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.

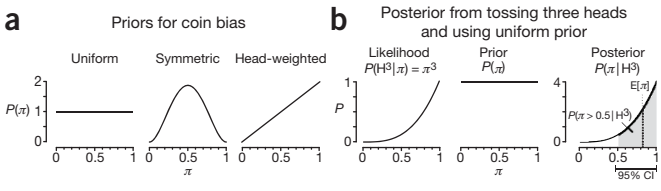

Figure 1 | Prior probability distributions represent knowledge about the coin before it is tossed. (**a**) Three different prior distributions of π, the probability of heads. (**b**) Toss outcomes are combined with the prior to create the posterior distribution used to make inferences about the coin. 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.

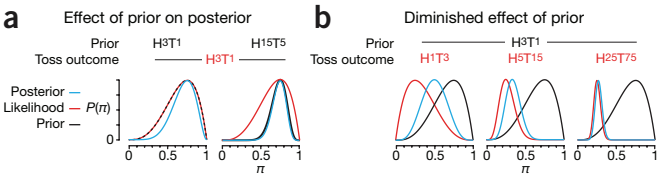

Figure 2 | Effect of choice of prior and amount of data collected on the posterior. All curves are beta(a,b) distributions labeled by their equivalent toss outcome, $H^{a-1}T^{b-1}$. (**a**) Posteriors for a toss outcome of H^3T^1 using weakly (H^3T^1) and strongly ($H^{15}T^5$) head-weighted priors. (**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

Corrected after print 24 September 2015.

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.

Corrigendum: Bayesian statistics

Jorge López Puga, Martin Krzywinski & Naomi Altman
Nat. Methods **12,** 377–378 (2015); **published online 29 April 2015; corrected after print 24 September 2015.**

In the version of this article initially published, the curves (in red) showing the likelihood distribution in Figure 2 were incorrectly drawn in some panels. The error has been corrected in the HTML and PDF versions of the article.