sample estimate, respectively. It is more convenient to model uncertainty using variance, which is the square of the s.d. and denoted by Var() (or σ^2) and s^2 for the population and sample, respectively. Using this notation, the relationship between the uncertainty in the population of sample means and that of the population is $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$ for samples

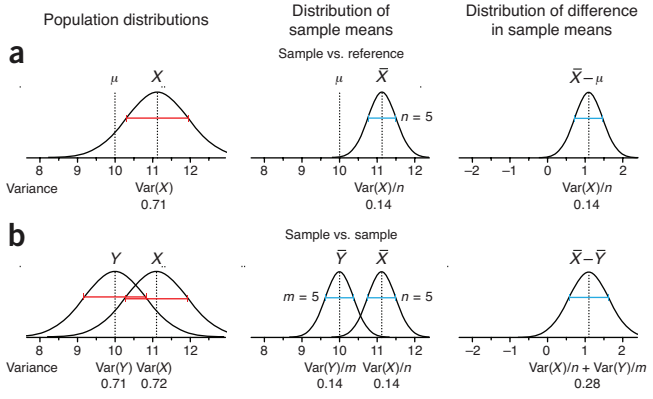

Figure 1 | The uncertainty in a sum or difference of random variables is the sum of the variables' individual uncertainties, as measured by the variance. Numerical values reflect sample estimates from **Figure 2**. Horizontal error bars show s.d., which is √Var. (**a**) Comparing a sample to a reference value involves only one measure of uncertainty: the variance of the sample's underlying population, Var(*X*). The variance of the sample mean is reduced in proportion to the sample size as Var(*X*)/*n*, which is also the uncertainty in the estimate of the difference between sample and reference. (**b**) When the reference is replaced by sample *Y* of size *m*, the variance of *Y* contributes to the uncertainty in the difference of means.

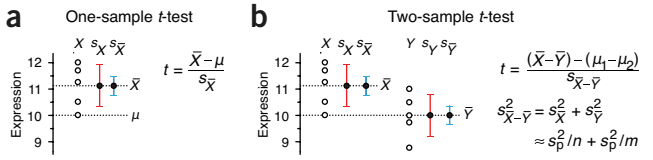

Figure 2 | In the two-sample test, both samples contribute to the uncertainty in the difference of means. (**a**) The difference between a sample ($n = 5$, $\bar{X} = 11.1$, $s_X = 0.84$) and a reference value ($\mu = 10$) can be assessed with a one-sample *t*-test. (**b**) When the reference value is itself a sample ($\bar{Y} = 10$, $s_Y = 0.85$), the two-sample version of the test is used, in which the *t*-statistic is based on a combined spread of *X* and *Y*, which is estimated using the pooled variance, s_p^2.

of size *n*. The equivalent statement for sample data is $s_{\bar{X}}^2 = s_X^2/n$, where $s_{\bar{X}}$ is the s.e.m. and s_X is the sample s.d.

Recall our example of the one-sample *t*-test in which the expression of a protein was compared to a reference value[1]. Our goal will be to extend this approach, in which only one quantity had uncertainty, to accommodate a comparison of two samples, in which both quantities now have uncertainty. **Figure 1a** encapsulates the relevant distributions for the one-sample scenario. We assumed that our sample *X* was drawn from a population, and we used the sample mean $\bar{X}$ to estimate the population mean. We defined the *t*-statistic (*t*) as the difference between the sample mean and the reference value, μ, in units of uncertainty in the mean, given by the s.e.m., and showed that *t* follows the Student's *t*-distribution[1] when the reference value is the mean of the population. We computed the probability that the difference between the sample and reference was due to the uncertainty in the sample mean. When this probability was less than a fixed type I error level, α, we concluded that the population mean differed from μ.

Let's now replace the reference with a sample *Y* of size *m* (**Fig. 1b**). Because the sample means are an estimate of the population means, the difference $\bar{X} - \bar{Y}$ serves as our estimate of the difference in the mean of the populations. Of course, populations can vary not only in their means, but for now we'll focus on this parameter. Just as in the one-sample case, we want to evaluate the difference in units of its uncertainty. The additional uncertainty introduced by replacing the reference with *Y* will need to be taken into account. To estimate the uncertainty in $\bar{X} - \bar{Y}$, we can turn to a useful result in probability theory.

For any two uncorrelated random quantities, *X* and *Y*, we have the following relationship: $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. In other words, the expected uncertainty in a difference of values is the sum of individual uncertainties. If we have reason to believe that the variances of the two populations are about the same, it is customary to use the average of sample variances as an estimate of both population variances. This is called the pooled variance, s_p^2. If the sample sizes are equal, it is computed by a simple average, $s_p^2 = (s_X^2 + s_Y^2)/2$. If not, it is an average weighted by $n - 1$ and $m - 1$, respectively. Using the pooled variance and applying the addition of variances rule to the variance of sample means gives $\mathrm{Var}(\bar{X} - \bar{Y}) = s_p^2/n + s_p^2/m$. The uncertainty in $\bar{X} - \bar{Y}$ is given by its s.d., which is the square root of this quantity.

To illustrate with a concrete example, we have reproduced the protein expression one-sample *t*-test example[1] in **Figure 2a** and contrast it to its two-sample equivalent in **Figure 2b**. We have adjusted sample values slightly to better illustrate the difference between these two tests. For the one-sample case, we find $t = 2.93$ and a corresponding *P* value of 0.04. At a type I error cutoff of $\alpha = 0.05$, we can conclude that the protein expression is significantly elevated relative to the refer-

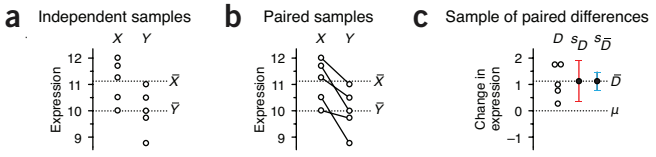

Figure 3 | The paired *t*-test is appropriate for matched-sample experiments. (**a**) When samples are independent, within-sample variability makes differences between sample means difficult to discern, and we cannot say that *X* and *Y* are different at $\alpha = 0.05$. (**b**) If *X* and *Y* represent paired measurements, such as before and after treatment, differences between value pairs can be tested, thereby removing within-sample variability from consideration. (**c**) In a paired test, differences between values are used to construct a new sample, to which the one-sample test is applied ($\bar{D} = 1.1$, $s_D = 0.65$).

ence. For the two-sample case, $t = 2.06$ and $P = 0.073$. Now, when the reference is replaced with a sample, the additional uncertainty in our difference estimate has resulted in a smaller *t* value that is no longer significant at the same α level. In the lookup between *t* and *P* for a two-sample test, we use d.f. $= n + m - 2$ degrees of freedom, which is the sum of d.f. values for each sample.

Our inability to reject the null hypothesis in the case of two samples is a direct result of the fact that the uncertainty in $\bar{X} - \bar{Y}$ is larger than in $\bar{X} - \mu$ (**Fig. 1b**) because now $\mathrm{Var}(\bar{Y})$ is a contributing factor. To reach significance, we would need to collect additional measurements. Assuming the sample means and s.d. do not change, one additional measurement would be sufficient—it would decrease $\mathrm{Var}(\bar{X} - \bar{Y})$ and increase the d.f. The latter has the effect of reducing the width of the *t*-distribution and lowering the *P* value for a given *t*.

This reduction in sensitivity is accompanied by a reduction in power[2]. The two-sample test has a lower power than the one-sample equivalent, for the same variance and number of observations per group. Our one-sample example with a sample size of 5 has a power of 52% for an expression change of 1.0. The corresponding power for the two-sample test with five observations per sample is 38%. If the sample variance remained constant, to reach the 52% power, the two-sample test would require larger samples ($n = m = 7$).

When assumptions are met, the two-sample *t*-test is the optimal procedure for comparing means. The robustness of the test is of interest because these assumptions may be violated in empirical data. One way departure from optimal performance is reported is by the difference between α—the type I error rate we think we are testing at—and the actual type I error rate, τ. If all assumptions are satisfied, $\alpha = \tau$, and our chance of committing a type I error is indeed equal to α. However, failing to satisfy assumptions can result in $\tau > \alpha$, causing us to commit a type I error more often than we think. In other words, our rate of false positives will be larger than planned for. Let's examine the assumptions of the *t*-test in the context of robustness.

First, the *t*-test assumes that samples are drawn from populations that are normal in shape. This assumption is the least burdensome. Systematic simulations of a wide range of practical distributions find that the type I error rate is stable within $0.03 < \tau < 0.06$ for $\alpha = 0.05$ for $n \geq 5$ (ref. 3).

Next, sample populations are required to have the same variance (**Fig. 1b**). Fortunately, the test is also extremely robust with respect to this requirement—more so than most people realize[3]. For example, when the sample sizes are equal, testing at $\alpha = 0.05$ (or $\alpha = 0.01$) gives $\tau < 0.06$ ($\tau < 0.015$) for $n \geq 15$, regardless of the difference in population variances. If these sample sizes are impractical, then we can fall back on the result that $\tau < 0.064$ when testing at $\alpha = 0.01$ regardless of *n* or difference in variance. When sample sizes are unequal, the impact of a variance difference is much larger, and τ can depart from α substantially. In these cases, the Welch's variant of the *t*-test is recommended, which uses actual sample variances, $s_X^2/n + s_Y^2/m$, in place of the pooled estimate. The test statistic is computed as usual, but the d.f. for the reference distribution depends on the estimated variances.

The final, and arguably most important, requirement is that the samples be uncorrelated. This requirement is often phrased in terms of independence, though the two terms have different technical definitions. What is important is that their Pearson correlation coefficient (ρ) be 0, or close to it. Correlation between samples can arise when data are obtained from matched samples or repeated measurements. If samples are positively correlated (larger values in first sample are associated with larger values in second sample), then the test performs more conservatively ($\tau < \alpha$)[4], whereas negative correlations increase the real type I error ($\tau > \alpha$). Even a small amount of correlation can make the test difficult to interpret—testing at $\alpha = 0.05$ gives $\tau < 0.03$ for $\rho > 0.1$ and $\tau > 0.08$ for $\rho < -0.1$.

If values can be paired across samples, such as measurements of the expression of the same set of proteins before and after experimental intervention, we can frame the analysis as a one-sample problem to increase the sensitivity of the test.

Consider the two samples in **Figure 3a**, which use the same values as in **Figure 2b**. If samples *X* and *Y* each measure different sets of proteins, then we have already seen that we cannot confidently conclude that the samples are different. This is because the spread within each sample is large relative to the differences in sample means. However, if *Y* measures the expression of the same proteins as *X*, but after some intervention, the situation is different (**Fig. 3b**), now we are concerned not with the spread of expression values within a sample but with the change of expression of a protein from one sample to another. By constructing a sample of differences in expression (*D*; **Fig. 3c**), we reduce the test to a one-sample *t*-test in which the sole source of uncertainty is the spread in differences. The spread within *X* and *Y* has been factored out of the analysis, making the test of expression difference more sensitive. For our example, we can conclude that expression has changed between *X* and *Y* at $P = 0.02$ ($t = 3.77$) by testing $\bar{D}$ against the null hypothesis that $\mu = 0$. This method is sometimes called the paired *t*-test.

We will continue our discussion of sample comparison next month, when we will discuss how to approach carrying out and reporting multiple comparisons. In the meantime, **Supplementary Table 1** can be used to interactively explore two-sample comparisons.

Martin Krzywinski & Naomi Altman

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2858).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1041–1042 (2013).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
3. Ramsey, P.H. *J. Educ. Stat.* **5**, 337–349 (1980).
4. Wiederman, W. & von Eye, A. *Psychol. Test Assess. Model.* **55**, 39–61 (2013).

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

POINTS OF SIGNIFICANCE

Comparing samples—part II

When a large number of tests are performed, *P* values must be interpreted differently.

It is surprising when your best friend wins the lottery but not when a random person in New York City wins. When we are monitoring a large number of experimental results, whether it is expression of all the features in an 'omics experiment or the outcomes of all the experiments done in the lifetime of a project, we expect to see rare outcomes that occur by chance. The use of *P* values, which assign a measure of rarity to a single experimental outcome, is misleading when many experiments are considered. Consequently, these values need to be adjusted and reinterpreted. The methods that achieve this are called multiple-testing corrections. We discuss the basic principles of this analysis and illustrate several approaches.

Recall the interpretation of the *P* value obtained from a single two-sample *t*-test: the probability that the test would produce a statistic at least as extreme, assuming that the null hypothesis is true. Significance is assigned when $P \leq \alpha$, where α is the type I error rate set to control false positives. Applying conventional $\alpha = 0.05$, we expect a 5% chance of making a false positive inference. This is the per-comparison error rate (PCER).

When we now perform *N* tests, this relatively small PCER can result in a large number of false positive inferences, αN. For example, if $N = 10{,}000$, as is common in analyses that examine large gene sets, we expect 500 genes to be incorrectly associated with an effect for $\alpha = 0.05$. If the effect chance is 10% and test power is 80%, we'll conclude that 1,250 genes show an effect, and we will be wrong 450 out of 1,250 times. In other words, roughly 1 out of 3 'discoveries' will be false. For cases in which the effect chance is even lower, our list of significant genes will be over-run with false positives: for a 1% effect chance, 6 out of 7 (495 of 575) discoveries are false. The role of multiple-testing correction methods is to mitigate these issues—a large number of false positives and large fraction of false discoveries—while ideally keeping power high.

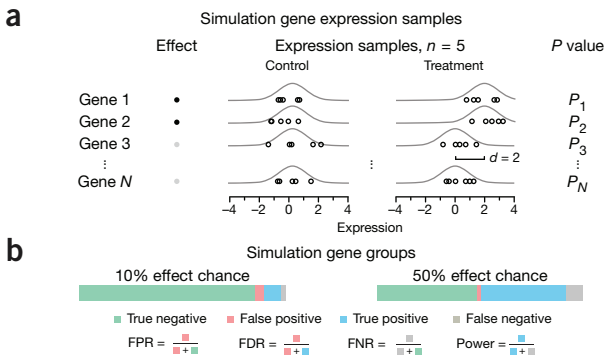

Figure 1 | The experimental design of our gene expression simulation. (**a**) A gene's expression was simulated by a control and treatment sample ($n = 5$ each) of normally distributed values ($\mu = 0$, $\sigma = 1$). For a fraction of genes, an effect size $d = 2$ (80% power) was simulated by setting $\mu = 2$. (**b**) Gene data sets were generated for 10% and 50% effect chances. *P* values were tested at $\alpha = 0.05$, and inferences were categorized as shown by the color scheme. For each data set and correction method, false positive rate (FPR), false detection rate (FDR) and power were calculated. FNR is the false negative rate.

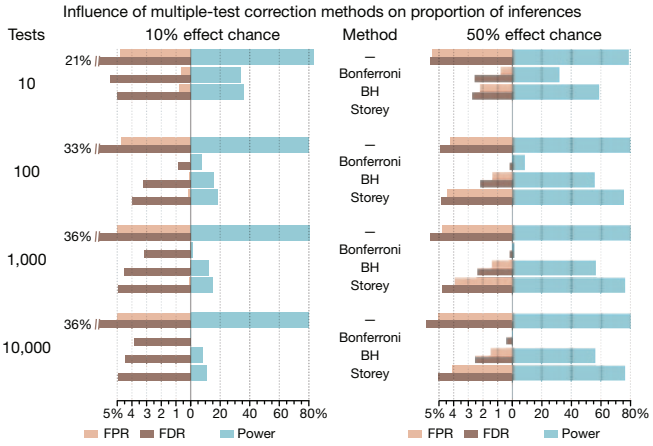

Figure 2 | Family-wise error rate (FWER) methods such as Bonferroni's negatively affect statistical power in comparisons across many tests. False discovery rate (FDR)-based methods such as Benjamini-Hochberg (BH) and Storey's are more sensitive. Bars show false positive rate (FPR), FDR and power for each combination of effect chance and *N* on the basis of inference counts using *P* values from the gene expression simulation (**Fig. 1**) adjusted with different methods (unadjusted (—), Bonferroni, BH and Storey). Storey's method did not provide consistent results for $N = 10$ because a larger number of tests is needed.

There are many adjustment methods; we will discuss common ones that adjust the *P* value. To illustrate their effect, we performed a simulation of a typical 'omics expression experiment in which *N* genes are tested for an effect between control and treatment (**Fig. 1a**). Some genes were simulated to have differential expression with an effect size $d = 2$, which corresponded to a test power of 80% at $\alpha = 0.05$. The *P* value for the difference in expression between control and treatment samples was computed with a two-sample *t*-test. We created data sets with $N = 10$, 100, 1,000 and 10,000 genes and an effect chance (percentage of genes having a nonzero effect) of 10% and 50% (**Fig. 1b**). We performed the simulation 100 times for each combination of *N* and effect chance to reduce the variability in the results to better illustrate trends, which are shown in **Figure 2**.

Figure 1b defines useful measures of the performance of the multiple-comparison experiment. Depending on the correction method, one or more of these measures are prioritized. The false positive rate (FPR) is the chance of inferring an effect when no effect is present. Without *P* value adjustment, we expect FPR to be close to α. The false discovery rate (FDR) is the fraction of positive inferences that are false. Technically, this term is reserved for the expected value of this fraction over all samples—for any given sample, the term false discovery percentage (FDP) is used, but either can be used if there is no ambiguity. Analogously to the FDR, the false nondiscovery rate (FNR) measures the error rate in terms of false negatives. Together the FDR and FNR are the multiple-test equivalents of type I and type II error levels. Finally, power is the fraction of real effects that are detected[1]. The performance of popular correction methods is illustrated using FPR, FDR and power in **Figure 2**.

The simplest correction method is Bonferroni's, which adjusts the *P* values by multiplying them by the number of tests, $P' = PN$, up to a maximum value of $P' = 1$. As a result, a *P* value may lose its significance in the context of multiple tests. For example, for $N = 10{,}000$ tests, an observed $P = 0.00001$ is adjusted $P' = 0.1$. The effect of this

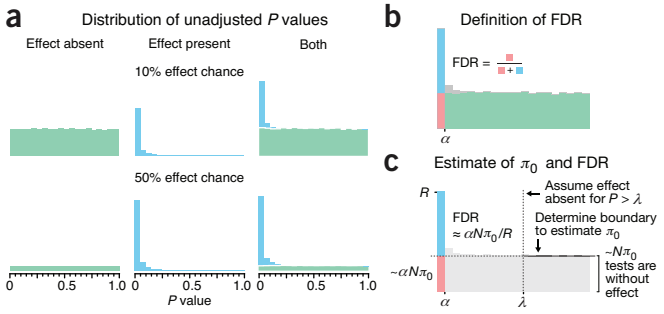

Figure 3 | The shape of the distribution of unadjusted *P* values can be used to infer the fraction of hypotheses that are null and the false discovery rate (FDR). (**a**) *P* values from null are expected to be distributed uniformly, whereas those for which the null is false will have more small values. Shown are distributions from the simulation for $N = 1{,}000$. (**b**) Inference types using color scheme of **Figure 1b** on the *P* value histogram. The FDR is the fraction of $P < \alpha$ that correspond to false positives. (**c**) Storey's method first estimates the fraction of comparisons for which the null is true, π_0, by counting the number of *P* values larger than a cutoff λ (such as 0.5) relative to $(1 - \lambda)N$ (such as $N/2$), the count expected when the distribution is uniform. If *R* discoveries are observed, about $\alpha N\pi_0$ are expected to be false positives, and FDR can be estimated by $\alpha N\pi_0/R$.

correction is to control the probability of committing even one type I error across all tests. The chance of this is called the family-wise error rate (FWER), and Bonferroni's correction ensures that FWER < α.

FWER methods such as Bonferroni's are extremely conservative and greatly reduce the test's power in order to control the number of false positives, particularly as the number of tests increases (**Fig. 2**). For $N = 10$ comparisons, our simulation shows a reduction in power for Bonferroni from 80% to ~33% for both 10% and 50% effect chance. These values drop to ~8% for $N = 100$, and by the time we are testing a large data set with $N = 10{,}000$, our power is ~0.2%. In other words, for a 10% effect chance, out of the 1,000 genes that have an effect, we expect to find only 2! Unless the cost of a false positive greatly outweighs the cost of a false negative, applying Bonferroni correction makes for an inefficient experiment. There are other FWER methods (such as Holm's and Hochberg's) that are designed to increase power by applying a less stringent adjustment to the *P* values. The benefits of these variants are realized when the number of comparisons is small (for example, <20) and the effect rate is high, but neither method will rescue the power of the test for a large number of comparisons.

In most situations, we are willing to accept a certain number of false positives, measured by FPR, as long as the ratio of false positives to true positives is low, measured by FDR. Methods that control FDR—such as Benjamini-Hochberg (BH), which scales *P* values in inverse proportion to their rank when ordered—provide better power characteristics than FWER methods. Our simulation shows that their power does not decrease as quickly as Bonferroni's with *N* for a small effect chance (for example, 10%) and actually increases with *N* when the effect chance is high (**Fig. 2**). At $N = 1{,}000$, whereas Bonferroni correction has a power of <2%, BH maintains 12% and 56% power at 10% and 50% effect rate while keeping FDR at 4.4% and 2.2%, respectively. Now, instead of identifying two genes at $N = 10{,}000$ and effect rate 10% with Bonferroni, we find 88 and are wrong only four times.

The final method shown in **Figure 2** is Storey's, which introduces two useful measures: π_0 and the *q* value. This approach is based on the observation that if the requirements of the *t*-test are met, the distribution of its *P* values for comparisons for which the null is true is expected to be uniform (by definition of the *P* value). In contrast, comparisons corresponding to an effect will have more *P* values close to 0 (**Fig. 3a**). In a real-world experiment we do not know which comparisons truly correspond to an effect, so all we see is the aggregate distribution, shown as the third histogram in **Figure 3a**. If the effect rate is low, most of our *P* values will come from cases in which the null is true, and the peak near 0 will be less pronounced than for a high effect chance. The peak will also be attenuated when the power of the test is low.

When we perform the comparison $P \leq \alpha$ on unadjusted *P* values, any values from the null will result in a false positive (**Fig. 3b**). This results in a very large FDR: for the unadjusted test, FDR = 36% for $N = 1{,}000$ and 10% effect chance. Storey's method adjusts *P* values with a rank scheme similar to that of BH but incorporates the estimate of the fraction of tests for which the null is true, π_0. Conceptually, this fraction corresponds to part of the distribution below the optimal boundary that splits it into uniform (*P* under true null) and skewed components (*P* under false null) (**Fig. 3b**). Two common estimates of π_0 are twice the average of all *P* values (Pound and Cheng's method) and $2/N$ times the number of *P* values greater than 0.5 (Storey's method). The latter is a specific case of a generalized estimate in which a different cutoff, λ, is chosen (**Fig. 3c**). Although π_0 is used in Storey's method in adjusting *P* values, it can be estimated and used independently. Storey's method performs very well, as long as there are enough comparisons to robustly estimate π_0. For all simulation scenarios, power is better than BH, and FDR is more tightly controlled at 5%. Use the interactive graphs in **Supplementary Table 1** to run the simulation and explore adjusted *P*-value distributions.

The consequences of misinterpreting the *P* value are repeatedly raised[2,3]. The appropriate measure to report in multiple-testing scenarios is the *q* value, which is the FDR equivalent of the *P* value. Adjusted *P* values obtained from methods such as BH and Storey's are actually *q* values. A test's *q* value is the minimum FDR at which the test would be declared significant. This FDR value is a collective measure calculated across all tests with FDR ≤ *q*. For example, if we consider a comparison with $q = 0.01$ significant, then we accept an FDR of at most 0.01 among the set of comparisons with $q \leq 0.01$. This FDR should not be confused with the probability that any given test is a false positive, which is given by the local FDR. The *q* value has a more direct meaning to laboratory activities than the *P* value because it relates the proportion of errors in the quantity of interest—the number of discoveries.

The choice of correction method depends on your tolerance for false positives and the number of comparisons. FDR methods are more sensitive, especially when there are many comparisons, whereas FWER methods sacrifice sensitivity to control false positives. When the assumptions of the *t*-test are not met, the distribution of *P* values may be unusual and these methods lose their applicability—we recommend always performing a quick visual check of the distribution of *P* values from your experiment before applying any of these methods.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2900).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
2. Nuzzo, R. *Nature* **506**, 150–152 (2014).
3. Anonymous. Trouble at the lab. *Economist* 26–30 (19 October 2013).

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

POINTS OF SIGNIFICANCE

Nonparametric tests

Nonparametric tests robustly compare skewed or ranked data.

We have seen that the *t*-test is robust with respect to assumptions about normality and equivariance[1] and thus is widely applicable. There is another class of methods—nonparametric tests—more suitable for data that come from skewed distributions or have a discrete or ordinal scale. Nonparametric tests such as the sign and Wilcoxon rank-sum tests relax distribution assumptions and are therefore easier to justify, but they come at the cost of lower sensitivity owing to less information inherent in their assumptions. For small samples, the performance of these tests is also constrained because their *P* values are only coarsely sampled and may have a large minimum. Both issues are mitigated by using larger samples.

These tests work analogously to their parametric counterparts: a test statistic and its distribution under the null are used to assign significance to observations. We compare in **Figure 1** the one-sample *t*-test[2] to a nonparametric equivalent, the sign test (though more sensitive and sophisticated variants exist), using a putative sample *X* whose source distribution we cannot readily identify (**Fig. 1a**). The null hypothesis of the sign test is that the sample median m_X is equal to the proposed median, $M = 0.4$. The test uses the number of sample values larger than *M* as its test statistic, *W*—under the null we expect to see as many values below the median as above, with the exact probability given by the binomial distribution (**Fig. 1c**). The median is a more useful descriptor than the mean for asymmetric and otherwise irregular distributions. The sign test makes no assumptions about the distribution—only that sample values be independent. If we propose that the population median is $M = 0.4$ and we observe *X*, we find $W = 5$ (**Fig. 1b**). The chance of observing a value of *W* under the null that is at least as extreme ($W \le 1$ or $W \ge 5$) is $P = 0.22$, using both tails of the binomial distribution (**Fig. 1c**). To limit the test to whether the median of *X* was biased towards values larger than *M*, we would consider only the area for $W \ge 5$ in the right tail to find $P = 0.11$.

The *P* value of 0.22 from the sign test is much higher than that from the *t*-test ($P = 0.04$), reflecting that the sign test is less sensitive. This is because it is not influenced by the actual distance between the sample values and *M*—it measures only 'how many' instead of 'how much'. Consequently, it needs larger sample sizes or more supporting evidence than the *t*-test. For the example of *X*, to obtain $P < 0.05$ we would need to have all values larger than *M* ($W = 6$). Its large *P* values and straightforward application makes the sign test a useful diagnostic. Take, for example, a hypothetical situation slightly different from that in **Figure 1**, where $P > 0.05$ is reported for the case where a treatment has lowered blood pressure in 6 out of 6 subjects. You may think this *P* seems implausibly large, and you'd be right because the equivalent scenario for the sign test ($W = 6$, $n = 6$) gives a two-tailed $P = 0.03$.

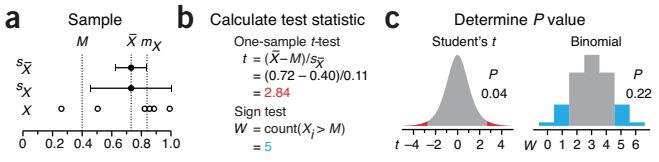

Figure 1 | A sample can be easily tested against a reference value using the sign test without any assumptions about the population distribution. (**a**) Sample *X* ($n = 6$) is tested against a reference $M = 0.4$. Sample mean $\bar{X}$ is shown with s.d. (s_X) and s.e.m. error bars ($s_{\bar{X}}$). m_x is sample median. (**b**) The *t*-statistic compares $\bar{X}$ to *M* in units of s.e.m. The sign test's *W* is the number of sample values larger than *M*. (**c**) Under the null, *t* follows Student's *t*-distribution with five degrees of freedom, whereas *W* is described by the binomial with 6 trials and $P = 0.5$. Two-tailed *P* values are shown.

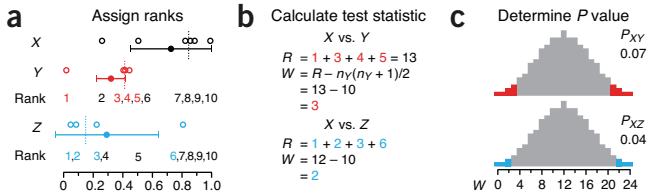

Figure 2 | Many nonparametric tests are based on ranks. (**a**) Sample comparisons of *X* vs. *Y* and *X* vs. *Z* start with ranking pooled values and identifying the ranks in the smaller-sized sample (e.g., 1, 3, 4, 5 for *Y*; 1, 2, 3, 6 for *Z*). Error bars show sample mean and s.d., and sample medians are shown by vertical dotted lines. (**b**) The Wilcoxon rank-sum test statistic *W* is the difference between the sum of ranks and the smallest possible observed sum. (**c**) For small sample sizes the exact distribution of *W* can be calculated. For samples of size (6, 4), there are only 210 different rank combinations corresponding to 25 distinct values of *W*.

To compare two samples, the Wilcoxon rank-sum test is widely used and is sometimes referred to as the Mann-Whitney or Mann-Whitney-Wilcoxon test. It tests whether the samples come from distributions with the same median. It doesn't assume normality, but as a test of equality of medians, it requires both samples to come from distributions with the same shape. The Wilcoxon test is one of many methods that reduce the dynamic range of values by converting them to their ranks in the list of ordered values pooled from both samples (**Fig. 2a**). The test statistic, *W*, is the degree to which the sum of ranks is larger than the lowest possible in the sample with the lower ranks (**Fig. 2b**). We expect that a sample from a population with a smaller median will be converted to a set of smaller ranks.

Because there is a finite number (210) of combinations of rank-ordering for *X* ($n_x = 6$) and *Y* ($n_Y = 4$), we can enumerate all outcomes of the test and explicitly construct the distribution of *W* (**Fig. 2c**) to assign a *P* value to *W*. The smallest value of $W = 0$ occurs when all values in one sample are smaller than those in the other. When they are all larger, the statistic reaches a maximum, $W = n_X n_Y = 24$. For *X* versus *Y*, $W = 3$, and there are 14 of 210 test outcomes with $W \le 3$ or $W \ge 21$. Thus, $P_{XY} = 14/210 = 0.067$. For *X* versus *Z*, $W = 2$, and $P_{XZ} = 8/210 = 0.038$. For cases in which both samples are larger than 10, *W* is approximately normal, and we can obtain the *P* value from a *z*-test of $(W - \mu_W)/\sigma_W$, where $\mu_W = n_1(n_1 + n_2 + 1)/2$ and $\sigma_W = \sqrt{(\mu_W n_2/6)}$.

The ability to enumerate all outcomes of the test statistic makes calculating the *P* value straightforward (**Figs. 1c** and **2c**), but there is an important consequence: there will be a minimum *P* value, P_{min}. Depending on the size of samples, P_{min} can be relatively large. For comparisons of samples of size $n_X = 6$ and $n_Y = 4$ (**Fig. 2a**), $P_{min} = 1/210 = 0.005$ for a one-tailed test, or 0.01 for a two-tailed test, corresponding to $W = 0$. Moreover, because there are only 25 distinct values of *W* (**Fig. 2c**), only two other two-tailed *P* values are <0.05: $P = 0.02$ ($W = 1$) and $P = 0.038$ ($W = 2$). The next-largest *P* value ($W = 3$) is $P = 0.07$. Because there is no *P* with value 0.05, the test cannot be set to reject the null at a type I rate of 5%. Even if we test at $\alpha = 0.05$, we will be rejecting the null at the

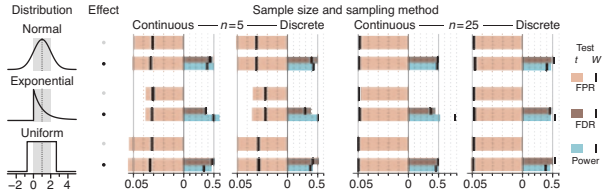

Figure 3 | The Wilcoxon rank-sum test can outperform the *t*-test in the presence of discrete sampling or skew. Data were sampled from three common analytical distributions with $\mu = 1$ (dotted lines) and $\sigma = 1$ (gray bars, $\mu \pm \sigma$). Discrete sampling was simulated by rounding values to the nearest integer. The FPR, FDR and power of Wilcoxon tests (black lines) and *t*-tests (colored bars) for 100,000 sample pairs for each combination of sample size ($n = 5$ and 25), effect chance (0 and 10%) and sampling method. In the absence of an effect, both sample values were drawn from a given distribution type with $\mu = 1$. With effect, the distribution for the second sample was shifted by d ($d = 1.4$ for $n = 5$; $d = 0.57$ for $n = 25$). The effect size was chosen to yield 50% power for the *t*-test in the normal noise scenario. Two-tailed P at $\alpha = 0.05$.

next lower P—for an effective type I error of 3.8%. We will see how this affects test performance for small samples further on. In fact, it may even be impossible to reach significance at $\alpha = 0.05$ because there is a limited number of ways in which small samples can vary in the context of ranks, and no outcome of the test happens less than 5% of the time. For example, samples of size 4 and 3 offer only 35 arrangements of ranks and a two-tailed $P_{min} = 2/35 = 0.057$. Contrast this to the *t*-test, which can produce any P value because the test statistic can take on an infinite number of values.

This has serious implications in multiple-testing scenarios discussed in the previous column[3]. Recall that when N tests are performed, multiple-testing corrections will scale the smallest P value to NP. In the same way as a test may never yield a significant result ($P_{min} > \alpha$), applying multiple-testing correction may also preclude it ($NP_{min} > \alpha$). For example, making $N = 6$ comparisons on samples such as X and Y shown in **Figure 2a** ($n_X = 6$, $n_Y = 4$) will never yield an adjusted P value lower than $\alpha = 0.05$ because $P_{min} = 0.01 > \alpha/N$. To achieve two-tailed significance at $\alpha = 0.05$ across $N = 10$, 100 or 1,000 tests, we require sample sizes that produce at least 400, 4,000 or 40,000 distinct rank combinations. This is achieved for sample pairs of size of (5, 6), (7, 8) and (9, 9), respectively.

The P values from the Wilcoxon test ($P_{XY} = 0.07$, $P_{XZ} = 0.04$) in **Figure 2a** appear to be in conflict with those obtained from the *t*-test ($P_{XY} = 0.04$, $P_{XZ} = 0.06$). The two methods tell us contradictory information—or do they? As mentioned, the Wilcoxon test concerns the median, whereas the *t*-test concerns the mean. For asymmetric distributions, these values can be quite different, and it is conceivable that the medians are the same but the means are different. The *t*-test does not identify the difference in means of X and Z as significant because the standard deviation, s_Z, is relatively large owing to the influence of the sample's largest value (0.81). Because the *t*-test reacts to any change in any sample value, the presence of outliers can easily influence its outcome when samples are small. For example, simply increasing the largest value in X (1.00) by 0.3 will increase s_X from 0.28 to 0.35 and result in a P_{XY} value that is no longer significant at $\alpha = 0.05$. This change does not alter the Wilcoxon P value because the rank scheme remains unaltered. This insensitivity to changes in the data—outliers and typical effects alike—reduces the sensitivity of rank methods.

The fact that the output of a rank test is driven by the probability that a value drawn from distribution A will be smaller (or larger) than one drawn from B without regard to their absolute difference has an interesting consequence: we cannot use this probability (pairwise preferences, in general) to impose an order on distributions. Consider a case of three equally prevalent diseases for which treatment A has cure times of 2, 2 and 5 days for the three diseases, and treatment B has 1, 4 and 4. Without treatment, each disease requires 3 days to cure—let's call this control C. Treatment A is better than C for the first two diseases but not the third, and treatment B is better only for the first. Can we determine which of the three options (A, B, C) is better? If we try to answer this using the probability of observing a shorter time to cure, we find $P(A < C) = 67\%$ and $P(C < B) = 67\%$ but also that $P(B < A) = 56\%$—a rock-paper-scissors scenario.

The question about which test to use does not have an unqualified answer—both have limitations. To illustrate how the *t*- and Wilcoxon tests might perform in a practical setting, we compared their false positive rate (FPR), false discovery rate (FDR) and power at $\alpha = 0.05$ for different sampling distributions and sample sizes ($n = 5$ and 25) in the presence and absence of an effect (**Fig. 3**). At $n = 5$, Wilcoxon FPR = $0.032 < \alpha$ because this is the largest P value it can produce smaller than α, not because the test inherently performs better. We can always reach this FPR with the *t*-test by setting $\alpha = 0.032$, where we'll find that it will still have slightly higher power than a Wilcoxon test that rejects at this rate. At $n = 5$, Wilcoxon performs better for discrete sampling—the power (0.43) is essentially the same as the *t*-test's (0.46), but the FDR is lower. When both tests are applied at $\alpha = 0.032$, Wilcoxon power (0.43) is slightly higher than *t*-test power (0.39). The differences between the tests for $n = 25$ diminishes because the number of arrangements of ranks is extremely large and the normal approximation to sample means is more accurate. However, one case stands out: in the presence of skew (e.g., exponential distribution), Wilcoxon power is much higher than that of the *t*-test, particularly for continuous sampling. This is because the majority of values are tightly spaced and ranks are more sensitive to small shifts. Skew affects *t*-test FPR and power in a complex way, depending on whether one- or two-tailed tests are performed and the direction of the skew relative to the direction of the population shift that is being studied[4].

Nonparametric methods represent a more cautious approach and remove the burden of assumptions about the distribution. They apply naturally to data that are already in the form of ranks or degree of preference, for which numerical differences cannot be interpreted. Their power is generally lower, especially in multiple-testing scenarios. However, when data are very skewed, rank methods reach higher power and are a better choice than the *t*-test.

Corrected after print 23 May 2014.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 215–216 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1041–1042 (2013).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 355–356 (2014).
4. Reineke, D.M, Baggett, J. & Elfessi, A. *J. Stat. Educ.* **11** (2003).

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

Corrigendum: Nonparametric tests

Martin Krzywinski & Naomi Altman
Nat. Methods **11, 467–468 (2014); published online 29 April 2014; corrected after print 23 May 2014**

In the version of this article initially published, the expression X ($n_X = 6$) was incorrectly written as X ($n_Y = 6$). The error has been corrected in the HTML and PDF versions of the article.

POINTS OF VIEW

Designing comparative experiments

Good experimental designs limit the impact of variability and reduce sample-size requirements.

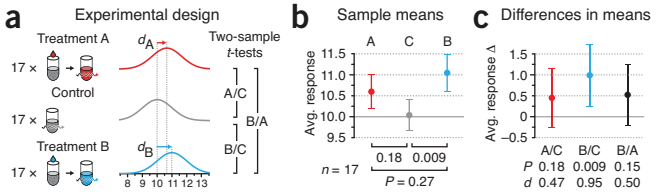

Figure 1 | Design and reporting of a single-factor experiment with three levels using a two-sample *t*-test. (**a**) Two treated samples (A and B) with $n = 17$ are compared to a control (C) with $n = 17$ and to each other using two-sample *t*-tests. (**b**) Simulated means and *P* values for samples in **a**. Values are drawn from normal populations with $\sigma = 1$ and mean response of 10 (C), 10.6 (A) and 11 (B). (**c**) The preferred reporting method of results shown in **b**, illustrating difference in means with CIs, *P* values and effect size, *d*. All error bars show 95% CI.

In a typical experiment, the effect of different conditions on a biological system is compared. Experimental design is used to identify data-collection schemes that achieve sensitivity and specificity requirements despite biological and technical variability, while keeping time and resource costs low. In the next series of columns we will use statistical concepts introduced so far and discuss design, analysis and reporting in common experimental scenarios.

In experimental design, the researcher-controlled independent variables whose effects are being studied (e.g., growth medium, drug and exposure to light) are called factors. A level is a subdivision of the factor and measures the type (if categorical) or amount (if continuous) of the factor. The goal of the design is to determine the effect and interplay of the factors on the response variable (e.g., cell size). An experiment that considers all combinations of *N* factors, each with n_i levels, is a factorial design of type $n_1 \times n_2 \times \ldots \times n_N$. For example, a 3 × 4 design has two factors with three and four levels each and examines all 12 combinations of factor levels. We will review statistical methods in the context of a simple experiment to introduce concepts that apply to more complex designs.

Suppose that we wish to measure the cellular response to two different treatments, A and B, measured by fluorescence of an aliquot of cells. This is a single factor (treatment) design with three levels (untreated, A and B). We will assume that the fluorescence (in arbitrary units) of an aliquot of untreated cells has a normal distribution with $\mu = 10$ and that real effect sizes of treatments A and B are $d_A = 0.6$ and $d_B = 1$ (A increases response by 6% to 10.6 and B by 10% to 11). To simulate variability owing to biological variation and measurement uncertainty (e.g., in the number of cells in an aliquot), we will use $\sigma = 1$ for the distributions. For all tests and calculations we use $\alpha = 0.05$.

We start by assigning samples of cell aliquots to each level (**Fig. 1a**). To improve the precision (and power) in measuring the mean of the response, more than one aliquot is needed[1]. One sample will be a control (considered a level) to establish the baseline response, and capture biological and technical variability. The other two samples will be used to measure response to each treatment. Before we can carry out the experiment, we need to decide on the sample size.

We can fall back to our discussion about power[1] to suggest *n*. How large an effect size (*d*) do we wish to detect and at what sensitivity? Arbitrarily small effects can be detected with large enough sample size, but this makes for a very expensive experiment. We will need to balance our decision based on what we consider to be a biologically meaningful response and the resources at our disposal. If we are satisfied with an 80% chance (the lowest power we should accept) of detecting a 10% change in response, which corresponds to the real effect of treatment B ($d_B = 1$), the two-sample *t*-test requires $n = 17$. At this *n* value, the power to detect $d_A = 0.6$ is 40%. Power calculations are easily computed with software; typically inputs are the difference in means ($\Delta\mu$), standard deviation estimate (σ), α and the number of tails (we recommend always using two-tailed calculations).

Based on the design in **Figure 1a**, we show the simulated samples means and their 95% confidence interval (CI) in **Figure 1b**. The 95% CI captures the mean of the population 95% of the time; we recommend using it to report precision. Our results show a significant difference between B and control (referred to as B/C, $P = 0.009$) but not for A/C ($P = 0.18$). Paradoxically, testing B/A does not return a significant outcome ($P = 0.15$). Whenever we perform more than one test we should adjust the *P* values[2]. As we only have three tests, the adjusted B/C *P* value is still significant, $P' = 3P = 0.028$. Although commonly used, the format used in **Figure 1b** is inappropriate for reporting our results: sample means, their uncertainty and *P* values alone do not present the full picture.

A more complete presentation of the results (**Fig. 1c**) combines the magnitude with uncertainty (as CI) in the difference in means. The effect size, *d*, defined as the difference in means in units of pooled standard deviation, expresses this combination of measurement and precision in a single value. Data in **Figure 1c** also explain better that the difference between a significant result (B/C, $P = 0.009$) and a nonsignificant result (A/C, $P = 0.18$) is not always significant (B/A, $P = 0.15$)[3]. Significance itself is a hard boundary at $P = \alpha$, and two arbitrarily close results may straddle it. Thus, neither significance itself nor differences in significance status should ever be used to conclude anything about the magnitude of the underlying differences, which may be very small and not biologically relevant.

CIs explicitly show how close we are to making a positive inference and help assess the benefit of collecting more data. For example, the CIs of A/C and B/C closely overlap, which suggests that at our sample size we cannot reliably distinguish between the response to A and B (**Fig. 1c**). Furthermore, given that the CI of A/C just barely crosses zero, it is possible that A has a real effect that our test failed to detect. More information about our ability to detect an effect can be obtained from a *post hoc* power analysis, which assumes that the observed effect is the same as the real effect (normally unknown), and uses the observed difference in means and pooled variance. For A/C, the difference in means is 0.48 and the pooled s.d. (s_p) = 1.03, which yields a *post hoc* power of 27%; we have little power to detect this difference. Other than increasing sample size, how could we improve our chances of detecting the effect of A?

Our ability to detect the effect of A is limited by variability in the difference between A and C, which has two random components. If

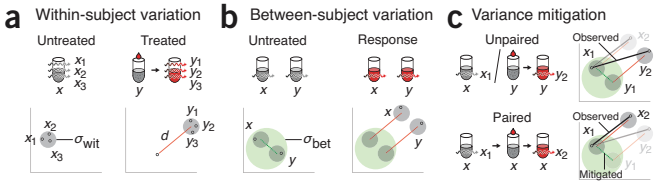

Figure 2 | Sources of variability, conceptualized as circles with measurements (x_i, y_i) from different aliquots (x,y) randomly sampled within them. (**a**) Limits of measurement and technical precision contribute to σ_{wit} (gray circle) observed when the same aliquot is measured more than once. This variability is assumed to be the same in the untreated and treated condition, with effect d on aliquot x and y. (**b**) Biological variation gives rise to σ_{bet} (green circle). (**c**) Paired design uses the same aliquot for both measurements, mitigating between-subject variation.

we measure the same aliquot twice, we expect variability owing to technical variation inherent in our laboratory equipment and variability of the sample over time (**Fig. 2a**). This is called within-subject variation, σ_{wit}. If we measure two different aliquots with the same factor level, we also expect biological variation, called between-subject variation, σ_{bet}, in addition to the technical variation (**Fig. 2b**). Typically there is more biological than technical variability ($\sigma_{bet} > \sigma_{wit}$). In an unpaired design, the use of different aliquots adds both σ_{wit} and σ_{bet} to the measured difference (**Fig. 2c**). In a paired design, which uses the paired t-test[4], the same aliquot is used and the impact of biological variation (σ_{bet}) is mitigated (**Fig. 2c**). If differences in aliquots (σ_{bet}) are appreciable, variance is markedly reduced (to within-subject variation) and the paired test has higher power.

The link between σ_{bet} and σ_{wit} can be illustrated by an experiment to evaluate a weight-loss diet in which a control group eats normally and a treatment group follows the diet. A comparison of the mean weight after a month is confounded by the initial weights of the subjects in each group. If instead we focus on the change in weight, we remove much of the subject variability owing to the initial weight.

If we write the total variance as $\sigma^2 = \sigma_{wit}^2 + \sigma_{bet}^2$, then the variance of the observed quantity in **Figure 2c** is $2\sigma^2$ for the unpaired design but $2\sigma^2(1 - \rho)$ for the paired design, where $\rho = \sigma_{bet}^2/\sigma^2$ is the correlation coefficient (intraclass correlation). The relative difference is captured by ρ of two measurements on the same aliquot, which must be included because the measurements are no longer independent. If we ignore ρ in our analysis, we will overestimate the variance and obtain overly conservative P values and CIs. In the case where there is no additional variation between aliquots, there is no benefit to using the same aliquot: measurements on the same aliquot are uncorrelated ($\rho = 0$) and variance of the paired test is the same as the variance of the unpaired. In contrast, if there is no variation in measurements on the same aliquot except for the treatment effect ($\sigma_{wit} = 0$), we have perfect correlation ($\rho = 1$). Now, the difference measurement derived from the same aliquot removes all the noise; in fact, a single pair of aliquots suffices for an exact inference. Practically, both sources of variation are present, and it is their relative size—reflected in ρ—that determines the benefit of using the paired t-test.

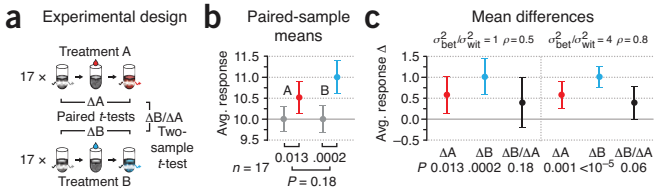

Figure 3 | Design and reporting for a paired, single-factor experiment. (**a**) The same $n = 17$ sample is used to measure the difference between treatment and background ($\Delta A = A_{after} - A_{before}$, $\Delta B = B_{after} - B_{before}$), analyzed with the paired t-test. Two-sample t-test is used to compare the difference between responses (ΔB versus ΔA). (**b**) Simulated sample means and P values for measurements and comparisons in **a**. (**c**) Mean difference, CIs and P values for two variance scenarios, $\sigma_{bet}^2/\sigma_{wit}^2$ of 1 and 4, corresponding to ρ of 0.5 and 0.8. Total variance was fixed: $\sigma_{bet}^2 + \sigma_{wit}^2 = 1$. All error bars show 95% CI.

We can see the improved sensitivity of the paired design (**Fig. 3a**) in decreased P values for the effects of A and B (**Fig. 3b** versus **Fig. 1b**). With the between-subject variance mitigated, we now detect an effect for A ($P = 0.013$) and an even lower P value for B ($P = 0.0002$) (**Fig. 3b**). Testing the difference between ΔA and ΔB requires the two-sample t-test because we are testing different aliquots, and this still does not produce a significant result ($P = 0.18$). When reporting paired-test results, sample means (**Fig. 3b**) should never be shown; instead, the mean difference and confidence interval should be shown (**Fig. 3c**). The reason for this comes from our discussion above: the benefit of pairing comes from reduced variance because $\rho > 0$, something that cannot be gleaned from **Figure 3b**. We illustrate this in **Figure 3c** with two different sample simulations with same sample mean and variance but different correlation, achieved by changing the relative amount of σ_{bet}^2 and σ_{wit}^2. When the component of biological variance is increased, ρ is increased from 0.5 to 0.8, total variance in difference in means drops and the test becomes more sensitive, reflected by the narrower CIs. We are now more certain that A has a real effect and have more reason to believe that the effects of A and B are different, evidenced by the lower P value for $\Delta B/\Delta A$ from the two-sample t-test (0.06 versus 0.18; **Fig. 3c**). As before, P values should be adjusted with multiple-test correction.

The paired design is a more efficient experiment. Fewer aliquots are needed: 34 instead of 51, although now 68 fluorescence measurements need to be taken instead of 51. If we assume $\sigma_{wit} = \sigma_{bet}$ ($\rho = 0.5$; **Fig. 3c**), we can expect the paired design to have a power of 97%. This power increase is highly contingent on the value of ρ. If σ_{wit} is appreciably larger than σ_{bet} (i.e., ρ is small), the power of the paired test can be lower than for the two-sample variant. This is because total variance remains relatively unchanged ($2\sigma^2(1 - \rho) \approx 2\sigma^2$) while the critical value of the test statistic can be markedly larger (particularly for small samples) because the number of degrees of freedom is now $n - 1$ instead of $2(n - 1)$. If the ratio of σ_{bet}^2 to σ_{wit}^2 is 1:4 ($\rho = 0.2$), the paired test power drops from 97% to 86%.

To analyze experimental designs that have more than two levels, or additional factors, a method called analysis of variance is used. This generalizes the t-test for comparing three or more levels while maintaining better power than comparing all sets of two levels. Experiments with two or more levels will be our next topic.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M.I. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
2. Krzywinski, M.I. & Altman, N. *Nat. Methods* **11**, 355–356 (2014).
3. Gelman, A. & Stern, H. *Am. Stat.* **60**, 328–331 (2006).
4. Krzywinski, M.I. & Altman, N. *Nat. Methods* **11**, 215–216 (2014).

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

POINTS OF SIGNIFICANCE

Analysis of variance and blocking

Good experimental designs mitigate experimental error and the impact of factors not under study.

Reproducible measurement of treatment effects requires studies that can reliably distinguish between systematic treatment effects and noise resulting from biological variation and measurement error. Estimation and testing of the effects of multiple treatments, usually including appropriate replication, can be done using analysis of variance (ANOVA). ANOVA is used to assess statistical significance of differences among observed treatment means based on whether their variance is larger than expected because of random variation; if so, systematic treatment effects are inferred. We introduce ANOVA with an experiment in which three treatments are compared and show how sensitivity can be increased by isolating biological variability through blocking.

Last month, we discussed a one-factor three-level experimental design that limited interference from biological variation by using the same sample to establish both baseline and treatment values[1]. There we used the *t*-test, which is not suitable when the number of factors or levels increases, in large part due to its loss of power as a result of multiple-testing correction. The two-sample *t*-test is a specific case of ANOVA, but the latter can achieve better power and naturally account for sources of error. ANOVA has the same requirements as the *t*-test: independent and randomly selected samples from approximately normal distributions with equal variance that is not under the influence of the treatments[2].

Here we continue with the three-treatment example[1] and analyze it with one-way (single-factor) ANOVA. As before, we simulated samples for $k = 3$ treatments each with $n = 6$ values (**Fig. 1a**). The ANOVA null hypothesis is that all samples are from the same distribution and have equal means. Under this null, between-group variation of sample means and within-group variation of sample values are predictably related. Their ratio can be used as a test statistic, F, which will be larger than expected in the presence of treatment effects. Although it appears that we are testing equality of variances, we are actually testing whether all the treatment effects are zero.

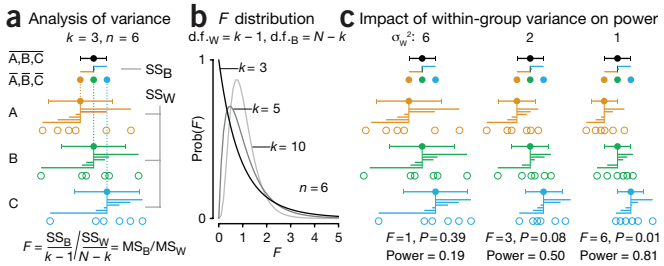

Figure 1 | ANOVA is used to determine significance using the ratio of variance estimates from sample means and sample values. (**a**) Between- and within-group variance is calculated from SS_B, the between treatment sum of squares, and SS_W, the within treatment sum of squares.. Deviations are shown as horizontal lines extending from grand and sample means. The test statistic, F, is the ratio mean squares MS_B and MS_W, which are SS_B and SS_W weighted by d.f. (**b**) Distribution of F, which becomes approximately normal as k and N increase, shown for $k = 3$, 5 and 10 samples each of size $n = 6$. $N = kn$ is the total number of sample values. (**c**) ANOVA analysis of sample sets with decreasing within-group variance ($\sigma_w^2 = 6,2,1$). $MS_B = 6$ in each case. Error bars, s.d.

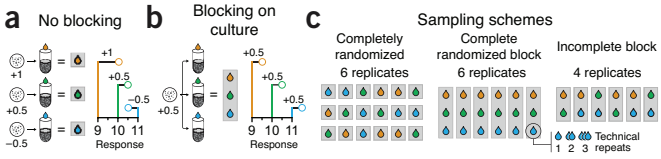

Figure 2 | Blocking improves sensitivity by isolating variation in samples that is independent from treatment effects. (**a**) Measurements from treatment aliquots derived from different cell cultures are differentially offset (e.g., 1, 0.5, −0.5) because of differences in cultures. (**b**) When aliquots are derived from the same culture, measurements are uniformly offset (e.g., 0.5). (**c**) Incorporating blocking in data collection schemes. Repeats within blocks are considered technical replicates. In an incomplete block design, a block cannot accommodate all treatments.

ANOVA calculations are summarized in an ANOVA table, which we provide for **Figures 1**, **3** and **4** (**Supplementary Tables 1–3**) along with an interactive spreadsheet (**Supplementary Table 4**). The sums of squares (SS) column shows sums of squared deviations of various quantities from their means. This sum is performed over each data point—each sample mean deviation (**Fig. 1a**) contributes to SS_B six times. The degrees of freedom (d.f.) column shows the number of independent deviations in the sums of squares; the deviations are not all independent because deviations of a quantity from its own mean must sum to zero. The mean square (MS) is SS/d.f. The F statistic, $F = MS_B/MS_W$, is used to test for systematic differences among treatment means. Under the null, F is distributed according to the F distribution for $k - 1$ and $N - k$ d.f. (**Fig. 1b**). When we reject the null, we conclude that not all sample means are the same; additional tests are required to identify which treatment means are different. The ratio $\eta^2 = SS_B/(SS_B + SS_W)$ is the coefficient of variation (also called R^2) and measures the fraction of the total variation resulting from differences among treatment means.

We previously introduced the idea that variance can be partitioned: within-group variance, σ_{wit}^2, was interpreted as experimental error and between-group variance, σ_{bet}^2, as biological variation[1]. In one-way ANOVA, the relevant quantities are MS_W and MS_B. MS_W corresponds to variance in the sample after other sources of variation have been accounted for and represents experimental error (σ_{wit}^2). If some sources of error are not accounted for (e.g., biological variation), MS_W will be inflated. MS_B is another estimate for MS_W, additionally inflated by average squared deviation of treatment means from the

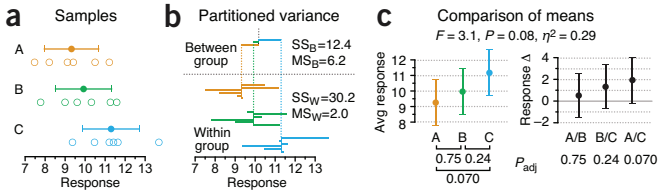

Figure 3 | Application of one-factor ANOVA to comparison of three samples. (**a**) Three samples drawn from normal distributions with $\sigma_{wit}^2 = 2$ and treatment means $\mu_A = 9$, $\mu_B = 10$ and $\mu_C = 11$. (**b**) Depiction of deviations with corresponding SS and MS values. (**c**) Sample means and their differences. P values for paired sample comparison are adjusted for multiple comparison using Tukey's method. Error bars, 95% CI.

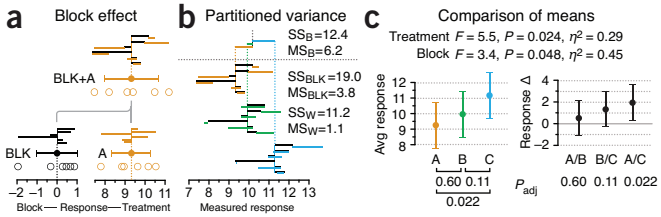

Figure 4 | Including blocking isolates biological variation from the estimate of within-group variance and improves power. (**a**) Blocking is simulated by augmenting each sample ($\sigma_{wit}^2 = 1$) with a fixed random component ($\mu_{blk} = 0$, $\sigma_{blk}^2 = 1$). (**b**) Variance is partitioned to treatment, block (black lines) and within-group. (**c**) Summary statistics for treatment and block effects in the same format as **Figure 3c**. In the presence of a sufficiently large blocking effect, MS_W is lowered and treatment test statistic $F = MS_B/MS_W$ is increased. Smaller error bars on sample mean differences reflect reduced MS_W.

grand mean, θ^2, times sample size if the null hypothesis is not true ($\sigma_{wit}^2 + n\theta^2$). Thus, the noisier the data (σ_{wit}^2), the more difficult it is to tease out σ_{treat}^2 and detect real effects, just like in the *t*-test, the power of which could be increased by decreasing sample variance[2]. To demonstrate this, we simulated three different sample sets in **Figure 1c** with $MS_B = 6$ and different MS_W values, for a scenario with fixed treatment effects ($\sigma_{treat}^2 = 1$), but progressively reduced experimental error ($\sigma_{wit}^2 = 6,2,1$). As noise within samples drops, a larger fraction variation is allocated to MS_B, and the power of the test improves. This suggests that it is beneficial to decrease MS_W. We can do this through a process called blocking to identify and isolate likely sources of sample variability.

Suppose that our samples in **Figure 1a** were generated by measuring the response to treatment of an aliquot of cells—a fixed volume of cells from a culture (**Fig. 2a**). Assume that it is not possible to derive all required aliquots from a single culture or that it is necessary to use multiple cultures to ensure that the results generalize. It is likely that aliquots from different cultures will respond differently owing to variation in cell concentration, growth rates, medium composition, among others. These so-called nuisance variables confound the real treatment effects: the baseline for each measurement unpredictably varies (**Fig. 2a**). We can mitigate this by using the same cell culture to create three aliquots, one for each treatment, to propagate these differences equally among measurements (**Fig. 2b**). Although measurements between cultures still would be shifted, the relative differences between treatments within the same culture remain the same. This process is called blocking, and its purpose is to remove as much variability as possible to make differences between treatments more evident. For example, the paired *t*-test implements blocking by using the same subject or biological sample.

Without blocking, cultures, aliquots and treatments are not matched—a completely randomized design (**Fig. 2c**)—which makes differences in cultures impossible to isolate. For blocking, we systematically assign treatments to cultures, such as in a randomized complete block design, in which each culture provides a replicate of each treatment (**Fig. 2c**). Each block is subjected to each of the treatments exactly once, and we can optionally collect technical repeats (repeating data collection from the measurement apparatus or multiple aliquots from the same culture) to minimize the impact of fluctuations in our measuring apparatus; these values would be averaged. In the case where a block cannot support all treatments (e.g., a culture yields only two aliquots), we would use combinations of treatment pairs with the requirement that each pair is measured equally often—a balanced incomplete block design. Let us look at how blocking can increase ANOVA sensitivity using the scenario from **Figure 1**.

We will start with three samples ($n = 6$) (**Fig. 3a**) that measure the effects of treatments A, B and C on aliquots of cells in a completely randomized scheme. We simulated the samples with $\sigma_{wit}^2 = 2$ to represent experimental error. Using ANOVA, we partition the variation (**Fig. 3b**) and find the mean squares for the components ($MS_B = 6.2$, $MS_W = 2.0$; **Supplementary Table 2**). MS_W reflects the value $\sigma_{wit}^2 = 2$ in the sample simulation, and it turns out that this variance is too high to yield a significant *F*; we find $F = 3.1$ ($P = 0.08$; **Fig. 3c**). Because we did not find a significant difference using ANOVA, we do not expect to obtain significant *P* values from two-sample *t*-tests applied pairwise to the samples. Indeed, when adjusted for multiple-test correction these P_{adj} values are all greater than 0.05 (**Fig. 3c**).

To illustrate blocking, we simulate samples to have the same values as in **Figure 3a** but with half of the variance due to differences in cultures. These differences in cultures (block effect) are simulated as normal with mean $\mu_{blk} = 0$ and variance $\sigma_{blk}^2 = 1$ (**Fig. 4a**), and are added to each of the sample values using the complete randomized block design (**Fig. 2c**). The variance within a sample is thus evenly split between the block effect and the remaining experimental error, which we presumably cannot partition further. The contribution of the block effect to the deviations is shown in **Figure 4b**, now a substantial component of the variance in each sample, unlike in **Figure 3b**, where blocking was not accounted for.

Having isolated variation owing to cell-culture differences, we increased sensitivity in detecting a treatment effect because our estimate of within-group variance is lower. Now $MS_W = 1.1$ and $F = 5.5$, which is significant at $P = 0.024$ and allows us to conclude that the treatment means are not all the same (**Fig. 4c**). By doing a *post hoc* pairwise comparison with the two-sample *t*-test, we can conclude that treatments A and C are different at an adjusted $P = 0.022$ (95% confidence interval (CI), 0.30–3.66) (**Fig. 4c**). We can calculate the *F* statistic for the blocking variable using $F = MS_{blk}/MS_W = 3.4$ to determine whether blocking had a significant effect. Mathematically, the blocking variable has the same role in the analysis as an experimental factor. Note that just because the blocking variable soaks up some of the variation we are not guaranteed greater sensitivity; in fact, because we estimate the block effect as well as the treatment effect, the within-group d.f. in the analysis is lower (e.g., changes from 15 to 10 in our case); our test may lose power if the blocks do not account for sufficient sample-to-sample variation.

Blocking increased the efficiency of our experiment. Without it, we would need nearly twice as large samples ($n = 11$) to reach the same power. The benefits of blocking should be weighed against any increase in associated costs and the decrease in d.f.: in some cases it may be more sensible to simply collect more data.

Note: Supplementary information is available in the online version of the paper (doi:10.1038/nmeth.3005).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 597–598 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 215–216 (2014).

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

POINTS OF SIGNIFICANCE

Replication

Quality is often more important than quantity.

Science relies heavily on replicate measurements. Additional replicates generally yield more accurate and reliable summary statistics in experimental work. But the straightforward question, 'how many and what kind of replicates should I run?' belies a deep set of distinctions and tradeoffs that affect statistical testing. We illustrate different types of replication in multilevel ('nested') experimental designs and clarify basic concepts of efficient allocation of replicates.

Replicates can be used to assess and isolate sources of variation in measurements and limit the effect of spurious variation on hypothesis testing and parameter estimation. Biological replicates are parallel measurements of biologically distinct samples that capture random biological variation, which may itself be a subject of study or a noise source. Technical replicates are repeated measurements of the same sample that represent independent measures of the random noise associated with protocols or equipment. For biologically distinct conditions, averaging technical replicates can limit the impact of measurement error, but taking additional biological replicates is often preferable for improving the efficiency of statistical testing.

Nested study designs can be quite complex and include many levels of biological and technical replication (**Table 1**). The distinction between biological and technical replicates depends on which sources of variation are being studied or, alternatively, viewed as noise sources.

An illustrative example is genome sequencing, where base calls (a statistical estimate of the most likely base at a given sequence position) are made from multiple DNA reads of the same genetic locus. These reads are technical replicates that sample the uncertainty in the sequencer readout but will never reveal errors present in the library itself. Errors in library construction can be mitigated by constructing technical replicate libraries from the same sample. If additional resources are available, one could potentially return to the source tissue and collect multiple samples to repeat the entire sequencing workflow. Such replicates would be technical if the samples were considered to be from the same aliquot or biological if considered to be from different aliquots of biologically distinct material[1]. Owing to historically high costs per assay, the field of genome sequencing has not demanded such replication. As the need for accuracy increases and the cost of sequencing falls, this is likely to change.

Table 1 | Replicate hierarchy in a hypothetical mouse single-cell gene expression RNA sequencing experiment

	Replicate type	Replicate category[a]
Animal study subjects	Colonies	B
	Strains	B
	Cohoused groups	B
	Gender	B
	Individuals	B
Sample preparation	Organs from sacrificed animals	B
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	B
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	V[b]
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

[a]Replicates are categorized as biological (B), technical (T) or of variable type (V). [b]Sequence reads serve diverse purposes depending on the application and how reads are used in analysis.

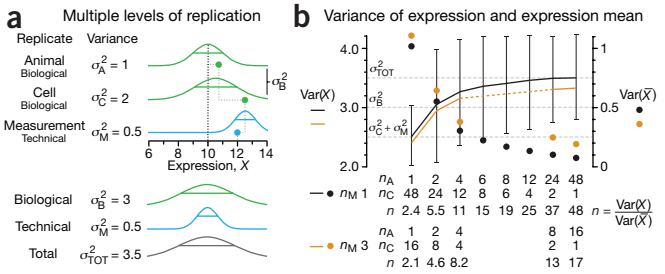

Figure 1 | Replicates do not contribute equally and independently to the measured variability, which can often underestimate the total variability in the system. (**a**) Three levels of replication (two biological, one technical) with animal, cell and measurement replicates normally distributed with a mean across animals of 10 and ratio of variances 1:2:0.5. Solid green (biological) and blue (technical) dots show how a measurement of the expression (X = 12) samples from all three sources of variation. Distribution s.d. is shown as horizontal lines. (**b**) Expression variance, Var(X), and variance of expression mean, Var($\bar{X}$), computed across 10,000 simulations of $n_A n_C n_M$ = 48 measurements for unique combinations of the number of animals (n_A = 1 to 48), cells per animal (n_C = 1 to 48) and technical replicate measurements per cell (n_M = 1 and 3). The ratio of Var(X) and Var($\bar{X}$) is the effective sample size, n, which corresponds to the equivalent number of statistically independent measurements. Horizontal dashed lines correspond to biological and total variation. Error bars on Var(X) show s.d. from the 10,000 simulated samples (n_M = 1).

How does one determine the types, levels and number of replicates to include in a study, and the extent to which they contribute information about important sources of variation? We illustrate the approach to answering these questions with a single-cell sequencing scenario in which we measure the expression of a specific gene in liver cells in mice. We simulated three levels of replication: animals, cells and measurements (**Fig. 1a**). Each level has a different variance, with animals ($\sigma_A^2 = 1$) and cells ($\sigma_C^2 = 2$) contributing to a total biological variance of $\sigma_B^2 = 3$. When technical variance from the assay ($\sigma_M^2 = 0.5$) is included, these distributions compound the uncertainty in the measurement for a total variance of $\sigma_{TOT}^2 = 3.5$. We next simulated 48 measurements, allocated variously between biological replicates (the number of animals, n_A and number of cells sampled per animal, n_C) and technical replicates (number of measurements taken per cell, n_M) for a total number of measurements $n_A n_C n_M = 48$. Although we will always make 48 measurements, the effective sample size, n, will vary from about 2 to 48, depending on how the measurements are allocated. Let us look at how this comes about.

Our ability to make accurate inferences will depend on our estimate of the variance in the system, Var(X). Different choices of n_A, n_C and n_M impact this value differently. If we sample n_C = 48 cells from a single animal (n_A = 1) and measure each n_M = 1 times, our estimate of the total variance σ_{TOT}^2 will be Var(X) = 2.5 (**Fig. 1b**). This reflects cell and measurement variances ($\sigma_C^2 + \sigma_M^2$) but not animal variation; with only one animal sampled we have no way of knowing what the animal variance is. Thus Var(X) certainly underestimates σ_{TOT}^2, but we would not know by

how much. Moreover, the uncertainty in Var(X) (error bar at $n_A = 1$; **Fig. 1b**) is the error in $\sigma_C^2 + \sigma_M^2$ and not σ_{TOT}^2. At another extreme, if all our measurements are technical replicates ($n_A = n_C = 1$, $n_M = 48$) we would find Var(X) = 0.5 (not represented in **Fig. 1**). This is only the technical variance; if we misinterpreted this as biological variation and used it for biological inference, we would have an excess of false positives. Be on the lookout: unusually small error bars on biological measurements may merely reflect measurement error, not biological variation. To obtain the best estimate of σ_{TOT}^2 we should sample $n_C = 1$ cells from $n_A = 48$ animals because each of the 48 measurements will independently sample each of the distributions in **Figure 1a**.

Our choice of the number of replicates also influences Var($\bar{X}$), the precision in the expression mean. The optimal way to minimize this value is to collect data from as many animals as possible ($n_A = 48$, $n_C = n_M = 1$), regardless of the ratios of variances in the system. This comes from the fact that n_A contributes to decreasing each contribution to Var($\bar{X}$), which is given by $\sigma_A^2/n_A + \sigma_C^2/n_A n_C + \sigma_M^2/n_A n_C n_M$. Although technical replicates allow us to determine σ_M^2, unless this is a quantity of interest, we should omit technical replicates and maximize n_A. Of course, good blocking practice suggests that samples from the different animals and cells should be mixed across the sequencing runs to minimize the effect of any systematic run-to-run variability (not present in simulated data here).

The value in additional measurements can be estimated by the prospective improvement in effective sample size. We have seen before that the variance in the mean of a random variable is related to its variance by Var(X) = nVar($\bar{X}$). The ratio of Var(X) to Var($\bar{X}$) can therefore be used as a measure of the equivalent number of independent samples. From **Figure 1b**, we can see that n = 48 only for n_A = 48 and drops to n = 25 for n_A, n_C = 12, 4 and is as low as about 2 for n_A = 1. In other words, even though we may be collecting additional measurements they do not all contribute equally to an increase in the precision of the mean. This is because additional cell and technical replicates do not correspond to statistically independent values: technical replicates are derived from the same cell and the cell replicates from the same animal. If it is necessary to summarize expression variability at the level of the animals, then cells from a given animal are pseudoreplicates—statistically correlated in a way that is unique to that animal and not representative of the population under study. Not all replicates yield statistically independent measures, and treating them as if they do can erroneously lower the apparent uncertainty of a result.

The number of replicates has a practical effect on inference errors in analysis of differences of means or variances. We illustrate this by enumerating inference errors in 10,000 simulated drug-treatment experiments in which we vary the number of animals and cells (**Fig. 2**). We assume a 10% effect chance for two scenarios: a twofold increase in variance, σ_C^2, or a 10% increase in mean, μ_A, using the same values for other variances and 48 total measurements as in **Figure 1**. Applying the t-test, we show false discovery rate (FDR) and power for detecting these differences (**Fig. 2**). If we want to detect a difference in variation across cells, it is best to choose $n_A \approx n_C$ in our range. On the other hand, when we are interested in changes in mean expression across mice, it is better to sample as many mice as possible. In either case, increasing the number of measurements from 48 to 144 by taking three technical replicates (n_M = 3) improves inference only slightly.

Biological replicates are preferable to technical replicates for inference about the mean and variance of a biological population.

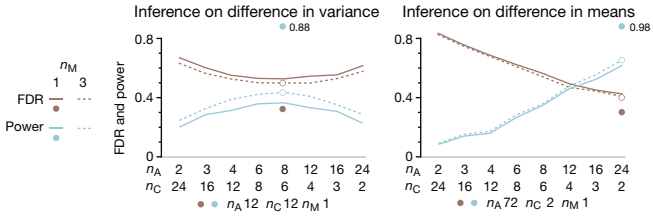

Figure 2 | The number of replicates affects FDR and power of inferences on the difference in variances and means. Shown are power and FDR profiles of a test of difference in cell variances (left) and animal means (right) for 48 (n_M = 1) or 144 (n_M = 3) measurements using different combinations of n_A and n_C. Vertical arrows indicate change in FDR and power when technical replicates are replaced by biological replicates, as shown by n_A,n_C,n_M, for the same number of measurements (144). Values generated from 10,000 simulations of a 10% chance of a treatment effect that increases cell variance $2\sigma_C^2$ or animal mean 1.1 × μ_A. Samples were tested with two-sample t-test (sample size n_A) at two-tailed α = 0.05.

(**Fig. 2**). For example, changing n_A,n_C,n_M from 8,6,3 (where power is highest) to 12,12,1 doubles the power (0.43 to 0.88) in detecting a twofold change in variance. In the case of detecting a 10% difference in means, changing n_A,n_C,n_M from 24,2,3 to 72,2,1 increases power by about 50% from 0.66 to 0.98. Practically, the cost difference between biological and technical replicates should be considered; this will affect the cost-benefit tradeoff of collecting additional replicates of one type versus the other. For example, if the cost units of animals to cells to measurements is 10:1:0.1 (biological replicates are likely more expensive than technical ones) then an experiment with n_A,n_C,n_M of 12,12,1 is about twice as expensive as that with 8,6,3 (278 versus 142 cost units). However, power in detecting a change in variance is doubled as well, so the cost increase is commensurate with increase in efficiency. In the case of detecting differences in means, 72,2,1 is about three times as expensive as 24,2,3 (878 versus 302 cost units) but increases power only by 50%, making this a lower-value proposition.

Typically, biological variability is substantially greater than technical variability, so it is to our advantage to commit resources to sampling biologically relevant variables unless measures of technical variability are themselves of interest, in which case increasing the number of measurements per cell, n_M, is valuable.

Good experimental design practice includes planning for replication. First, identify the questions the experiment aims to answer. Next, determine the proportion of variability induced by each step to distribute the capacity for replication of the experiment across steps. Be aware of the potential for pseudoreplication and aim to design statistically independent replicates.

As our capacity for higher-throughput assays increases, we should not be misled into thinking that more is always better. Clear thinking about experimental questions and sources of variability is still crucial to produce efficient study designs and valid statistical analyses.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Paul Blainey, Martin Krzywinski & Naomi Altman

1. Robasky, K., Lewis, N.E. & Church, G.M. *Nat. Rev. Genet.* **15**, 56–62 (2014).

Paul Blainey is an Assistant Professor of Biological Engineering at MIT and Core Member of the Broad Institute. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.

POINTS OF SIGNIFICANCE

Nested designs

For studies with hierarchical noise sources, use a nested analysis of variance approach.

Many studies are affected by random-noise sources that naturally fall into a hierarchy, such as the biological variation among animals, tissues and cells, or technical variation such as measurement error. With a nested approach, the variation introduced at each hierarchy layer is assessed relative to the layer below it. We can use the relative noise contribution of each layer to optimally allocate experimental resources using nested analysis of variance (ANOVA), which generally addresses replication and blocking, previously discussed *ad hoc*[1,2].

Recall that factors are independent variables whose values we control and wish to study[3] and which have systematic effects on the response. Noise limits our ability to detect effects, but known noise sources (e.g., cell culture) can be mitigated if used as blocking factors[2]. We can model the contribution of each blocking factor to the overall variability, isolate it and increase power[2]. Statisticians distinguish between fixed factors, typically treatments, and random factors, such as blocks.

The impact of fixed and random factors in the presence of experimental error is shown in **Figure 1**. For a fixed factor (**Fig. 1a**), each of its levels (for example, a specific drug) has the same effect in all experiments and an unmodeled uncertainty due to experimental error. The levels of a fixed factor can be exactly duplicated (level A1 in **Fig. 1a** is identical for each experiment) and are of specific interest, usually the effect on the population mean.

In contrast, when we repeat an experiment, the levels of a random factor are sampled from a population of all possible levels of the factor (replicates) and are different across all the experiments, emphasized by unique level labels (B1–B9; **Fig. 1b**). Because the levels cannot be exactly duplicated, their effect is random and they are not of specific interest. Instead, we use the sample of levels to model the uncertainty added by the random factor (for example, all mice).

Fixed and random factors may be crossed or nested (**Fig. 2**). When crossed, all combinations of factors are used to study the main effects and interactions of two or more factors (**Fig. 2a**). In contrast, nested designs apply a hierarchy—some level combinations are not studied because the levels cannot be duplicated or reused (**Fig. 2b**). Random factors (for example, mouse and cell) are nested within the fixed factor (drug) to measure noise due to individual mice and cells and to generalize the effects of the fixed

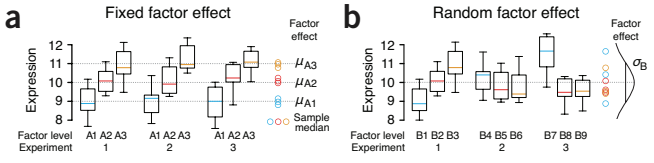

Figure 1 | Inferences about fixed factors are different than those about random factors, as shown by box-plots of $n = 10$ samples across three independent experiments. Circles indicate sample medians. Box-plot height reflects simulated measurement error ($\sigma_\varepsilon^2 = 0.5$). (**a**) Fixed factor levels are identical across experiments and have a systematic effect on the mean. (**b**) Random factor levels are samples from a population, have a random effect on the mean and contribute noise to the system ($\sigma_B^2 = 1$).

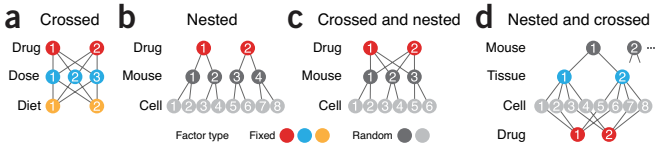

Figure 2 | Factors may be crossed or nested. (**a**) A crossed design examines every combination of levels for each fixed factor. (**b**) Nested design can progressively subreplicate a fixed factor with nested levels of a random factor that are unique to the level within which they are nested. (**c**) If a random factor can be reused for different levels of the treatment, it can be crossed with the treatment and modeled as a block. (**d**) A split plot design in which the fixed effects (tissue, drug) are crossed (each combination of tissue and drug are tested) but themselves nested within replicates.

factor on all mice and cells. If mice can be reused, we can cross them with the drug and use them as a random blocking factor[2] (**Fig. 2c**).

We will use the design in **Figure 2b** to illustrate the analysis of nested fixed and random factors using nested ANOVA, similar to the ANOVA discussed previously[2]. Now nesting is taken into account and the calculations have different interpretations because some of the factors are random. The fixed factor may have an effect on the mean, and the two random factors will add uncertainty. We will be able to estimate the amount of variance for each random factor and use it to better plan our replication strategy. We can maximize power (for example, within cost constraints) to detect a difference in means due to the top-level fixed factor or to detect variability due to random factors. The latter is biologically interesting when increased variance in cell response may be due to increased heterogeneity in the genotypes and implicated in drug resistance.

We will simulate the nested design in **Figure 2b** using three factors: A ($a = 2$ levels: control and treatment), B (mice, $b = 5$ levels, $\sigma_B^2 = 1$), C (cells, $c = 5$ levels, $\sigma_C^2 = 2$). Expression for each cell will be measured using three technical replicates ($\sigma_\varepsilon^2 = 0.5$, $n = 3$). The raw sample data of the simulation are shown in **Figure 3a**.

Nested ANOVA calculations begin with the sum of squared deviations (SS) to partition the variance among the factors, exactly as in regular ANOVA. For example, the first blue arrow in **Figure 3a** represents the difference between the averages of all points from mouse B4 (X_{14}..) and all points from the control (X_{1}...). Factor C has the largest deviations (**Fig. 3b**) because it was modeled to be the largest source of noise ($\sigma_C^2 = 2$). The distinction between regular and nested ANOVA is how the mean squares (MS) enter into the calculation of the *F*-ratio for each factor. The *F*-ratio is a ratio of MS values, and the denominator corresponds to the MS of the next nested factor (for example, MS_B/MS_C) and not MS_E (see **Supplementary Table 1** for nested ANOVA formulas and calculated values; see **Supplementary Table 2** for expected values of MS). The *F*-test uses the ratio of between-group sample variance (estimate of population variance from sample means) and within-group variance (estimate of population variance from sample variances) to test whether group means differ (for fixed factors). In the case of random factors, the interpretation is whether the factor contributes noise in addition to the noise due to the factor nested within it (for example, is there more mouse-to-mouse variability than would be expected from cell-to-cell variability?).

At the bottom of the nested hierarchy ($n = 3$ technical replicates per cell), we find $MS_E = 0.55$, which is an estimate of $\sigma_\varepsilon^2 = 0.5$ in our simulation. We find statistically significant (at $\alpha = 0.05$) contributions to noise from both mice (factor B) and cells (factor C) with estimated variance contributions of 0.84 and 2.1, respectively, which matches

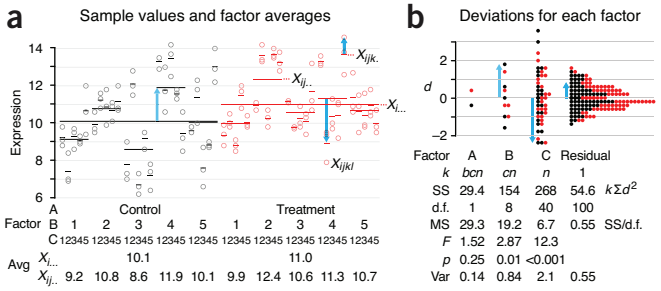

Figure 3 | Data and analysis for a simulated three-factor nested experiment. (**a**) Simulated expression levels, X_{ijkl}, measured for $a = 2$ levels of factor A (control and treatment, i), $b = 5$ of factor B (mice, j), $c = 5$ of factor C (cells, k) and $n = 3$ technical replicates (l). Averages across factor levels are shown as horizontal lines and denoted by dots in subscript for the factor's index. Blue arrows illustrate deviations used for calculation of sum of squares (SS). Data are simulated with $\mu_c = 10$ for control and $\mu_t = 11$ for treatment and $\sigma_B^2 = 1$, $\sigma_C^2 = 2$, $\sigma_\varepsilon^2 = 0.5$ for noise at mouse, cell and technical replicate levels, respectively. Values below the figure show factor levels and averages at levels of A ($X_{i...}$) and B ($X_{ij..}$). Labels for the levels of B and C are reused but represent distinct individual mice and cells. (**b**) Histogram of deviations (d) for each factor. Three deviations illustrated in **a** are identified by the same blue arrows. Nested ANOVA calculations show number of times (k) each deviation (d) contributes to SS, degrees of freedom (d.f.), mean squares (MS), F-ratio, P value and the estimated variance contribution of each factor.

our inputs $\sigma_B^2 = 1$ and $\sigma_C^2 = 2$. Because the top-layer factor is fixed and not considered a source of noise, its variance component is not a useful quantity—of interest is its effect on the mean. Unfortunately, we were unable to detect a difference in means for A ($P = 0.25$) because of poor power due to our allocation of replicates. It is useful to relate the F-test for factor A to a two-sample t-test to understand the statistical quantities involved and calculate power.

The F-test for the top-layer factor A ($F = \mathrm{MS}_A/\mathrm{MS}_B$) tests the difference between the variances of treatment and mouse means. Any treatment effect on the mean will show up as additional variance, which we stand a chance to detect. Because we have only two levels of factor A, the F-test, which has degrees of freedom (d.f.) of $a - 1 = 1$ and $a(b - 1) = 8$, is equivalent to the two-sample t-test for samples of size b, $2(b - 1)$ d.f. and with $t = \sqrt{F}$. This t-test is applied to the control and treatment samples formed using $b = 5$ averages X_{ij}.. (**Fig. 3a**) whose expected variance is $\mathrm{E}[\mathrm{Var}(X_{ij}..)] = \sigma_B^2 + \sigma_C^2/c + \sigma_\varepsilon^2/(cn) = 1.43$ (ref. 1). This quantity is estimated by $\mathrm{MS}_B/(cn) = 1.28$, which is exactly the average variance of the two sample variances 1.73 and 0.83 (**Supplementary Table 3**). These samples yield the control and treatment means of 10.1 and 11.0 (X_{i}...; **Fig. 3a**) and a t-statistic of $0.9/\sqrt{(2\mathrm{MS}_B/(bcn))} = 1.24$, which yields the same P value of 0.25 as from the F-test.

We can now calculate the t-test power for our scenario. For a difference in means of $d = 1$, the power using samples of size $b = 5$ is 0.21, using the expected variance 1.43. In practice, we might run a trial experiment to determine this value using $\mathrm{MS}_B/(cn)$. Clearly, our initial choice of b, c and n was an inadequate design—we should aim for a power of at least 0.8. If variance is kept at 1.43 ($c = 5$, $n = 5$), this power can be achieved for a sample size $b = 24$. With 24 mice, the expected variance of the average across mice would be $\mathrm{E}[\mathrm{Var}(X_{i}...)] = 1.43/24$. Dividing this into the total variance due to replication ($\sigma_B^2 + \sigma_C^2 + \sigma_\varepsilon^2 = 3.5$), we can calculate the effective sample size, 57 (ref. 1). As we've previously seen, this can be achieved with the fewest number of measurements if we have $b = 57$ mice and $c = n = 1$. If we assume the cost of mice, cells and technical replicates to be 100, 10 and 1, respectively, these designs would cost 3,960 ($b = 24$, $c = 5$, $n = 3$) and 6,327 ($b = 57$, $c = 1$, $n = 1$). Let's see if we can use fewer mice and increase replication to obtain the same power at a lower cost.

The nested analysis provides a general framework for these cost and power calculations. The optimum number of replicates at each level can be calculated on the basis of the cost of replication and the variance at the level of the factor. We want to minimize $\mathrm{Var}(X_{i}...) = \sigma_B^2/b + \sigma_C^2/(bc) + \sigma_\varepsilon^2/(bcn)$ within the cost constraint $K = bC_B + bcC_D + bcnC_D$ (C_X is cost per replicate at factor X) with the goal of finding values of b, c and n that provide the largest decrease in the variance per unit cost. The optimum number of technical replicates is $n^2 = C_C/C_D \times \sigma_\varepsilon^2/\sigma_C^2$. In other words, subreplicates are preferred to replicates when they are cheaper and their factor is a source of greater noise. With the costs as given above ($C_C/C_N = 10$) we find $n^2 = 10 \times 0.5/2 = 2.5$ and $n = 2$. We can apply the same equation for the number of cells, $c^2 = C_B/C_C \times \sigma_C^2/\sigma_B^2$, where C_B is the cost of a mouse. Using the same tenfold cost ratio, $c^2 = 10 \times 2/1 = 20$ and $c = 5$. For $c = 5$ and $n = 2$, $\mathrm{Var}(X_{ij}..)$ is 1.45, and we would reach a power of 0.8 if we had $b = 24$ mice. This experiment is slightly cheaper than the one with $n = 3$ (3,840 vs. 3,960).

Two components affect power in detecting differences in means. Subreplication at the cell and technical layer helps increase power by decreasing the variance of mouse averages, $\mathrm{Var}(X_{ij}..)$, used for t-test samples. The number of mice also increases power because it decreases the standard error of $X_{ij}..$ (the precision of mouse averages) because sample size is increased. To obtain the largest power to detect a treatment effect with the fewest number of measurements, it is always best to pick as many mice as possible: effective sample size is largest and variance of sample averages is lowest.

The number of replicates also affects our ability to detect the noise contribution from each random factor. If detecting and estimating variability in mice and cells is of interest, we should aim to increase the power of the associated F-tests (**Supplementary Table 1**). For example, under the alternative hypothesis of a nonzero contribution of cells to noise (σ_C^2), the F-statistic will be distributed as a multiple of the null hypothesis F-statistic, $F_{u,v} \times (n\sigma_C^2 + \sigma_\varepsilon^2)/\sigma_\varepsilon^2$. The multiplication factor is the ratio of expected MS values (**Supplementary Table 2**). For our simulation values, the multiple is 13 and the d.f. are $u = 40$ and $v = 100$. The critical F-value is 1.52, and our power is the P value for 1.52/13, which is essentially 1 (this is why the P value for factor C in **Fig. 2b** is very low). For level B we have $u = 8$, $v = 40$, a multiple of 3.3 (21.5/6.5) and a power of 0.72. The power of our design to detect noise within mice and cells was much higher than that for detecting an effect of the treatment on the means.

Nested designs are useful for understanding sources of variability in the hierarchy of the subsamples and can reduce the cost of the experiment when costs vary across the hierarchy. Statistical conclusions can be made only about the layers actually replicated—technical replication cannot replace biological replication for biological inference.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3137).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Martin Krzywinski, Naomi Altman & Paul Blainey

1. Blainey, P., Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 879–880 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 597–598 (2014).

Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University. Paul Blainey is an Assistant Professor of Biological Engineering at MIT and Core Member of the Broad Institute.

POINTS OF SIGNIFICANCE

Two-factor designs

When multiple factors can affect a system, allowing for interaction can increase sensitivity.

When probing complex biological systems, multiple experimental factors may interact in producing effects on the response. For example, in studying the effects of two drugs that can be administered simultaneously, observing all the pairwise level combinations in a single experiment is more revealing than varying the levels of one drug at a fixed level of the other. If we study the drugs independently we may miss biologically relevant insight about synergies or antisynergies and sacrifice sensitivity in detecting the drugs' effects.

The simplest design that can illustrate these concepts is the 2 × 2 design, which has two factors (A and B), each with two levels (*a*/*A* and *b*/*B*). Specific combinations of factors (*a*/*b*, *A*/*b*, *a*/*B*, *A*/*B*) are called treatments. When every combination of levels is observed, the design is said to be a complete factorial or completely crossed design. So this is a complete 2 × 2 factorial design with four treatments.

Our previous discussion about experimental designs was limited to the study of a single factor for which the treatments are the factor levels. We used ANOVA[1] to determine whether a factor had an effect on the observed variable and followed up with pairwise *t*-tests[2] to isolate the significant effects of individual levels. We now extend the ANOVA idea to factorial designs. Following the ANOVA analysis, pairwise *t*-tests can still be done, but often analysis focuses on a different set of comparisons: main effects and interactions.

Figure 1 illustrates some possible outcomes in a 2 × 2 factorial experiment (values in **Table 1**). Suppose that both factors correspond to drugs and the observed variable is liver glucose level. In **Figure 1a**, drugs A and B increase glucose levels by 1 unit. Because neither drug influences the effect of the other we say there is no interaction and that the effects are additive. In **Figure 1b**, the effect of A in the presence of B is larger than the sum of their effects when they are administered separately (3 vs. 0.5 + 1). When the effect of the levels of a factor depends on the levels of other factors, we say that there is an interaction between the factors. In this case, we need to be careful about defining the effects of each factor.

The main effect of factor A is defined as the difference in the means of the two levels of A averaged over all the levels of B. For **Figure 1b**, the average for level *a* is $\tau = (0 + 1)/2 = 0.5$ and for level *A* is $\tau = (0.5 + 3)/2 = 1.75$, giving a main effect of 1.75 – 0.5 = 1.25 (**Table 1**). Similarly, the main effect of B is 2 – 0.25 = 1.75. The interaction compares the differences in the mean of A at the two levels of B (2 – 0.5 = 1.5; in the Δ row) or, equivalently, the differences in the mean of B at the two levels of A (2.5 – 1 = 1.5). Interaction plots are useful to evaluate effects when the number of factors is small (line plots, **Fig 1b**). The *x* axis represents levels of one factor and lines correspond to levels of other factors. Parallel lines indicate no interaction. The more the lines diverge, or cross, the greater the interaction.

Figure 1c shows an interaction effect with no main effect. This can happen if one factor increases the response at one level of the other factor but decreases it at the other. Both factors have the same average value for each of their levels, $\tau = 0.5$. However, the two factors do interact because the effect of one drug is different depending on the presence of the other.

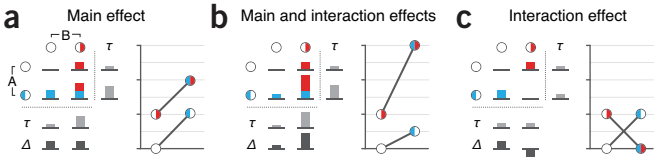

Figure 1 | When studying multiple factors, main and interaction effects can be observed, shown here for two factors (A, blue; B, red) with two levels each. (**a**) The main effect is the difference between τ values (light gray), which is the response for a given level of a factor averaged over the levels of other factors. (**b**) The interaction effect is the difference between effects of A at the different levels of B or vice versa (dark gray, Δ). (**c**) Interaction effects may mask main effects.

There are various ways in which effects can combine; their clear and concise reporting is important. For a 2 × 2 design with two levels per factor, effects can be estimated directly from treatment means. In this case, effects should be summarized with their estimated value and a confidence interval (CI) and graphically reported as a plot of means with error bars[2]. Optionally, a two-sample *t*-test can be used to provide a *P* value for the null hypothesis that the two treatments have the same effect—a zero difference in their means. For example, with levels *a*/*A* and *b*/*B* we have four treatment means μ_{ab}, μ_{Ab}, μ_{aB} and μ_{AB}. The effect of A at level *b* is $\mu_{Ab} - \mu_{ab}$, which is estimated by substituting the observed sample means. The standard error of this estimate is s.e. $= s\sqrt{(1/n_{Ab} + 1/n_{ab})}$, where *s* is the estimate of the population standard deviation, estimated by $\sqrt{MS_E}$, where MS_E is the residual mean square from the ANOVA, and n_{ij} is the observed sample size for treatment A = *i* and B = *j*. If the design is balanced, $n_{Ab} = n_{ab} = n$ and s.e. $= \sqrt{(2MS_E/n)}$. The *t*-statistic is $t = (\bar{x}_{Ab} - \bar{x}_{ab})/\text{s.e.}$ The CI can be constructed using $\bar{x}_{Ab} - \bar{x}_{ab} \pm t^* \times \text{s.e.}$, where t^* is the critical value for the *t*-statistic at the desired α. Note, however, that the degrees of freedom (d.f.) are the error d.f. from the ANOVA, not $2(n - 1)$ as in the usual two-sample *t*-test[2], because the MS_E rather than the sample variances is used in the s.e. computation.

When there are more factors or more levels, the main effects and interactions are summarized over many comparisons as sums of squares (SS) and usually only the test statistic (*F*-test), its d.f. and the *P* value are reported. If there are statistically significant interactions, pairwise comparisons of different levels of one factor for fixed levels of the other factors (sometimes called simple main effects) are often computed in the manner described above. If the interactions are not significant, we typically compute differences between levels of one factor averaged over the levels of the other factor. Again, these are pairwise comparisons between means that are handled as just described, except that the sample sizes are also summed over the levels.

To illustrate the two-factor design analysis, we'll use a simulated data set in which the effect of levels of the drug and diet were tested in two different designs, with 8 mice and 8 observations (**Fig. 2a**). We'll assume an experimental protocol in which a mouse liver tissue sample is tested for glucose levels using two-way ANOVA. Our simulated simple effects are shown in **Figure 1b**—the increase in the response variable is 0.5 (*A*/*b*), 1 (*a*/*B*) and 3 (*A*/*B*). The two drugs are synergistic—A is 4× as potent in the presence of B, as can be seen by $(\mu_{AB} - \mu_{aB})/(\mu_{Ab} - \mu_{ab}) = \Delta_B/\Delta_b = 2/0.5 = 4$ (**Table 1**). We'll assume the same variation due to mice and measurement error, $\sigma^2 = 0.25$.

Table 1 | Quantities used to determine main and interaction effects from data in **Figure 1**

	Main effect			Main and interaction effects			Interaction effect		
	b	***B***	τ	***b***	***B***	τ	***b***	***B***	τ
a	0	1	0.5	0	1	0.5	0	1	0.5
A	1	2	1.5	0.5	3	1.75	1	0	0.5
τ	0.5	1.5		0.25	2		0.5	0.5	
Δ	1	1		0.5	2		1	−1	

Treatment values shown are means for *a/b*, *a/B*, *A/b* and *A/B* level combinations. A main effect is observed if the difference between τ values (e.g., 1.5 − 0.5 = 1) is nonzero. An interaction effect is observed if Δ, the difference between the mean levels of A, varies across levels of B or vice versa.

We'll use a completely randomized design with each of the 8 mice randomly assigned to one of the four treatments in a balanced fashion each providing a single liver sample (**Fig. 2a**). First, let's test the effect of the two factors separately using one-way ANOVA, averaging over the values of the other factor. If we consider only A, the effects of B are considered part of the residual error and we do not detect any effect ($P = 0.48$, **Fig. 2b**). If we consider only B, we can detect an effect ($P = 0.04$) because B has a larger main effect (2.0 − 0.25 = 1.75) than A (1.75 − 0.5 = 1.25).

When we test for multiple factors, the ANOVA calculation partitions the total sum of squares, SS_T, into components that correspond to A (SS_A), B (SS_B) and the residual (SS_E) (**Fig. 2b**). The additive two-factor model assumes that there is no interaction between A and B—the effect of a given level of A does not depend on a level of B. In this case, the interaction component is assumed to be part of the error. If this assumption is relaxed, we can partition the total variance into four components, now accounting for how the response of A varies with B. In our example, the SS_A and SS_B terms remain the same, but SS_E is reduced by the amount of SS_{AB} (4.6), to 2.0 from 6.6. The resulting reduction in MS_E (0.5 vs. 1.3) corresponds to the variance explained by the interaction between the two factors. When interaction is accounted for, the sensitivity of detecting an effect of A and B is increased because the *F*-ratio, which is inversely proportional to MS_E, is larger.

To calculate the effect and interaction CIs, as described above, we start with the treatment means $\bar{x}_{ab} = 0.27$, $\bar{x}_{Ab} = -0.39$, $\bar{x}_{aB} = 0.86$ and $\bar{x}_{AB} = 3.23$, each calculated from two values. To calculate the main effects of A and B, we average over four measurements to find $\bar{x}_a = 0.57$, $\bar{x}_A = 1.42$, $\bar{x}_b = -0.06$ and $\bar{x}_B = 2.05$. The residual error $MS_E = 0.5$ is used to calculate the s.e. of main effects: $\sqrt{(2MS_E/n)} = \sqrt{(2 \times 0.5/4)} = 0.5$. The critical *t*-value at $\alpha = 0.05$ and d.f. = 4 is 2.78, giving a 95% CI for the main effect of A to be 0.9 ± 1.4 ($F_{1,4} = 2.9$), where d.f. = (1,4) and of B to be 2.1 ± 1.4 ($F_{1,4} = 17.6$). The CIs reflect that we detected the main effect of B but not of A. For the interaction, we find $(\bar{x}_{AB} - \bar{x}_{aB}) - (\bar{x}_{Ab} - \bar{x}_{ab}) = 3.0$ with s.e. = 1 and a CI of 3.0 ± 2.8 ($F_{1,4} = 9.1$).

To improve the sensitivity of detecting an effect of A, we can mitigate biological variability in mice by using a randomized complete block approach[1] (**Fig. 2a**). If the mice share some characteristic, such as litter or weight which contributes to response variability, we could control for some of the variation by assigning one complete replicate to each batch of similar mice. The total number of observations will still be 8, and we will track the mouse batch across measurements and use the batch as a random blocking factor[2]. Now, in addition to the effect of interaction, we can further reduce the MS_E by the amount of variance explained by the block (**Fig. 2b**).

The sum-of-squares partitioning and *P* values for the blocking scenario are shown in **Figure 2b**. In each case, the SS_E value is proportionately lower than in the completely randomized design, which makes the tests more sensitive. Once we incorporate blocking and interaction, we are able to detect both main and interaction effects and account for nearly all of the variance due to sources other than measurement error ($SS_E = 0.8$, $MS_E = 0.25$). The interpretation of $P = 0.01$ for the blocking factor M is that the biological variation due to the blocking factor has a nonzero variance. Effects and CIs are calculated just as for the completely randomized design—although the means have two sources of variance (block effect and MS_E), their difference has only one (MS_E) because the block effect cancels.

With two factors, more complicated designs are also possible. For example, we might expose the whole mouse to a drug (factor A) *in vivo* and then expose two liver samples to different *in vitro* treatments (factor B). In this case, the two liver samples from the same mouse form a block that is nested in mouse.

We might also consider factorial designs with more levels per factor or more factors. If the response to our two drugs depends on genotype, we might consider using three genotypes in a 2 × 2 × 3 factorial design with 12 treatments. This design allows for the possibility of interactions among pairs of factors and also among all three factors. The smallest factorial design with *k* factors has two levels for each factor, leading to 2^k treatments. Another set of designs, called fractional factorial designs, used frequently in manufacturing, allows for a large number of factors with a smaller number of samples by using a carefully selected subset of treatments.

Complete factorial designs are the simplest designs that allow us to determine synergies among factors. The added complexity in visualization, summary and analysis is rewarded by an enhanced ability to understand the effects of multiple factors acting in unison.

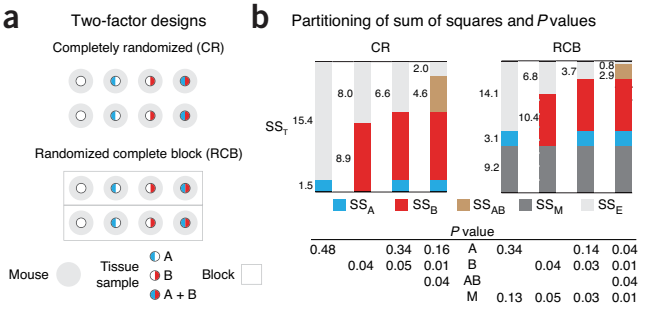

Figure 2 | In two-factor experiments, variance is partitioned between each factor and all combinations of interactions of the factors. (**a**) Two common two-factor designs with 8 measurements each. In the CR scenario, each mouse is randomly assigned a single treatment. Variability among mice can be mitigated by grouping mice by similar characteristics (e.g., litter or weight). The group becomes a block. Each block is subject to all treatments. (**b**) Partitioning of the total sum of squares (SS_T; CR, 16.9; RCB, 26.4) and *P* values for the CR and RCB designs in **a**. M represents the blocking factor. Vertical axis is relative to the SS_T. The total d.f. in both cases = 7; all other d.f. = 1.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Martin Krzywinski & Naomi Altman

1. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 215–216 (2014).
3. Montgomery, D.C. *Design and Analysis of Experiments* 8th edn. (Wiley, 2012).

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

POINTS OF SIGNIFICANCE

Sources of variation

To generalize conclusions to a population, we must sample its variation.

Variability is inevitable in experiments owing to both biological and technical effects. Whereas technical variability should be tightly controlled to enhance the internal validity of the results, some types of biological variability need to be maintained to allow generalization of the results to the population of interest. Experimental control, randomization, blocking and replication are the tools that allow replicable and meaningful results to be obtained in the face of variability.

In previous columns we have given examples of how variation limits our ability to detect effects by reducing the power of tests. This month we go into more detail about variability and how it affects our ability to replicate the experimental results (internal validity) and generalize from our experiment to the population (external validity).

Let's start with an idealized experiment, which we will then expand upon. Suppose that we are able to culture a single murine cell under tightly controlled conditions so that the response of different aliquots of the culture is identical. Also, suppose that our measuring device is so accurate that the difference between measurements of an aliquot is below the detection limit. If measurement does not disrupt the cell culture, we require only a single aliquot: we measure the baseline response, apply the treatment and measure the treatment response. No replication is needed because differences between the measurements can only be due to the treatment.

This idealized system has perfect internal validity—the response variable solely reflects the treatment effect, and repeating the experiment on another aliquot from the same cell culture will give identical results. However, the system lacks external validity—it tells us about only a specific cell from a specific mouse. We know that cells vary within a single tissue, and that tissues vary from mouse to mouse, but we cannot use this ideal system to make inferences about other cell cultures or other mice because we have no way of determining how much variability to expect. To do so requires that we sample the biological variation across relevant experimental variables (**Fig. 1**).

A well-designed experiment is a compromise between internal and external validity. Our goal is to observe a reproducible effect that can be due only to the treatment (avoiding confounding and bias) while simultaneously measuring the variability required to estimate how much we expect the effect to differ if the measurements are repeated with similar but not identical samples (replicates).

When administering the treatment *in vivo*, we can never control the many sources of biological variability in the mice sufficiently to achieve identical measurements for different animals. However, with careful design, we can reduce the impact of this variability on our measurements by controlling some of these factors.

Genotype and gender are examples of sources of variability that are under complete experimental control. We can eliminate the source entirely by selecting a single level or select several levels so that the effects can be determined. For gender we can observe all the possible levels, so we can treat gender as a fixed factor in our experiment. Genotype can be a fixed effect (specific genotypes of interest, such as a mutant and its background wild type) or a random (noise) effect (several wild-type strains representing the wild-type population). Only by deliberately introducing variability can we make general statements about treatment effect—and then only across factors that were varied.

Other sources of variability, such as diet, temperature and other housing effects, are under partial experimental control. Noise factors that cannot be controlled, or are unknown, can be handled by random assignment[1] (to avoid bias), replication[2] (to increase precision) and blocking[3] (to isolate noise).

When dealing with variation, two principles apply: the precision with which we can characterize a sample (e.g., s.e.m.) and the manner in which variances from different sources combine together[4]. The s.e.m. of a random sample is $\sigma/\sqrt{n}$, where σ is the s.d. of the population (also written as $\mathrm{Var}(\overline{X}) = \mathrm{Var}(X)/n$). With sufficient replication (large n), our precision in measuring the mean as measured by the s.e.m. can be made arbitrarily small (**Fig. 2a**). When multiple independent sources of variation are present, the variance of the measurement is the sum of individual variances.

These two principles can be combined to obtain the variation of the mean in a nested replication scenario[2] (**Supplementary Fig. 1**). Suppose that variances due to mouse, cell and measurement are M, C and ε (Var() is omitted for brevity). The variance of the measurement of a single cell will be $M + C + \varepsilon$, the sum of the individual variances. If we measure the same cell n_ε times, the variance of the average measurement will be $M + C + \varepsilon/n_\varepsilon$. If we measure n_C cells, each n_ε times, the variance will be $M + C/n_C + \varepsilon/(n_C \times n_\varepsilon)$. Finally, if we repeat the procedure for n_M mice, the variance will be reduced to $M/n_M + C/(n_M \times n_C) + \varepsilon/(n_M \times n_C \times n_\varepsilon)$. In general, the variance of each source is divided by the number of times that source is independently sampled. This is illustrated in **Figure 2b** for $M = 1$, $C = 4$ and $\varepsilon = 0.25$. As we have already seen[2], the number of replicates at each layer (n_M, n_C, n_ε) can be controlled to optimally reduce variation (increase power) within practical constraints (cost). For example, to reduce the total variance to 25% of the total $M + C + \varepsilon$, we can sample using $n_M = 4$, $n_C = 1$ or $n_M = n_C = 3$ (**Fig. 2b**). Sampling a single mouse allows us to reduce variance only to M, but it would not allow us to estimate the variation at the mouse layer and therefore would not allow for inference about the population of mice. For our example, technical variation is much smaller than biological variation, and technical replicates are of little value—variance is reduced by only 5% for $n_M = n_C = 1$ and $n_\varepsilon = 10$ (**Fig. 2b**, gray trace) and can be reduced only to $M + C$.

When measurements themselves are an average of a large number of contributing factors, biological variability of the components can be underestimated. For example, measuring two samples from the

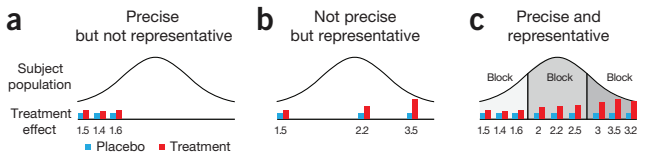

Figure 1 | Internal and external validity relate respectively to how precise and representative the results are of the population of interest. (**a**) Sampling only a part of the population may create precise measurements, but generalizing to the rest of the population can result in bias. (**b**) Better representation can be achieved by sampling across the population, but this can result in highly variable measurements. (**c**) Identifying blocks of similar subjects within the population increases the precision (within block) and captures population variability (between blocks).

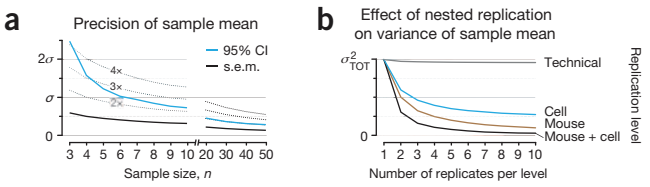

Figure 2 | In the presence of variability, the precision in sample mean can be improved by increasing the sample size, or the number of replicates in a nested design. (**a**) Increasing the sample size, n, improves the precision in the mean by $1/\sqrt{n}$ as measured by the s.e.m. The 95% CI is a more intuitive measure of precision: the range of values that are not significantly different at $\alpha = 0.05$ from the observed mean. The 95% confidence interval (CI) shrinks as $t^*/\sqrt{n}$, where t^* is the critical value of the Student's t-distribution at two-tailed $\alpha = 0.05$ and $n - 1$ degrees of freedom. t^* decreases from 4.3 ($n = 3$) to 2.0 ($n = 50$). Dotted lines represent constant multiples of the s.e.m. (**b**) For a nested design with mouse, cell and technical variances of $M = 1$, $C = 4$, $\varepsilon = 0.25$ ($\sigma^2_{TOT} = 5.25$), the variance of the mean decreases with the number of replicates at each layer.

same homogenized tissue, gives us the average of all cells. There is essentially no biological variation in these measurements because n in the s.e.m. term is very large—the only variability that we are likely to find is due to measurement error. We must not confuse the reproducibility of the tissue average with response of individual cells, which can be quite variable.

Blocking[3] on a noise variable allows us to remove a noise effect by taking a difference of two measurements that share the same value of the noise (e.g., same sample before and after treatment). Blocking enhances external validity—within the block, variability is controlled as tightly as possible for internal validity. The blocks themselves are chosen to cover the range of variability needed to estimate the response variability in the population of interest (**Fig. 1c**). This is the approach taken by the paired t-test, in which the block is a subject. For another example, a heterogeneous tissue could not be homogenized and a block would be defined by a spatial boundary between different cells. Neglecting to account for this would disregard the block boundaries in **Figure 1c** and would reduce sensitivity.

There can also be multiple sources of technical variability, such as reagents, measurement platforms and personnel. The same principles apply as for biological inference, measures of technical variability are seldom of interest—the usual objective is to minimize it. Blocking may still be used to eliminate known sources of noise—for example, collaborating labs may each do one complete replicate of an experiment to provide sufficient replication while eliminating any variability due to lab effects in the treatment comparisons.

Consider an experiment that assesses the effect of a drug on the livers of male mice of a specific genotype, at both the animal and cell layers. If the drug is administered *in vivo*, the animal is euthanized and the response measured on many cells, animals exposed to the drug cannot be their own controls. So, we expect variability at both the mouse layer and the cell (within mouse) layer. As well, we expect variability due to cell culture and maternal effects.

In the simplest experiment, we have a nested design, with mice selected at random for the treatment and the control. After dissection, cells are sampled from each liver, and their response to the drug is measured. The total variation of the measurement is the sum of variances of each effect, weighted by the number of times the effect was independently sampled (**Fig. 2b**). Using the same variances as above and $(n_M, n_C, n_\varepsilon) = (10, 5, 3)$ we find $\mathrm{Var}(\bar{X}) = 1/10 + 4/50 + 0.25/150 = 0.18$. The variance of the difference in the means of two measurements (e.g., reference and drug) will be twice this, 0.36, and our power to detect an effect of $d = 1.5$ is 0.65 (**Supplementary Note**).

Suppose that we discover that the mouse variation, $M = 1$, has significant components from maternal and cell culture effects, given by variances M_{MAT} and M_{CELL}. In this context, we can partition $M = M_{MAT} + M_{CELL} + M_0$, where M_0 is the unique variance not attributable to maternal or cell culture effects. We can attempt to control maternal effects by using sibling pairs (a block) and subjecting one mouse from each pair to the drug and one to the control. As the pairs have the same mother, the maternal effects cancel. Similarly, variance due to cell culture effects can be minimized by concurrently euthanizing each sibling pair (another block) and jointly preparing the cell cultures.

Having blocked these two effects, although M_{MAT} and M_{CELL} still contribute to the variance for both control and drug, we have effectively removed them from the variance of the difference in means. If these effects account for half of the mouse variance, $M_{MAT} + M_{CELL} = M/2 = 0.5$ (using $M = 1$ as above), blocking reduces the variance in the difference by $2(M_{MAT} + M_{CELL})/10$ from 0.36 to 0.26 and increases our power to 0.79 (**Supplementary Note**).

We can use the concept of effective sample size, $n = \mathrm{Var}(X)/\mathrm{Var}(\bar{X})$, to demonstrate the effect of this blocking. In the nested replication design, n is typically smaller than the total number of measurements ($n_M \times n_C \times n_\varepsilon$) because we do not independently sample each source of variation in each measurement[2] (it is largest for $n_C = n_\varepsilon = 1$). As a result, replication at the cell and technical layers decreases $\mathrm{Var}(\bar{X})$ proportionally more slowly than replication at the topmost mouse layer. When both maternal and cell culture effects are included, $\mathrm{Var}(X) = M + C + \varepsilon = 5.25$ and the effective sample size is $n = 5.25/0.36 = 15$. When maternal and cell effects are blocked, $\mathrm{Var}(X)$ remains the same, but now $\mathrm{Var}(\bar{X})$ is reduced to 0.26 and $n = 5.25/0.26 = 20$.

Given the choice, we should always block at the top layer because the noise in this layer is independently sampled the fewest times. We can use the effective sample size n to illustrate this. Blocking at mouse layer decreased M from 1 to 0.5 (by 50%) and increased n from 15 to 20 (power from 0.65 to 0.79). In contrast, a proportional reduction in C from 4 to 2 increases n to 19 (power to 0.76), whereas a reduction in ε has essentially no effect on n.

We need to distinguish between sources of variation that are nuisance factors in our goal to measure mean biological effects from those that are required to assess how much effects vary in the population. Whereas the former should be minimized to optimize the power of the experiment, the latter need to be sampled and quantified so that we can both generalize our conclusions and robustly determine the uncertainty in our estimates.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3224).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

1. Krzywinski, M. & Altman, *Nat. Methods* **11**, 597–598 (2014).
2. Blainey, P., Krzywinski, M. & Altman, *Nat. Methods* **11**, 879–880 (2014).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 809–810 (2013).

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.

POINTS OF SIGNIFICANCE

Split plot design

When some factors are harder to vary than others, a split plot design can be efficient.

We have already seen that varying two factors simultaneously provides an effective experimental design for exploring the main (average) effects and interactions of the factors[1]. However, in practice, some factors may be more difficult to vary than others at the level of experimental units. For example, drugs given orally are difficult to administer to individual tissues, but observations on different tissues may be done by biopsy or autopsy. When the factors can be nested, it is more efficient to apply a difficult-to-change factor to the units at the top of the hierarchy and then apply the easier-to-change factor to a nested unit. This is called a split plot design.

The term "split plot" derives from agriculture, where fields may be split into plots and subplots. It is instructive to review completely randomized design (CRD) and randomized complete block design (RCBD)[2] and show how these relate to split plot design. Suppose we are studying the effect of irrigation amount and fertilizer type on crop yield. We have access to eight fields, which can be treated independently and without proximity effects (**Fig. 1a**). If applying irrigation and fertilizer is equally easy, we can use a complete 2 × 2 factorial design and assign levels of both factors randomly to fields in a balanced way (each combination of factor levels is equally represented).

If our land is divided into two large fields that may differ in some way, we can use the field as a blocking factor (**Fig. 1b**). Within each block, we again perform a complete 2 × 2 factorial design: irrigation and fertilizer are assigned to each of the four smaller fields within the large field, leading to an RCBD with field as the block. Each combination of irrigation and fertilizer is balanced within the large field.

So far, we have not considered whether managing levels of irrigation and fertilizer require the same effort. If varying irrigation on a small scale is difficult, it makes more sense to irrigate larger areas of land than in **Figure 1a** and then vary the fertilizer accordingly to maintain a balanced design. If our land is divided into four fields (whole plots), each of which can be split into two subplots (**Fig. 1c**), we would assign irrigation to whole plots using CRD. Within a whole plot, fertilizer would be distributed across subplots using RCBD,

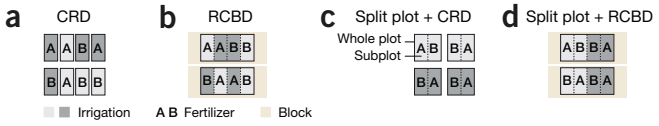

Figure 1 | Split plot design examples from agriculture. (**a**) In CRD, levels of irrigation and fertilizer are assigned to plots of land (experimental units) in a random and balanced fashion. (**b**) In RCBD, similar experimental units are grouped (for example, by field) into blocks and treatments are distributed in a CRD fashion within the block. (**c**) If irrigation is more difficult to vary on a small scale and fields are large enough to be split, a split plot design becomes appropriate. Irrigation levels are assigned to whole plots by CRD and fertilizer is assigned to subplots using RCBD (irrigation is the block). (**d**) If the fields are large enough, they can be used as blocks for two levels of irrigation. Each field is composed of two whole plots, each composed of two subplots. Irrigation is assigned to whole plots using RCBD (blocked by field) and fertilizer assigned to subplots using RCBD (blocked by irrigation).

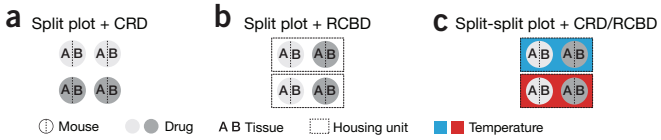

Figure 2 | In biological experiments using split plot designs, whole plot experimental units can be individual animals or groups. (**a**) A two-factor, split plot animal experiment design. The whole plot is represented by a mouse assigned to drug, and tissues represent subplots. (**b**) Biological variability coming from nuisance factors, such as weight, can be addressed by blocking the whole plot factor, whose levels are now sampled using RCBD. (**c**) With three factors, the design is split-split plot. The housing unit is the whole plot experimental unit, each subject to a different temperature. Temperature is assigned to housing using CRD. Within each whole plot, the design shown in **b** is performed. Drug and tissue are subplot and sub-subplot units. Replication is done by increasing the number of housing units.

randomly and balanced within whole plots with a given irrigation level. Irrigation is the whole plot factor and fertilizer is the subplot factor. It is important to note that all split plot experiments include at least one RCBD subexperiment, with the whole plot factor acting as a block.

Assigning levels of irrigation to fields at random neglects any heterogeneity among the fields. For example, if the land is divided into two large fields (**Fig. 1b**), it is best to consider each as a block. Within each block, we consider half of the field as a whole plot and irrigate using RCBD (**Fig. 1d**). As before, the fertilizer is assigned to subplots using RCBD. The designs in **Figure 1c** and **Figure 1d** vary only in how the whole plot factor levels are assigned: by CRD or RCBD.

Because split plot designs are based on RCBD, the two can be easily confused. For example, why is **Figure 1b** not considered a split plot design with field index being the whole plot factor? The answer involves whether we are interested in specific levels of the factor or are using it for blocking purposes. In **Figure 1b**, the field is a blocking factor because it is used to control the variability of the plots, not as a systematic effect. We use these two fields to generalize to all fields. In **Figure 1c**, irrigation is a whole plot factor and not a blocking factor because we are studying the specific levels of irrigation.

The terms "whole plot" and "subplot" translate naturally from agricultural to biological context, where split plot designs are common. Many factors, such as diet or housing conditions, are more easily applied to large groups of experimental subjects, making them suitable at the whole plot level. In other experiments, factors that are sampled hierarchically or from the same individual (tissue, cell or time points) can act as subplot factors. **Figure 2** illustrates split plot designs in a biological context.

Suppose that we wish to determine the *in vivo* effect of a drug on gene expression in two tissues. We assign mice to one of two drug treatments using CRD. The mouse is the whole plot experimental unit and the drug is the whole plot factor. Both tissues are sampled from each mouse. The tissue is the subplot factor and each mouse acts as a block for the tissue subplot factor; this is the RCBD component (**Fig. 2a**). The mouse itself can be considered a random factor used to sample biological variability and increase the external validity of the experiment. If we suspect environmental variability, we can group the mice by their housing unit (**Fig. 2b**), just as we did whole plots by field (**Fig. 1d**). The housing unit is now a blocking factor for the drug, which is applied to mice using RCBD. Other ways to group mice might be by weight, familial relationship or genotype.

Sensitivity in detecting effects of the subplot factor as well as interactions is generally greater than for a corresponding completely

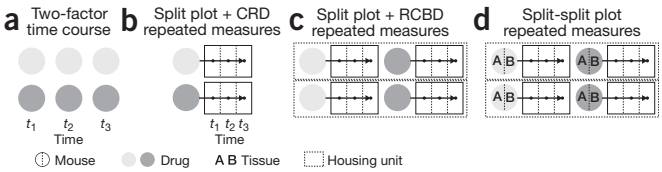

Figure 3 | The split plot design with CRD is commonly applied to a repeated measures time course design. (**a**) Basic time course design, in which time is one of the factors. Each measurement uses a different mouse. (**b**) In a repeated measures design, mice are followed longitudinally. Drug is assigned to mice using CRD. Time is the subplot factor. (**c**) Drug is blocked by housing. (**d**) A three-factor, repeated measures split-split plot design, now including tissue. Tissue is subplot and time is sub-subplot.

Table 1 | Split plot ANOVA table for two-factor split plot designs

	CRD			RCBD		
	d.f.	MS	*F*-ratio	d.f.	MS	*F*-ratio
Block, *bl*				n'	MS_{bl}	MS_{bl}/MS_{wp}
A	a'	MS_A	MS_A/MS_{wp}	a'	MS_A	MS_A/MS_{wp}
Error wp	an'	MS_{wp}		$n'a'$	MS_{wp}	
B	b'	MS_B	MS_B/MS_{sp}	b'	MS_B	MS_B/MS_{sp}
$A \times B$	$a'b'$	$MS_{A\times B}$	MS_{AB}/MS_{sp}	$a'b'$	$MS_{A\times B}$	$MS_{A\times B}/MS_{sp}$
Error sp	$ab'n'$	MS_{sp}		$ab'n'$	MS_{sp}	
Total	$abn - 1$			$abn - 1$		

Split plot ANOVA table for two factor split plot designs using CRD (**Fig. 1c**) and RCBD (**Fig. 1d**) with a levels of whole plot factor A and b levels of subplot factor B. For CRD n is measurements per subplot and for RCBD n is number of blocks. Whole plot and subplot errors are indicated by wp and sp subscripts, respectively. For RCBD, interaction between blocking factor *bl* and B is usually included in the subplot error term. $a' = a - 1$, $b' = b - 1$, $n' = n - 1$. d.f., degrees of freedom; *F*-ratio, test statistic for *F* test.

randomized factorial design in which only one tissue is measured in each mouse. This is because tissue comparisons are within mouse. However, because comparing the whole plot factor (drug) is done between subjects, the sensitivity for the whole plot factor is similar to that of a completely randomized design. Applying blocking at the whole plot level, such as housing (**Fig. 2b**), can improve sensitivity for the whole plot factor similarly to using a RCBD. Compared to a split plot design, the completely randomized design is both more expensive (twice as many mice are required) and less efficient (mouse variability will not cancel, and thus the tissue and interaction effects will include mouse-to-mouse variability).

The experimental unit at the whole plot level does not have to correspond to an individual. It can be one level above the individual in the hierarchy, such as a group or enclosure. For example, suppose we are interested in adding temperature as one of the factors to the study in **Figure 2b**. Since it is more practical to control the temperature of the housing unit than of individual mice, we use cage as the whole plot (**Fig. 2c**). Temperature is the whole plot factor and cage is the experimental unit at the whole plot level. As in **Figure 2a**, we use CRD to assign the whole plot factor (temperature) levels to whole plots (cages). Mice are now experimental units at the subplot level and the drug is now a subplot factor. Because we have three layers in the hierarchy of factors, tissue is at the sub-subplot level and the design is split-split plot. In **Figure 2b**, the cage is a block used to control variability because the effects of housing are not of specific interest to us. By contrast, in **Figure 2c**, specific levels of the temperature factor are of interest so it is part of the plot factor hierarchy.

Care must be taken to not mistake a split plot design for CRD. For example, an inadvertent split plot[3] can result if some factor levels are not changed between experiments. If the analysis treats all experiments as independent, then we can expect mistakes in conclusions about the significance of effects.

With two factors, more complicated designs are also possible. For example, we might expose the whole mouse to a drug (factor *A*) *in vivo* and then expose two liver samples to different *in vitro* treatments (factor *B*). In this case, the two liver samples from the same mouse form a block, which is nested in mouse[4].

The split plot CRD design (**Fig. 2a**) is commonly used as the basis for a repeated measures design, which is a type of time course design. The most basic time course includes time as one of the factors in a two-factor design. In a completely randomized time course experiment, different mice are used at each of the measurement times t_1, t_2 and t_3 after initial treatment (**Fig. 3a**). If the same mouse is used at each time and the mice are assigned at random to the levels of a (time-invariant) factor, the design becomes a repeated measures design (**Fig. 3b**) because the measurements are nested within mouse. The time of measurement is the subplot factor. The corresponding repeated measures of the design that uses housing as a block in **Figure 2b** is shown in **Figure 3c**. As before, housing is the block and drug is the whole plot factor, but now time is the subplot factor. If we include tissue type, the design becomes a split-split plot, with tissue being subplot and time sub-subplot (**Fig. 3d**).

Split plot designs are analyzed using ANOVA. Because comparisons at the whole plot level have different variability than those at the subplot level, the ANOVA table contains two sources of error, MS_{wp} and MS_{sp}, the mean square associated with whole plots and subplots, respectively (**Table 1**). This difference occurs because the subplot factor is always compared within a block, while the whole plot factor is compared between the whole plots. For example, in **Figure 2a**, variation between mice cancels out when comparing tissues but not when comparing drugs. Analogously to a two-factor ANOVA[1], we calculate the sums of squares and mean squares in a split plot ANOVA. For example, in a split plot with RCBD, given n blocks of blocking factor *bl* (**Table 1**) at the whole plot level and a and b levels of whole plot factor A and subplot factor B, $MS_{bl} = SS_{bl}/(n-1)$, where SS_{bl} is the sum of squared deviations of the average across each block relative to the grand mean times the number of measurements contributing to each average ($a \times b$). Similarly, SS_A uses the average across levels of A and the multiple is $n \times b$. The analysis at the whole plot level is essentially the same as in a one-way ANOVA with blocking: the subplot values are considered subsamples. The associated MS_{sp} is usually lower than in a factorial design, which improves the sensitivity in detecting $A \times B$ interactions.

Split plot designs are helpful when it is difficult to vary all factors simultaneously, and, if factors that require more time or resources can be identified, split plot designs can offer cost savings. This type of design is also useful for cases when the investigator wishes to expand the scope of the experiment: a factor can be added at the whole plot level without sacrificing sensitivity in the subplot factor.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

1. Krzywinski, M., Altman, N. & Blainey, P. *Nat. Methods* **11**, 1187–1188 (2014).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
3. Ganju, J. & Lucas, J.M. *J. Stat. Plan. Infer.* **81**, 129–140 (1999).
4. Krzywinski, M., Altman, N. & Blainey, P. *Nat. Methods* **11**, 977–978 (2014).

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.

POINTS OF SIGNIFICANCE

Bayes' theorem

Incorporate new evidence to update prior information.

Observing, gathering knowledge and making predictions are the foundations of the scientific process. The accuracy of our predictions depends on the quality of our present knowledge and accuracy of our observations. Weather forecasts are a familiar example—the more we know about how weather works, the better we can use current observations and seasonal records to predict whether it will rain tomorrow and any disagreement between prediction and observation can be used to refine the weather model. Bayesian statistics embodies this cycle of applying previous theoretical and empirical knowledge to formulate hypotheses, rank them on the basis of observed data and update prior probability estimates and hypotheses using observed data[1]. This will be our first of a series of columns about Bayesian statistics. This month, we'll introduce the topic using one of its key concepts—Bayes' theorem—and expand to include topics such as Bayesian inference and networks in future columns.

Bayesian statistics is often contrasted with classical (frequentist) statistics, which assumes that observed phenomena are generated by an unknown but fixed process. Importantly, classical statistics assumes that population parameters are unknown constants, given that complete and exact knowledge about the sample space is not available[2]. For estimation of population characteristics, the concept of probability is used to describe the outcomes of measurements.

In contrast, Bayesian statistics assumes that population parameters, though unknown, are quantifiable random variables and that our uncertainty about them can be described by probability distributions. We make subjective probability statements, or 'priors', about these parameters based on our experience and reasoning about the population. Probability is understood from this perspective as a degree of belief about the values of the parameter under study. Once we collect data, we combine them with the prior to create a distribution called the 'posterior' that represents our updated information about the parameters, as a probability assessment about the possible values of the parameter. Given that experience, knowledge, and reasoning process vary among individuals, so do their priors—making specification of the prior one of the most controversial topics in Bayesian statistics. However, the influence of the prior is usually diminished as we gather knowledge and make observations.

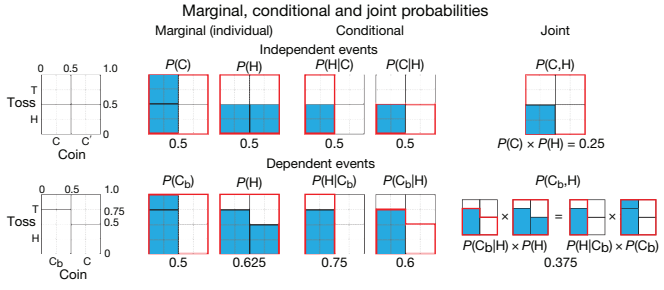

Figure 1 | Marginal, joint and conditional probabilities for independent and dependent events. Probabilities are shown by plots[3], where columns correspond to coins and stacked bars within a column to coin toss outcomes, and are given by the ratio of the blue area to the area of the red outline. The choice of one of two fair coins (C, C′) and outcome of a toss are independent events. For independent events, marginal and conditional probabilities are the same and joint probabilities are calculated using the product of probabilities. If one of the coins, C_b, is biased (yields heads (H) 75% of the time), the events are dependent, and joint probability is calculated using conditional probabilities.

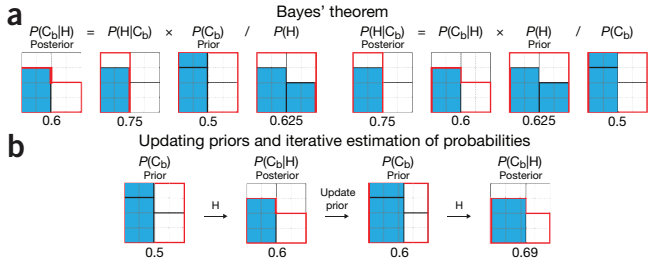

Figure 2 | Graphical interpretation of Bayes' theorem and its application to iterative estimation of probabilities. (**a**) Relationship between conditional probabilities given by Bayes' theorem relating the probability of a hypothesis that the coin is biased, $P(C_b)$, to its probability once the data have been observed, $P(C_b|H)$. (**b**) The probability of the identity of the chosen coin can be inferred from the toss outcome. Observing a head increases the chances that the coin is biased from $P(C_b) = 0.5$ to 0.6, and further to 0.69 if a second head is observed.

At the core of Bayesian statistics is Bayes' theorem, which describes the outcome probabilities of related (dependent) events using the concept of conditional probability. To illustrate these concepts, we'll start with independent events—tossing one of two fair coins, C and C′. The toss outcome probability does not depend on the choice of coin—the probability of heads is always the same, $P(H) = 0.5$ (**Fig. 1**). The joint probability of choosing a given coin (e.g., C) and toss outcome (e.g., H) is simply the product of their individual probabilities, $P(C, H) = P(C) \times P(H)$. But if we were to replace one of the coins with a biased coin, C_b, that yields heads 75% of the time, the choice of coin would affect the toss outcome probability, making the events dependent. We express this using conditional probabilities by $P(H|C) = 0.5$ and $P(H|C_b) = 0.75$, where "|" means "given" or "conditional upon" (**Fig. 1**).

If $P(H|C_b)$ is the probability of observing heads given the biased coin, how can we calculate $P(C_b|H)$, the probability that the coin is biased having observed heads? These two conditional probabilities are generally not the same—failing to distinguish them is known as the prosecutor's fallacy. $P(H|C_b)$ is a property of the biased coin and, unlike $P(C_b|H)$, is unaffected by the chance of the coin being biased.

We can relate these conditional probabilities by first writing the joint probability of selecting C_b and observing H: $P(C_b, H) = P(C_b|H) \times P(H)$ (**Fig. 1**). The fact that this is symmetric, $P(C_b|H) \times P(H) = P(H|C_b) \times P(C_b)$, leads us to Bayes' theorem, which is a rearrangement of this equality: $P(C_b|H) = P(H|C_b) \times P(C_b)/P(H)$ (**Fig. 2a**). $P(C_b)$ is our guess of the coin being biased before data are collected (the prior), and $P(C_b|H)$ is our guess once we have observed heads (the posterior).

If both coins are equally likely to be picked, $P(C_b) = P(C) = 0.5$. We also know that $P(H|C_b) = 0.75$, which is a property of the biased coin. To apply Bayes' theorem, we need to calculate $P(H)$, which is the probability of all the ways of observing heads—picking the fair coin and observing heads and picking the biased coin and observing heads. This is $P(H) = P(H|C) \times P(C) + P(H|C_b) \times P(C_b) = 0.5 \times 0.5 + 0.75 \times 0.5 = 0.625$. By substituting these values in Bayes' theorem, we can compute the probability that the coin is biased

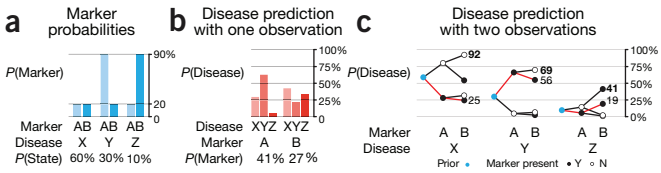

Figure 3 | Disease predictions based on presence of markers. (**a**) Independent conditional probabilities of observing each marker (A, B) given a disease (X, Y, Z) (e.g., $P(A|Y) = 0.9$). (**b**) Posterior probability of each disease given a single observation that confirms the presence of one of the markers (e.g., $P(Y|A) = 0.66$). (**c**) Evolution of disease probability predictions with multiple assays. For a given disease, each path traces (left to right) the value of the posterior that incorporates all the assay results up to that point, beginning at the prior probability for the disease (blue dot). The assay result is encoded by an empty (marker absent) or a solid (marker present) dot. The red path corresponds to presence of A and B. The highest possible posterior is shown in bold.

after observing a head, $P(C_b|H) = P(H|C_b) \times P(C_b)/P(H) = 0.75 \times 0.5/0.625 = 0.6$ (**Fig. 2a**).

Bayes' theorem can be applied to such inverse probability problems iteratively—when we need to update probabilities step by step as we gain evidence. For example, if we toss the coin a second time, we can update our prediction that the coin is biased. On the second toss we no longer use $P(C_b) = 0.5$ because the first toss suggested that the biased coin is more likely to be picked. The posterior from the first toss becomes the new prior, $P(C_b) = 0.6$. If the second toss yields heads, we compute $P(H) = 0.5 \times 0.4 + 0.75 \times 0.6 = 0.65$ and apply Bayes' theorem again to find $P(C_b|HH) = 0.75 \times 0.6/0.65 = 0.69$ (**Fig. 2b**). We can continue tossing to further refine our guess—each time we observe a head, the assessment of the posterior probability that the coin is biased is increased. For example, if we see four heads in a row, there is an 84% posterior probability that the coin is biased (see **Supplementary Table 1**).

We have computed the probability that the coin is biased given that we observed two heads. Up to this point we have not performed any statistical inference because all the probabilities have been specified. Both Bayesians and frequentists agree that $P(C_b|HH) = 0.69$ and $P(HH|C) = 0.25$. Statistical inference arises when there is an unknown, such as $P(H|C_b)$. The difference between frequentist and Bayesian inference will be discussed more fully in the next column.

Let's extend the simple coin example to include multiple event outcomes. Suppose a patient has one of three diseases (X, Y, Z) whose prevalence is 0.6, 0.3 or 0.1, respectively—X is relatively common, whereas Z is rare. We have access to a diagnostic test that measures the presence of protein markers (A, B). Both markers can be present, and the probabilities of observing a given marker for each disease are known and independent of each other in each disease state (**Fig. 3a**). We can ask: if we see marker A, can we predict the state of the patient? Also, how do our predictions change if we subsequently assay for B?

Let's first calculate the probability that the patient has disease X given that marker A was observed: $P(X|A) = P(A|X) \times P(X)/P(A)$. We know the prior probability for X, which is the prevalence $P(X) = 0.6$, and the probability of observing A given X, $P(A|X) = 0.2$ (**Fig. 3a**). To apply Bayes' theorem we need to calculate $P(A)$, which is the total probability of observing A regardless of the state of the patient. To find $P(A)$ we sum over the product of the probability of each disease and finding A in that disease, which is all the ways in which A can be observed: $P(A) = 0.6 \times 0.2 + 0.3 \times 0.9 + 0.1 \times 0.2 = 0.41$ (**Fig. 3b**). Bayes' theorem gives us $P(X|A) = 0.2 \times 0.6/0.41 = 0.29$. Because marker A is more common in another disease, Y, this new estimate that the patient has disease X is much lower than the original of 0.6. Similarly, we can calculate the posteriors for Y and Z as $P(Y|A) = 0.66$ and $P(Z|A) = 0.05$ (see **Supplementary Table 1**). With a single assay that confirms A, it is most likely (66%) that the patient has disease Y.

Instead, if we confirm B is present, the probabilities of X, Y and Z are 44%, 22% and 33%, respectively (**Fig. 3b**), and our best guess is that the patient has X. Even though marker B is nearly always present in disease Z—$P(B|Z) = 0.9$—detecting it raises the probability of Z only to $P(Z|B) = 0.33$, which is still lower than the probability of X. The reason for this is that Z itself is rare, and observing B is also possible for the more common diseases X and Y. This phenomenon is captured by Carl Sagan's words: "extraordinary claims require extraordinary evidence." In this case, observing B is not "extraordinary" enough to significantly advance our claim that the patient has disease Z. Even if B were always present in Z, i.e., $P(B|Z) = 1$, and present in X and Y at only 1%, $P(B|X) = P(B|Y) = 0.01$, observing B would only allow us to say that there is a 92% chance that the patient has Z. If we failed to account for different prevalence rates, we would grossly overestimate the chances that the patient has Z. For example, if instead we supposed that all three diseases are equally likely, $P(X) = P(Y) = P(Z) = 1/3$, observing B would lead us to believe that the chances of Z are 69%.

Having observed A, we could refine our predictions by testing for B. As with the coin example, we use the posterior probability of the disease after observing A as the new prior. The posterior probabilities for diseases X, Y and Z given that A and B are both present are 0.25, 0.56 and 0.19, respectively, making Y the most likely. If the assay for B is negative, the calculations are identical but use complementary probabilities (e.g., $P(\text{not } B|X) = 1 - P(B|X)$) and find 0.31, 0.69 and 0.01 as the probabilities for X, Y and Z. Observing A but not B greatly decreases the chances of disease Z, from 19% to 1%. **Figure 3c** traces the change in posterior probabilities for each disease with each possible outcome as we assay both markers in turn. If we find neither A nor B, there is a 92% probability that the patient has disease X—the marker profile with the highest probability for predicting X. The most specific profile for Y is A^+B^- (69%) and for Z is A^-B^+ (41%).

When event outcomes map naturally onto conditional probabilities, Bayes' theorem provides an intuitive method of reasoning and convenient computation. It allows us to combine prior knowledge with observations to make predictions about the phenomenon under study. In Bayesian inference, all unknowns in a system are modeled by probability distributions that are updated using Bayes' theorem as evidence accumulates. We will examine Bayesian inference and compare it with frequentist inference in our next discussion.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3335).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Jorge López Puga, Martin Krzywinski & Naomi Altman

1. Eddy, S.R. *Nat. Biotechnol.* **22**, 1177–1178 (2004).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 809–810 (2013).
3. Oldford, R.W. & Cherry, W.H. Picturing probability: the poverty of Venn diagrams, the richness of eikosograms. http://sas.uwaterloo.ca/~rwoldfor/papers/venn/eikosograms/paperpdf.pdf (University of Waterloo, 2006)

Jorge López Puga is a Professor of Research Methodology at Universidad Católica de Murcia (UCAM). Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.

POINTS OF SIGNIFICANCE

Bayesian statistics

Today's predictions are tomorrow's priors.

One of the goals of statistics is to make inferences about population parameters from a limited set of observations. Last month, we showed how Bayes' theorem is used to update probability estimates as more data are collected[1]. We used the example of identifying a coin as fair or biased based on the outcome of one or more tosses. This month, we introduce Bayesian inference by treating the degree of bias as a population parameter and using toss outcomes to model it as a distribution to make probabilistic statements about its likely values.

How are Bayesian and frequentist inference different? Consider a coin that yields heads with a probability of π. Both the Bayesian and the frequentist consider π to be a fixed but unknown constant and compute the probability of a given set of tosses (for example, k heads, H^k) based on this value (for example, $P(H^k \mid \pi) = \pi^k$), which is called the likelihood. The frequentist calculates the probability of different data generated by the model, $P(\text{data} \mid \text{model})$, assuming a probabilistic model with known and fixed parameters (for example, coin is fair, $P(H^k) = 0.5^k$). The observed data are assessed in light of other data generated by the same model.

In contrast, the Bayesian uses probability to quantify uncertainty and can make more precise probability statements about the state of the system by calculating $P(\text{model} \mid \text{data})$, a quantity that is meaningless in frequentist statistics. The Bayesian uses the same likelihood as the frequentist, but also assumes a probabilistic model (prior distribution) for possible values of π based on previous experience. After observing the data, the prior is updated to the posterior, which is used for inference. The data are considered fixed and possible models are assessed on the basis of the posterior.

Let's extend our coin example from last month to incorporate inference and illustrate the differences in frequentist and Bayesian approaches to it. Recall that we had two coins: coin C was fair, $P(H \mid C) = \pi_0 = 0.5$, and coin C_b was biased toward heads, $P(H \mid C_b) = \pi_b = 0.75$. A coin was selected at random with equal probability and tossed. We used Bayes' theorem to compute the probability that the biased coin was selected given that a head was observed; we found $P(C_b \mid H) = 0.6$. We also saw how we could refine our guess by updating this probability with the outcome of another toss: seeing a second head gave us $P(C_b \mid H^2) = 0.69$.

In this example, the parameter π is discrete and has two possible values: fair ($\pi_0 = 0.5$) and biased ($\pi_b = 0.75$). The prior probability of each before tossing is equal, $P(\pi_0) = P(\pi_b) = 0.5$, and the data-generating process has the likelihood $P(H^k \mid \pi) = \pi^k$. If we observe a head, Bayes' theorem gives the posterior probabilities as $P(\pi_0 \mid H) = \pi_0/(\pi_0 + \pi_b) = 0.4$ and $P(\pi_b \mid H) = \pi_b/(\pi_0 + \pi_b) = 0.6$. Here all the probabilities are known and the frequentist and Bayesian agree on the approach and the results of computation.

In a more realistic inference scenario, nothing is known about the coin and π could be any value in the interval [0,1]. What can be inferred about π after a coin toss produces H^3 (where H^kT^{n-k} denotes the outcome of n tosses that produced k heads and $n-k$ tails)? The frequentist and the Bayesian agree on the data generation model $P(H^3 \mid \pi) = \pi^3$, but they will use different methods to encode experience from other coins and the observed outcomes.

In part, this compatibility arises because, for the frequentist, only the data have a probability distribution. The frequentist may test whether the coin is fair using the null hypothesis, H_0: $\pi = \pi_0 = 0.5$. In this case, H^3 and T^3 are the most extreme outcomes, each with probability 0.125. The P value is therefore $P(H^3 \mid \pi_0) + P(T^3 \mid \pi_0) = 0.25$. At the nominal level of $\alpha = 0.05$, the frequentist fails to reject H_0 and accepts that $\pi = 0.5$. The frequentist might estimate π using the sample percentage of heads or compute a 95% confidence interval for π, $0.29 < \pi \leq 1$. The interval depends on the outcome, but 95% of the intervals will include the true value of π.

The frequentist approach can only tell us the probability of obtaining our data under the assumption that the null hypothesis is the true data-generating distribution. Because it considers π to be fixed, it does not recognize the legitimacy of questions like "What is the probability that the coin is biased towards heads?" The coin either is or is not biased toward heads. For the frequentist, probabilistic questions about π make sense only when selecting a coin by a known randomization mechanism from a population of coins.

By contrast, the Bayesian, while agreeing that π has a fixed true value for the coin, quantifies uncertainty about the true value as a probability distribution on the possible values called the prior distribution. For example, if she knows nothing about the coin, she could use a uniform distribution on [0,1] that captures her assessment that any value of π is equally likely (**Fig. 1a**). If she thinks that the coin is most likely to be close to fair, she can pick a bell-shaped prior distribution (**Fig. 1a**). These distributions can be imagined as the histogram of the values of π from a large population of coins from which the current coin was selected at random. However, in the Bayesian model, the investigator chooses the prior based on her knowledge about the coin at hand, not some imaginary set of coins.

Given the toss outcome of H^3, the Bayesian applies Bayes' theorem to combine the prior, $P(\pi)$, with the likelihood of observing the data, $P(H^3 \mid \pi)$, to obtain the posterior $P(\pi \mid H^3) = P(H^3 \mid \pi) \times P(\pi) / P(H^3)$ (**Fig. 1b**). This is analogous to $P(A \mid B) = P(B \mid A) \times P(A)/P(B)$, except now A is the model parameter, B is the observed data and, because π is continuous $P(\cdot)$ is interpreted as a probability density. The term corresponding to the denominator $P(B)$, the marginal likelihood $P(H^3)$, becomes the normalizing constant so that the total probability (area under the curve) is 1. As long as this is finite, it is often left out and the numerator is used to express the shape of density. That is the reason why it is commonly said that posterior distribution is proportional to the prior times the likelihood.

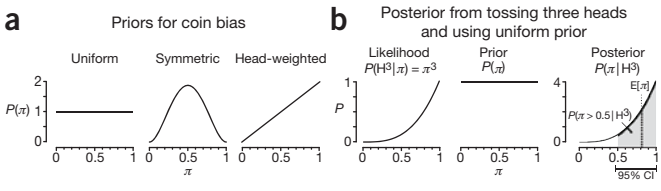

Figure 1 | Prior probability distributions represent knowledge about the coin before it is tossed. (**a**) Three different prior distributions of π, the probability of heads. (**b**) Toss outcomes are combined with the prior to create the posterior distribution used to make inferences about the coin. The likelihood is the probability of observing a given toss outcome, which is π^3 for a toss of H^3. The gray area corresponds to the probability that the coin is biased toward heads. The error bar is the 95% credible interval (CI) for π. The dotted line is the posterior mean, $E(\pi)$. The posterior is shown normalized to $4\pi^3$ to make its area 1.

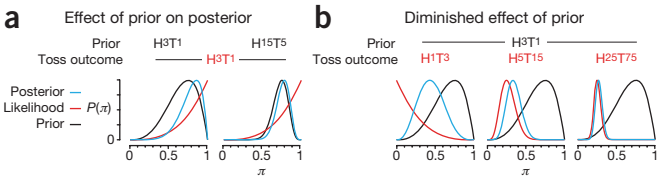

Figure 2 | Effect of choice of prior and amount of data collected on the posterior. All curves are beta(a,b) distributions labeled by their equivalent toss outcome, $H^{a-1}T^{b-1}$. (**a**) Posteriors for a toss outcome of H^3T^1 using weakly (H^3T^1) and strongly ($H^{15}T^5$) head-weighted priors. (**b**) The effect of a head-weighted prior, H^3T^1, diminishes with more tosses (4, 20, 100) indicative of a tail-weighted coin (75% tails).

Suppose the Bayesian knows little about the coin and uses the uniform prior, $P(\pi) = 1$. The relationship between posterior and likelihood is simplified to $P(\pi \mid H^3) = P(H^3 \mid \pi) = \pi^3$ (**Fig. 1b**). The Bayesian uses the posterior distribution for inference, choosing the posterior mean ($\pi = 0.8$), median ($\pi = 0.84$) or value of π for which posterior is maximum ($\pi = 1$, mode) for a point estimate of π.

The Bayesian can also calculate 95% credible region, the smallest interval over which we find 95% of the area under the posterior—which is [0.47,1] (**Fig. 1b**). Like the frequentist, the Bayesian cannot conclude that the coin is not biased, because $\pi = 0.5$ falls within the credible interval. Unlike the frequentist, they can make statements about the probability that the coin is biased toward heads (94%) using the area under the posterior distribution for $\pi > 0.5$ (**Fig. 1b**). The probability that the coin is biased toward tails is $P(\pi < 0.5 \mid H^3) = 0.06$. Thus, given the choice of prior, the toss outcome H^3 overwhelmingly supports the hypothesis of head bias, which is 0.94/0.06 = 16 times more likely than tail bias. This ratio of posterior probabilities is called the Bayes factor and its magnitude can be associated with degree of confidence[2]. By contrast, the frequentist would test $H_0\ \pi_0 \leq 0.5$ versus $H_A\ \pi_0 > 0.5$ using the P value based on a one-tailed test at the boundary ($\pi_0 = 0.5$) and obtain $P = 0.125$ and would not reject the null hypothesis. Conversely, the Bayesian cannot test the hypothesis that the coin is fair because, in using the uniform prior, statements about P are limited to intervals and cannot be made for single values of π (which always have zero prior and posterior probabilities).

Suppose now that we suspect the coin to be head-biased and want a head-weighted prior (**Fig. 1a**). What would be a justifiable shape? It turns out that if we consider the general case of n tosses with outcome H^kT^{n-k}, we arrive at a tidy solution. With a uniform prior, this outcome has a posterior probability proportional to $\pi^k(1-\pi)^{n-k}$. The shape and interpretation of the prior is motivated by considering n' more tosses that produce k' heads, $H^{k'}T^{n-k'}$. The combined toss outcome is $H^{k+k'}T^{(n+n')-(k+k')}$, which, with a uniform prior, has a posterior probability proportional to $\pi^{k+k'}(1-\pi)^{(n+n')-(k+k')}$. Another way to think about this posterior is to treat the first set of tosses as the prior, $\pi^k(1-\pi)^{n-k}$, and the second set as the likelihood, $\pi^{k'}(1-\pi)^{n-k'}$. In fact, if we extrapolate this pattern back to 0 tosses (with outcome H^0T^0), the original uniform prior is exactly the distribution that corresponds to this: $\pi^0(1-\pi)^0 = 1$. This iterative updating by adding powers treats the prior as a statement about the coin based on the outcomes of previous tosses.

Let's look how different shapes of priors might arise from this line of reasoning. Suppose we suspect that the coin is biased with $\pi = 0.75$. In a large number of tosses we expect to see 75% heads. If we are uncertain about this, we might let this imaginary outcome be H^3T^1 and set the prior proportional to $\pi^3(1-\pi)^1$ (**Fig. 2a**). If our suspicion is stronger, we might use $H^{15}T^5$ and set the prior proportional to $\pi^{15}(1-\pi)^5$. In either case, the posterior distribution is obtained simply by adding the number of observed heads and tails to the exponents of π and $(1-\pi)$, respectively. If our toss outcome is H^3T^1, the posteriors are proportional to $\pi^6(1-\pi)^2$ and $\pi^{18}(1-\pi)^6$.

As we collect data, the impact of the prior is diminished and the posterior is shaped more like the likelihood. For example, if we use a prior that corresponds to H^3T^1, suggesting that the coin is head-biased, and collect data that indicates otherwise and see tosses of H^1T^3, H^5T^{15} and $H^{25}T^{75}$ (75% tails), our original misjudgment about the coin is quickly mitigated (**Fig. 2b**).

In general, a distribution on π in [0,1] proportional to $\pi^{a-1}(1-\pi)^{b-1}$ is called a beta(a,b) distribution. The parameters a and b must be positive, but they do not need to be whole numbers. When $a \geq 1$ and $b \geq 1$, then $(a + b - 2)$ is like a generalized number of coin tosses and controls the tightness of the distribution around its mode (location of maximum of the density), and $(a - 1)$ is like the number of heads and controls the location of the mode.

All of the curves in **Figure 2** are beta distributions. Priors corresponding to a previous toss outcomes of H^kT^{n-k} are beta distributions with $a = k + 1$ and $b = n - k + 1$. For example, the prior for $H^{15}T^5$ has a shape of beta(16,6). For a prior of beta(a,b), a toss outcome of H^kT^{n-k} will have a posterior of beta($a + k$, $b + n - k$). For example, the posterior for a toss outcome of H^3T^1 using a $H^{15}T^5$ prior is beta(19,7).

In general, when the posterior comes from the same family of distributions as the prior with an update formula for the parameter, we say that the prior is conjugate to the distribution generating the data. Conjugate priors are convenient when they are available for data-generating models because the posterior is readily computed. The beta distributions are conjugate priors for binary outcomes such as H or T and come in a wide variety of shapes, flat, skewed, bell- or U- shaped. For a prior on the interval [0,1], it is usually possible to pick values of (a,b) for a suitable head probability prior for coin tosses (or the success probability for independent binary trials).

Frequentist inference assumes that the data-generating mechanism is fixed and that only the data have a probabilistic component. Inference about the model is therefore indirect, quantifying the agreement between the observed data and the data generated by a putative model (for example, the null hypothesis). Bayesian inference quantifies the uncertainty about the data-generating mechanism by the prior distribution and updates it with the observed data to obtain the posterior distribution. Inference about the model is therefore obtained directly as a probability statement based on the posterior. Although the inferential philosophies are quite different, advances in statistical modeling, computing and theory have led many statisticians to keep both sets of methodologies in their data analysis toolkits.

ACKNOWLEDGMENTS
The authors gratefully acknowledge M. Lavine for contributions to the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Jorge López Puga, Martin Krzywinski & Naomi Altman

1. Puga, J.L., Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 277–278 (2015).
2. Kass, R.E. & Raftery, A.E. *J. Am. Stat. Assoc.* **90**, 791 (1995).

Jorge López Puga is a Professor of Research Methodology at UCAM Universidad Católica de Murcia. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.