Matters of significance

Sound experimental design and analysis require improved statistical training.

Concerns about data quality and reproducibility in biomedical research have been rising. This May, Nature Publishing Group put in place new reporting standards for the research we publish. At the core of these standards is a document that asks authors to disclose technical and statistical information about their study. But because these reporting requirements come after research is completed and the manuscript is written and submitted for publication, they do not actually affect experimental design but rather serve to better expose the existing level of rigor to reviewers and readers.

In an effort to give these topics the attention they deserve and help researchers at the stage of experimental planning and design, we debut a new column: Points of Significance. The column will present important concepts and practical advice about statistics and experimental design in an easily digestible format.

As biological researchers apply increasingly refined techniques such as targeted genome engineering that are likely to yield smaller but more biologically meaningful effects, study design and analysis decisions are more important than ever. Experimentalists are also examining systems at a depth that is orders of magnitude greater than that of just five years ago. Analyzing such data will require pushing the envelope in experimental design standards and analysis.

Fortunately, the necessary understanding of basic concepts of variability, effect size and experimental design essential for guiding good experimental practice can be gained with minimal mathematical sophistication. For much bench research, these fundamental principles inform the design of valid replicable experiments that can be analyzed using standard techniques. But scientists should also know enough to realize when their level of training is insufficient and it is time to talk to a statistician. For large studies, a discussion with a statistician at the study design stage—as is commonly done for clinical studies—can save resources and money and prevent angst.

A considerable amount of basic research flies by the seat of its pants, performed while techniques are still being developed and while it isn't yet known whether usable data will be forthcoming. In these cases, in which statistical consultation may be difficult or inefficient, a basic understanding of statistical concepts can help guide the experimental process and allow the researcher to avoid unproductive or misleading paths of investigation.

Because our intuition about probability can be misguided (p. 809), some form of training is essential for developing a good grasp of fundamental statistical concepts and practice. But despite perennial grumbling about inadequate statistical competence among biological researchers, statistics training is often not part of the core course requirements in biological graduate programs.

A challenge to providing universal training is the difficulty in offering a single course that covers the technical requirements of different fields of biology and provides engaging teaching examples that all students can relate to. Statistics training is thus often relegated to required discipline-specific methodology courses, or students must take available statistics electives. As a result, a substantial number of practicing researchers in biology end up with no formal statistics training.

Basic training in experimental design and statistics should be required in all graduate programs that frequently lead to careers in biological research. Scientific ethics courses are now a core part of many graduate biomedical programs. These could integrate instruction about experimental design and basic statistics to complement ethical considerations in making important decisions at each stage of a research project regarding trade-offs among sample size, methodologies and available time and resources. Because those decisions are intimately connected to the reliability of the results, they possess an inherent ethical element that may not be appreciated by researchers anxious to get results. Tying experimental design and statistics to discussion of scientific ethics could lead to greater appreciation of their importance. However, the move in the United States to external online courses for ethics training makes this fusion difficult; thus, a better solution is a dedicated experimental design course that presents design, statistics and ethics in a holistic manner.

But what should practicing researchers with no formal training in statistics do? There is no shortage of statistics books targeted at biologists. A search of Amazon.com using "statistics for biology" gives no less than 3,000 results, with the top result appropriately titled *Statistics for Terrified Biologists*. Online courses are another option for obtaining the necessary minimum training, but it is difficult for someone immersed in research to make the time commitment either option requires.

We hope that by following in the footsteps of the successful Points of View column on visualization—now organized for browsing on Methagora—that the new Points of Significance column can fill a need and encourage both busy researchers and students to think more about statistics and gain a deeper appreciation of how they can improve the experimental rigor of their work.

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

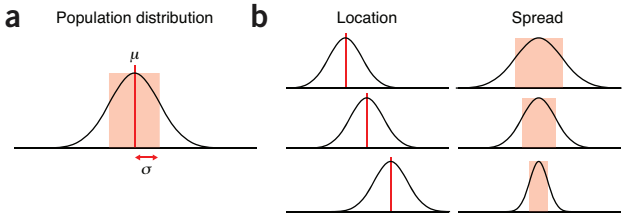

Figure 1 | The mean and s.d. are commonly used to characterize the location and spread of a distribution. When referring to a population, these measures are denoted by the symbols μ and σ.

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

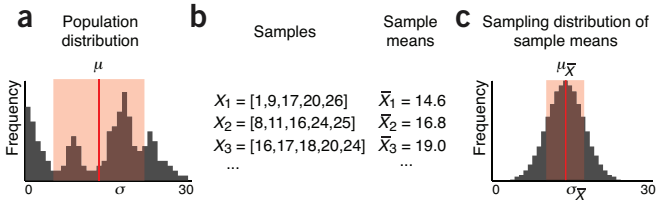

Figure 2 | Population parameters are estimated by sampling. (**a**) Frequency histogram of the values in a population. (**b**) Three representative samples taken from the population in **a**, with their sample means. (**c**) Frequency histogram of means of all possible samples of size $n = 5$ taken from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

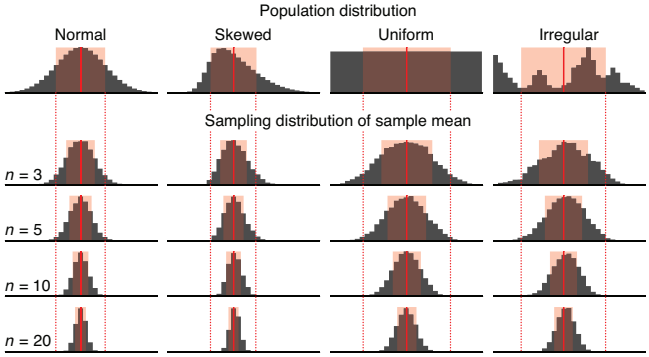

Figure 3 | The distribution of sample means from most distributions will be approximately normally distributed. Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in **Figure 1**.

(**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

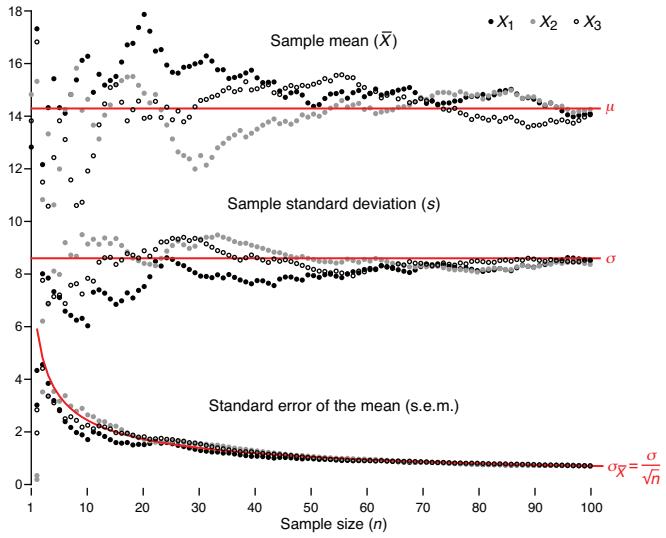

Figure 4 | The mean ($\bar{X}$), s.d. (s) and s.e.m. of three samples of increasing size drawn from the distribution in **Figure 2a**. As n is increased, $\bar{X}$ and s more closely approximate μ and σ. The s.e.m. ($s/\sqrt{n}$) is an estimate of $\sigma_{\bar{X}}$ and measures how well the sample mean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

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

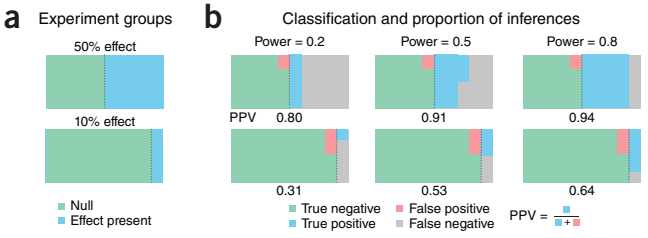

Figure 1 | When unlikely hypotheses are tested, most positive results of underpowered studies can be wrong. (**a**) Two sets of experiments in which 50% and 10% of hypotheses correspond to a real effect (blue), with the rest being null (green). (**b**) Proportion of each inference type within the null 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.

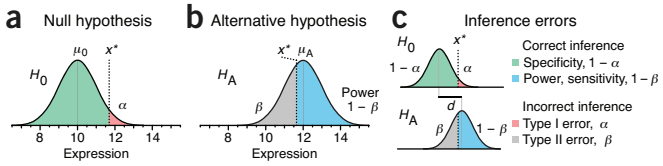

Figure 2 | Inference errors and statistical power. (**a**) Observations are assumed to be from the null distribution (H_0) with mean μ_0. We reject H_0 for values larger than x^* with an error rate α (red area). (**b**) The alternative hypothesis (H_A) is the competing scenario with a different mean 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

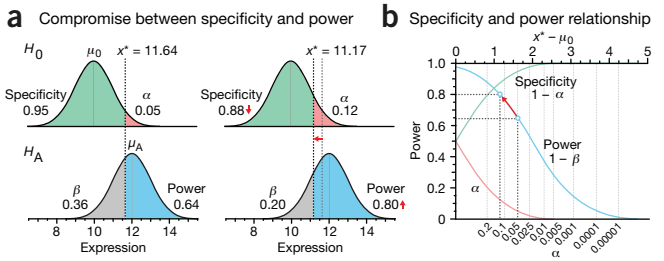

Figure 3 | Decreasing specificity increases power. H_0 and H_A are assumed normal with $\sigma = 1$. (**a**) Lowering specificity decreases the H_0 rejection cutoff x^*, capturing a greater fraction of H_A beyond x^*, and increases the power from 0.64 to 0.80. (**b**) The relationship between specificity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

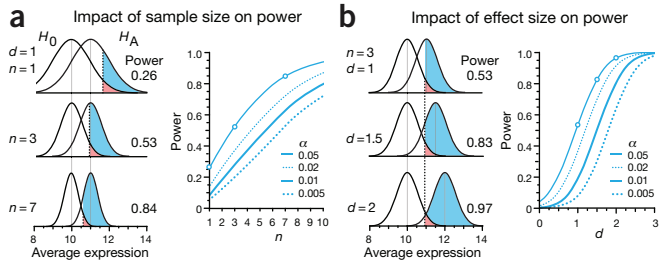

Figure 4 | Impact of sample (n) and effect size (d) on power. H_0 and H_A are assumed normal with $\sigma = 1$. (**a**) Increasing n decreases the spread of the distribution of sample averages in proportion to $1/\sqrt{n}$. Shown are scenarios at $n = 1$, 3 and 7 for $d = 1$ and $\alpha = 0.05$. Right, 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.

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**.

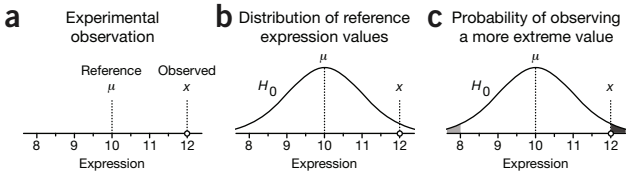

Figure 1 | The mechanism of statistical testing. (**a**–**c**) The significance of the difference between observed (x) and reference (μ) values (**a**) is calculated by assuming that observations are sampled from a distribution H_0 with mean μ (**b**). The statistical significance of the observation 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.

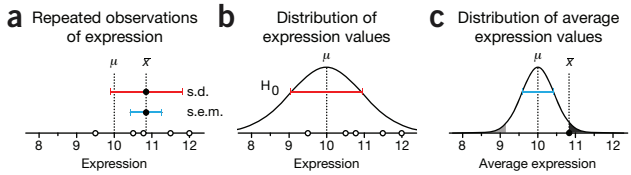

Figure 2 | Repeated independent observations are used to estimate the s.d. of the null distribution and derive a more robust *P* value. (**a**) A sample of $n = 5$ observations is taken and characterized by the mean $\bar{x}$, with error bars showing s.d. (s_x) and s.e.m. ($s_x/\sqrt{n}$). (**b**) The 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

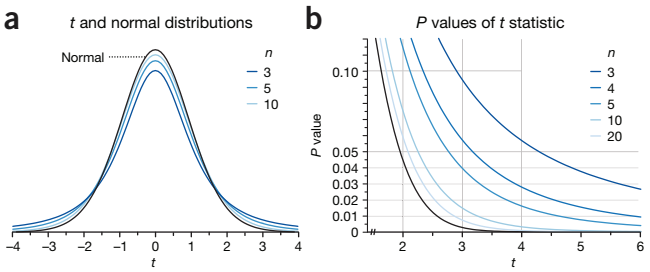

Figure 3 | The *t* and normal distributions. (**a**) The *t* distribution has higher tails that take into account that most samples will underestimate the variability in a population. The distribution is used to evaluate the significance of a *t* statistic derived from a sample of size *n* and is characterized 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

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

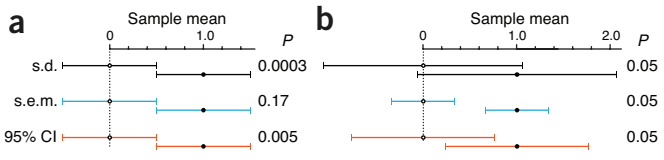

Figure 1 | Error bar width and interpretation of spacing depends on the error bar type. (**a**,**b**) Example graphs are based on sample means of 0 and 1 ($n = 10$). (**a**) When bars are scaled to the same size and abut, *P* values span a wide range. When s.e.m. bars touch, *P* is large ($P = 0.17$). (**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.

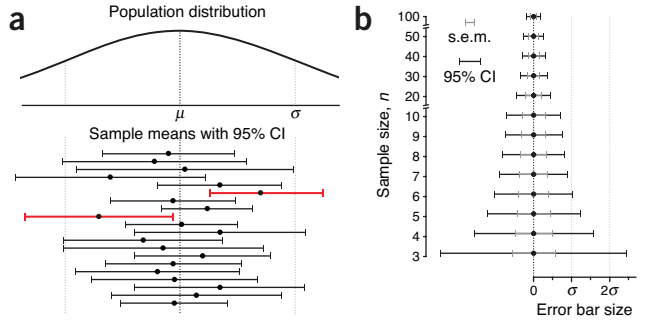

Figure 2 | The size and position of confidence intervals depend on the sample. On average, CI% of intervals are expected to span the mean—about 19 in 20 times for 95% CI. (**a**) Means and 95% CIs of 20 samples ($n = 10$) drawn from a normal population with mean μ and s.d. σ. By chance, two of the intervals (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. ($n > 15$).

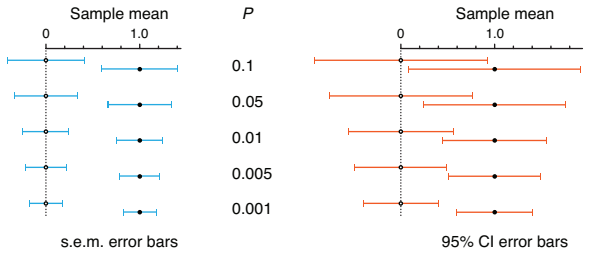

Figure 3 | Size and position of s.e.m. and 95% CI error bars for common *P* values. Examples are based on sample means of 0 and 1 ($n = 10$).

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.