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
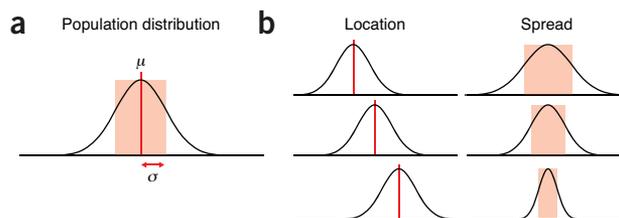

Figure 1 | The mean and s.d. are commonly used to characterize the location and spread of a distribution. When referring to a population, these measures are denoted by the symbols μ and σ.

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
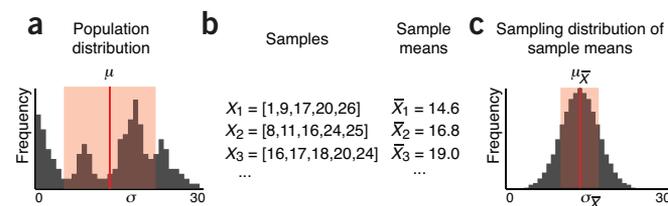

Figure 2 | Population parameters are estimated by sampling. (**a**) Frequency histogram of the values in a population. (**b**) Three representative samples taken from the population in **a**, with their sample means. (**c**) Frequency histogram of means of all possible samples of size $n = 5$ taken from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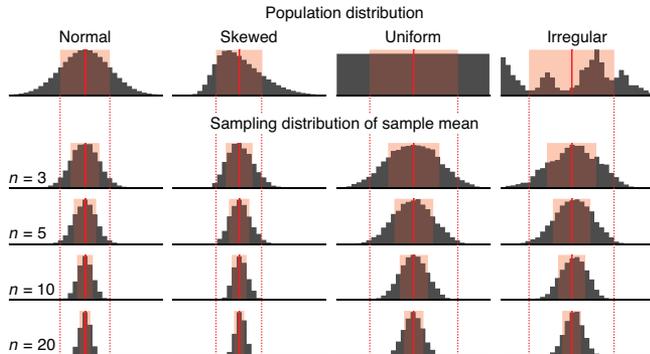

Figure 3 | The distribution of sample means from most distributions will be approximately normally distributed. Shown are sampling distributions of sample means for 10,000 samples for indicated sample sizes drawn from four different distributions. Mean and s.d. are indicated as in **Figure 1**.

(**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., $\text{SE}_{\bar{X}}$) 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
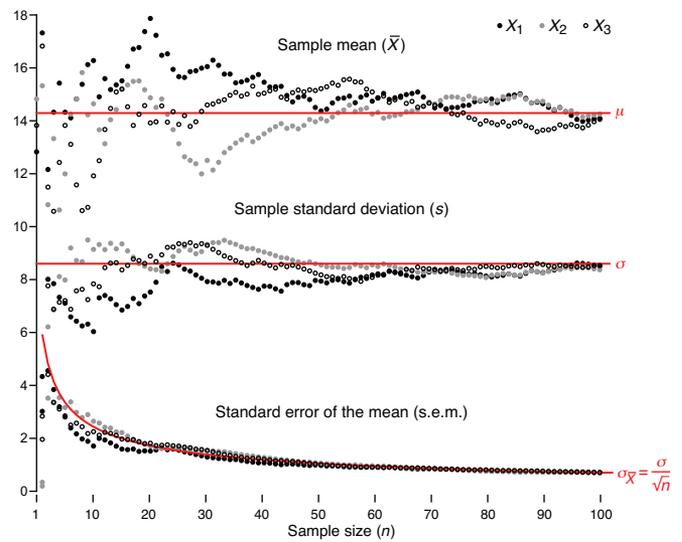

Figure 4 | The mean ($\bar{X}$), s.d. (s) and s.e.m. of three samples of increasing size drawn from the distribution in **Figure 2a**. As n is increased, $\bar{X}$ and s more closely approximate μ and σ. The s.e.m. ($s/\sqrt{n}$) is an estimate of $\sigma_{\bar{X}}$ and measures how well the sample mean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

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.