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.

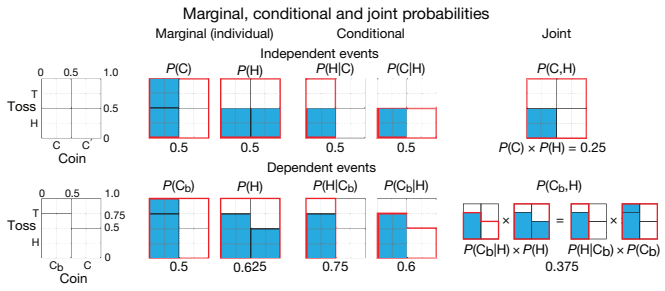

Figure 1 | Marginal, joint and conditional probabilities for independent and dependent events. Probabilities are shown by plots[3], where columns correspond to coins and stacked bars within a column to coin toss outcomes, and are given by the ratio of the blue area to the area of the red outline. The 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.

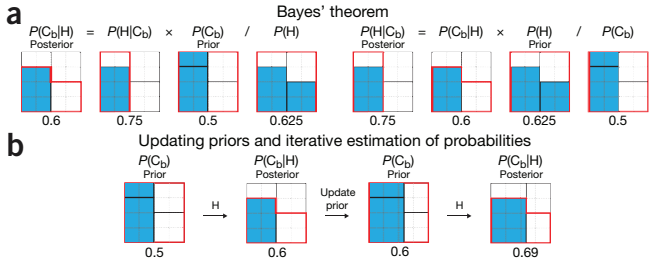

Figure 2 | Graphical interpretation of Bayes' theorem and its application to iterative estimation of probabilities. (**a**) Relationship between conditional probabilities given by Bayes' theorem relating the probability of a hypothesis that the coin is biased, $P(C_b)$, to its probability once the data 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

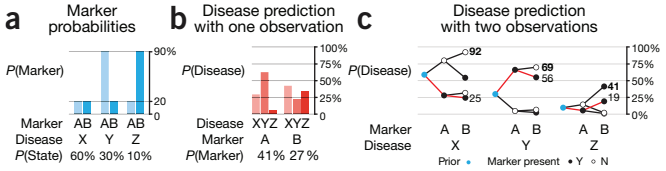

Figure 3 | Disease predictions based on presence of markers. (**a**) Independent conditional probabilities of observing each marker (A, B) given a disease (X, Y, Z) (e.g., $P(A|Y) = 0.9$). (**b**) Posterior probability of each disease given a single observation that confirms the presence of one of the 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.