Scott T. Youngquist, MD, MSc

Part 19: What is a *P* Value?

This review represents a nontechnical explanation of P values intended for the statistical novice.

After transporting several patients with the diagnosis of ruptured abdominal aortic aneurysm (AAA), a flight nurse notices a possible pattern: Patients who died during transport often had been given intravenous ketorolac, a nonsteroidal anti-inflammatory drug (NSAID), administered in the transferring emergency department for relief of undifferentiated back pain before the discovery of the aneurysm. She knows that NSAIDs may inhibit platelet aggregation and wonders whether this may have an effect on retroperitoneal bleeding and subsequent in-transport mortality. A quick review of her transport database reveals that 10 patients were transported with ruptured AAA in the past 5 years. Of these, half died during transport. She constructs a 2 × 2 table cross-classifying AAA according to in-flight mortality and toradol exposure status (Table 1).

Setting aside issues of systematic bias, covered in a prior article, a pattern is apparent. She calculates an odds ratio of 16 for the association between ketorolac exposure and in-flight death from AAA, a rather large effect size. However, the flight nurse wonders whether this association reflects a true causal relationship or the effects of "random variation," the influence of unmeasured and often unquantifiable factors that are associated with the outcome (but not exposure status) yet are assumed to be randomly allotted or distributed between toradol-exposed and nonexposed subjects. A random distribution of these factors does not imply that they are *evenly* distributed between groups, except with high numbers of subjects in each group. Therefore, it is possible that the observed association reflects an uneven distribution of these factors as a result of collecting data on such a small number of subjects. Is there a way to quantify the probability of obtaining these results if they were a result of random variation?

The common use of statistics in modern medicine is an attempt to answer just such a question, by comparing measured outcomes with those theoretical outcomes that would be expected under conditions of random variation. The logic behind this comparison always requires a thought experiment about an alternative world of results that might have been but were never observed.

Let us illustrate with a simpler statistical problem than our analysis of the AAA data: the classic coin toss. Consider the results of 10 coin flips from which 9 heads and 1 tail are obtained. Given the high frequency of heads observed, we might suspect this is an unbalanced (or biased) coin. But how do we test this concern? One way is to compare the observed results with results we would expect if the coin were unbiased. We know that if we were to perform a single set of 10 coin flips with an unbiased coin, we might get 9 heads by chance alone. But what are the chances?

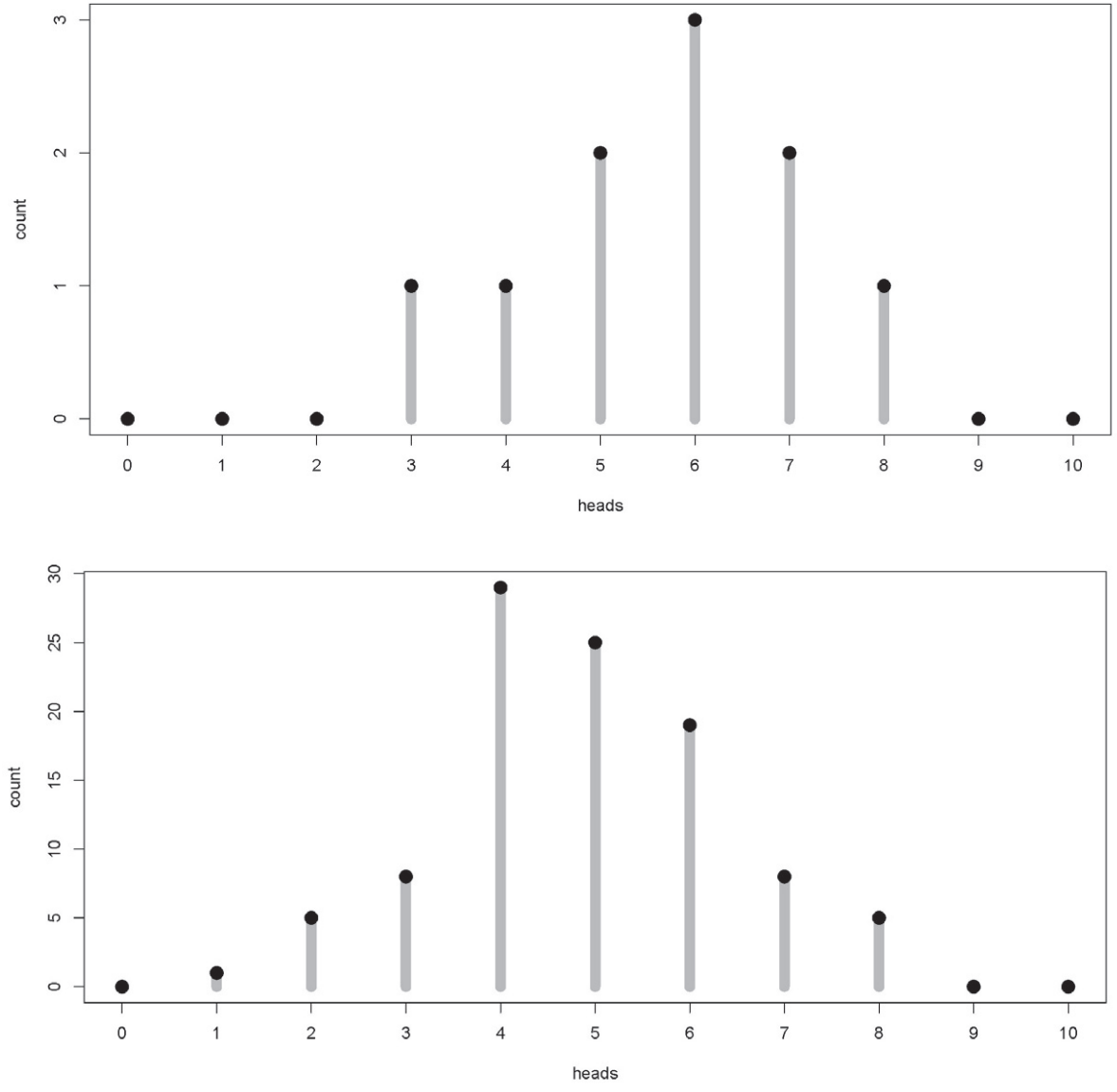
Let us perform the same trial (10 coin tosses) with an unbiased coin repeatedly, tallying the number of heads each time. The results will allow us to determine the relative frequency of obtaining 9 heads from such a fair coin. Figure 1 is a panel summarizing the hypothetical results of these trials. I generated these simulated results using the freely available R statistical software (R 2.14.0, R Foundation for Statistical Computing, Vienna, Austria). For the figure, imagine that each time a set of 10 coin tosses resulted in exactly 4 heads (and thus 6 tails), a casino chip or marker was placed over the number 4 place on the x-axis and likewise for any other number of heads obtained out of 10 tosses. With more and more such trials (Figure 1A–D), a stair-stepped pile of markers with the number of such markers at any location on the x-axis approximately proportional to the long-run probability of getting that many heads out of any given 10 tosses of the coin is built up. The height of each resulting marker stack is summarized by vertical spikes in the figure. One can see that with more and more trials of the fair coin, a pattern of unbiasedness takes shape, and the distribution of markers over potential outcomes stabilizes. After 1,000 repetitions of the trial, 5 of 10 heads is clearly the most likely outcome, with values departing from 5 in either direction having increasingly less probability.

The probability of obtaining 9 heads in 10 coin tosses can be approximated empirically in this example by simply counting the number of markers stacked on the 9s place and

Table 1. In-Flight Mortality and Toradol Exposure

	Died during Transport	Survived Transport
Ketorolac exposed	4	1
Ketorolac unexposed	1	4
Total	5	5

Figure 1. A, "Spike" histogram of results of 10 trials of 10 coin flips from an unbiased coin. If four heads were obtained in a single trial of 10 coin flips, the result contributes one frequency count to the bin labeled "4" and likewise for other values. Although 5/10 heads is the expected result from an unbiased coin, actual results vary because of random variation. With only a small number of trials, the shape of the distribution has yet to be clearly established. **B, Results from 100 trials of 10 coin flips from an unbiased coin.** Interpretation as in 1A.

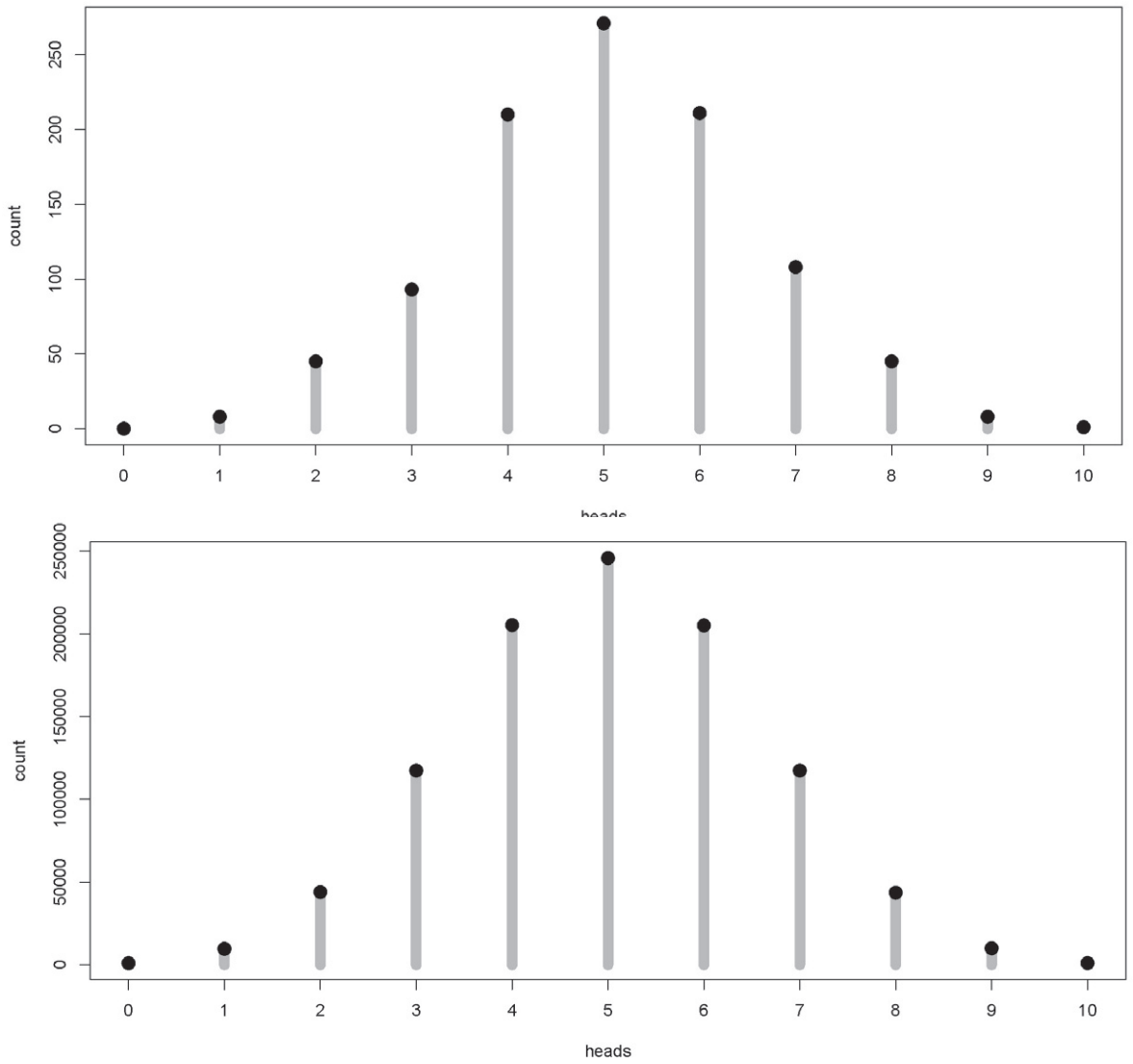


then dividing that number by the total number of markers placed. This value will get more and more accurate as we increase the sample size toward infinity and answers the question of how likely a result is 9/10 heads if this observation came from the single trial of an unbiased coin. The exact probability of such an outcome turns out to be extremely small. Out of 1,000 simulated trials, we observed an outcome of 9 heads exactly 10 times, giving a probability of approximately 1%. (The exact probability is 0.97656250% when this value is solved analytically.)

With only a 1% probability that this outcome would have resulted from an unbiased coin, we would probably like to try some more coin tosses before deciding whether to trust the coin, especially if we had a wager riding on the outcome of future tosses. But most experiments in biomedical research are not so simple to reproduce. We almost always have to draw a conclusion about the coin's lack of bias based on the results of a single trial of 10 coin tosses, so to speak. Therefore, we often simply conclude that the coin is biased, given the low likelihood of obtaining this single result if the coin were fair.

The 1% probability calculated above is an example of a *P* value. The *P* in *P* value stands for probability, and the same general approach to statistical inference applies to most so-called frequentist analyses of biomedical data. (The term *frequentist* refers to the endorsement of probabilities as the limiting value of long-run frequencies. In other words, these analyses are based on the

Figure 1. C, Results from 1,000 trials of 10 coin flips from an unbiased coin. Interpretation as in 1A. The probability mass distribution has taken on a fairly stable shape, with 5 heads being the most common outcome and results clustering around this expected outcome. D, Results from 1 million trials of 10 coin flips from an unbiased coin. Interpretation as in 1A. The shape of the probability mass distribution is virtually unchanged from that obtained by performing the experiment 1,000 times. The probability of obtaining any 9/10 heads with any 10 coin tosses is approximately equal to the number of outcomes in which 9 heads were obtained divided by the number of experiments performed.

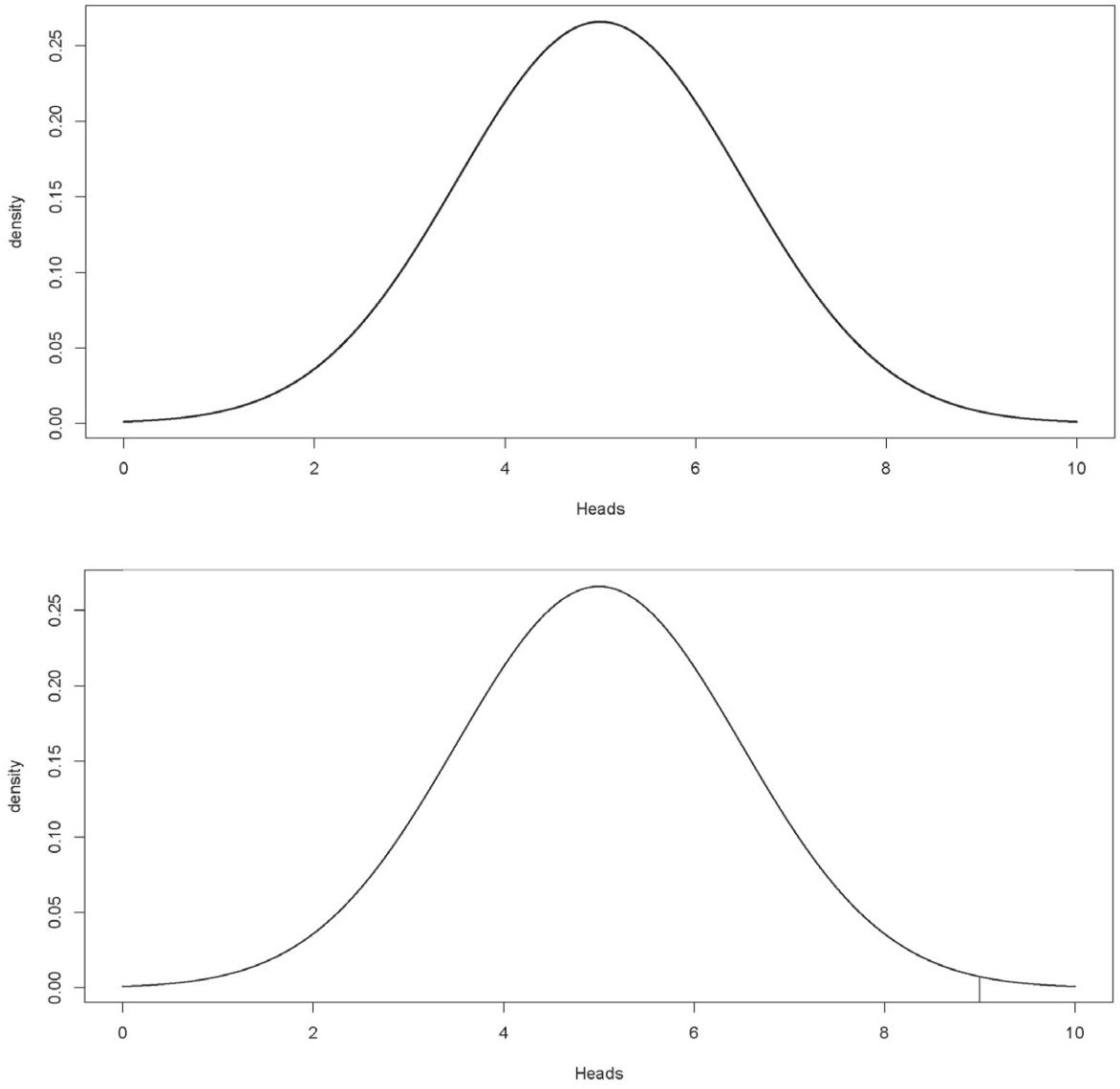


expected results of performing the experiment an infinite number of times with an unbiased coin, so to speak.) In a way, the calculation of a *P* value is really just a thought experiment, as stated previously. This thought experiment is constructed as follows:

First, assume that there is no true difference in outcomes over the long run. This is the so-called null hypothesis (H_0). In the coin toss example, we assumed there was no true difference in the chance of obtaining heads or tails, equivalent to saying the coin is fair. In the AAA example, the H_0 assumption would be that the true odds ratio (OR) for the association between ketorolac administration and death during transport from AAA is really unity (OR = 1, reflecting exactly equal odds of in-transport death among groups).

Second, assign a model probability distribution to the expected long-run distribution of H_0. A model probability distribution is best understood graphically as in the example of the coin toss, where markers were placed over values corresponding to the number of heads obtained per 10 coin tosses, with a single peak at the most likely value and other values assigned markers in proportion to their probability of outcome. A model probability distribution should similarly assign maximum probability to H_0, because we are assuming it is true, but also demonstrate the random variation in outcome we would expect if we were to repeat the very same study *ad infinitum*. The choice of a model probability distribution always requires implicit *a priori* assumptions about how the outcomes would

Figure 2. A, A normal distribution approximating the expected outcomes from the coin toss experiment. Instead of frequency counts on the y-axis, however, the relative density of the area under the curve is referenced. **B,** A line is drawn indicating the location of the outcome 9 heads out of 10 coin tosses. This line represents the exact probability of 9 heads under this probability distribution. Because a line has no width, the area under the curve at any single point is 0, and thus the exact probability of this single outcome approaches 0.

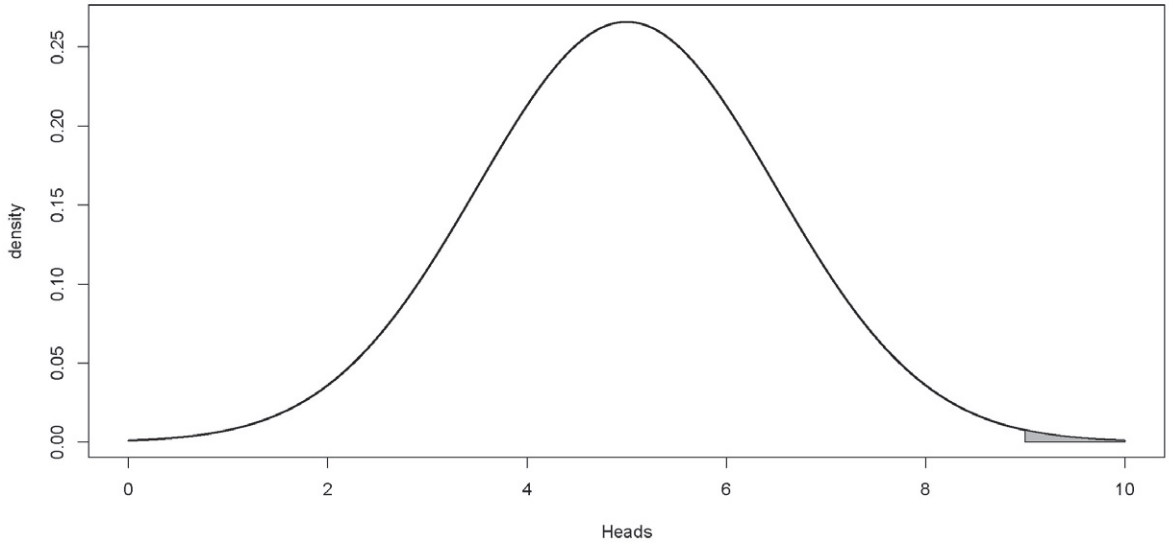


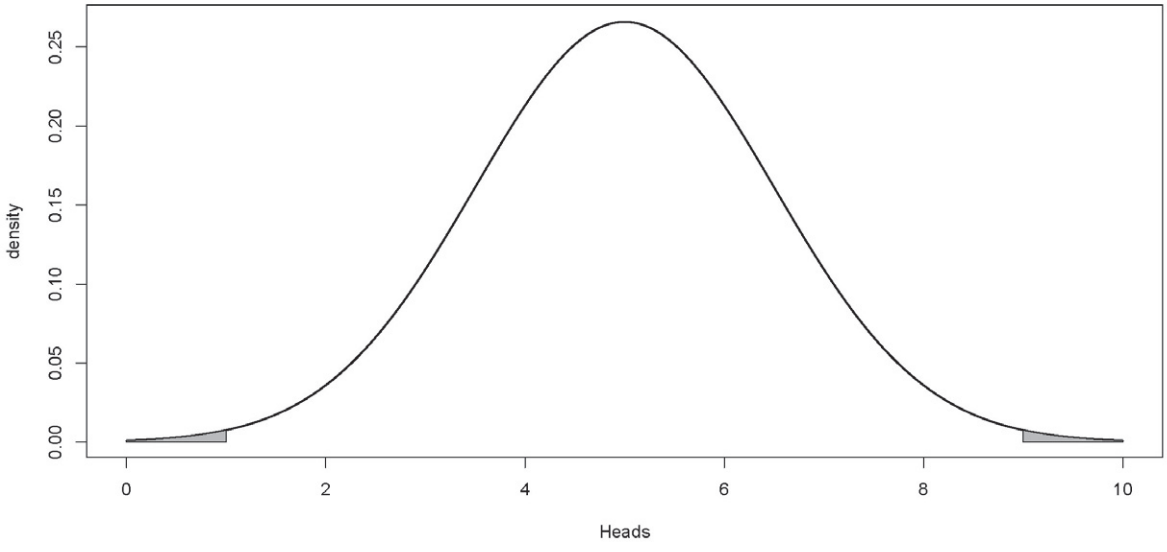
be distributed if the very same experiment were repeated *ad infinitum*. The observed result (such as the OR = 16) is then assumed to be a single sample randomly drawn from such a distribution. In reality, especially with observational data, what possible randomization mechanism could have led to the results is unclear, but this has seldom stopped anyone from using a *P* value to make inferences with them.

To summarize then, like the distribution of casino markers created by performing the coin toss experiment a million times, a probability distribution has a shape that serves to describe mathematically where probability lies. By far the most common type of shape adopted to model the expected probability of H_0 is of the type seen in Figure 2: the normal distribution, a mathematical function that models continuous probability and has one peak (located over the point on the x-axis where probability is highest) with a symmetrical decline on each side. This is also commonly called the bell-shaped curve.

Notice how the normal distribution we have created is like a smoothed outline of our stacked casino markers from Figure 1. I asked the R software to draw the normal distribution for paneled Figure 2 by telling it what the mean value should be (5 heads), to indicate where the peak probability density will occur (at 5 heads), as well as the value for the standard deviation, a mathematical summary of how far val-

Figure 2. C, Shading has been added indicating the cumulative area under the curve for results at least as extreme as those observed for the calculation when 9 heads is obtained. The P value is simply the proportion of the area under the curve that is shaded. **D,** Shading has been added indicating the cumulative area under the curve for results at least as extreme as those observed in both directions, allowing for the calculation of a two-sided P value.





ues are expected to vary from the average and which determines the horizontal spread of the distribution.

The coin toss example required a model of *discrete* probability, where *discrete* refers to the fact that values between the counting numbers are inadmissible and thus receive zero probability. (One cannot get 7.5 heads out of 10 tosses, for example, thus the gaps between spikes in Figure 1.) The normal probability distribution, however, is an example of a continuous probability distribution, appropriate for assigning probabilities to outcomes whose possible values are continuous along the number line, such as the difference in mean arterial pressure between groups.

However, for ease of calculation, we might (and often do) decide that the coin toss distribution is close enough to the normal distribution in its shape to use the latter's mathematically well-described properties to model the expected distribution of H_0. In other words, we might assume that if H_0 were true, the expected outcomes for that stepped mountain of markers produced by an infinite repetition of the coin toss experiment would look pretty much like the smoothed normal distribution in its shape and width. Then we can use the relatively straightforward formulas that describe the normal distribution to calculate our probabilities.

The third step in our effort is to do just that: calculate the probability of obtaining the observed result under the assumptions just stated. This is the *P* value. In the case of the coin toss, this probability was easy for us to obtain using simulated outcomes. A formula exists to do this for us algebraically, allowing us to forego the simulation exercise and giving us an exact probability associated with this discrete outcome. However, in the case of a continuous probability distribution, like one that might be used to assess the average difference in blood pressure measured between groups, the exact probability for any single outcome is actually 0.

To see why this is so, consider Figure 2. The total area under the curve represents all possible outcomes or 100% of the probability space. To determine the relative area under the curve associated with obtaining exactly 9 heads, we have to draw a straight line up from the number 9 on the x-axis until it reaches the curve above (Fig. 2B). But lines do not have any area, because, by definition, they lack width (area = width × length). Therefore, when using a continuous probability distribution, the exact probability of any single value is 0. We therefore have to compute a *cumulative* probability over a range of possible outcomes when the distribution is continuous to be able to calculate any probability at all.

By theory and convention, we employ values "as or more extreme" to generate this nonzero area under the curve. In simple terms, we take the continuous interval along the x-axis that includes the observed value on one end and all other values going away from the null value along the number line to calculate a *P* value. This has been represented by the shaded area of the distribution in Figure 2C. So-called two-sided *P* values additionally employ the symmetrical area on the opposite side of the curve for this calculation, as in Figure 2D.

Now that we have constructed a *P* value, how should it be interpreted? Unfortunately, the interpretation of *P* values is an area of philosophical uncertainty going back to their invention first by Sir R. A. Fisher and later by Jerzy Neyman and Egon Pearson. In practice, researchers generally reject the null if the *P* value is below a certain threshold, almost universally a value of $< .05$, embracing instead the alternative hypothesis: the coin is not fair.

There are many problems with this method of inference. First of all, there is no theoretical basis for concluding that 5% probability or less under the null hypothesis is universally strong enough evidence to reject *all* null hypotheses, regardless of subject matter. Nevertheless, this amount of exact or cumulative probability is widely used to determine whether a measured difference is "statistically significant." Like a Good Housekeeping seal of approval, a statistically significant *P* value is viewed as certifying the validity of reported findings. *P* values, however, mathematically shrink as the N becomes large or as the variability in the outcome (standard deviation) becomes small, regardless of whether a true difference exists.

Furthermore, the use of *P* values to reject the null hypothesis, as commonly used, is a form of scientific falsification, an approach popularized by the philosopher Karl Popper, entailing the systematic refutation of mutually exclusive and exhaustive hypotheses until the sole remaining hypothesis must be true. This "last man standing" approach to inductive reasoning cannot provide any direct support to alternative hypotheses, however, a fact that is frequently overlooked by researchers. When the null hypothesis has been rejected, an alternative to the null hypothesis must be embraced. But a common alternative to the null hypothesis, H_a: a true difference in outcome exists between groups, is such a broad proposition, often involving a universe of possible magnitude, as to represent only a small advancement in the establishment of scientific knowledge.

The construction of *P* values, specifically from continuous probability distributions, has also been criticized for its reliance on unobserved values "as or more extreme" than the one observed and more generally for its inability to account for evidence external to the experiment. The biostatistician Steven N. Goodman has criticized the mass misconception that "absent any consideration of biological plausibility and prior evidence, statistical methods can provide a number [the *P* value] that by itself reflects the probability of reaching erroneous conclusions."[1]

Perhaps the most common problems with the use of *P* values involve their misinterpretation. They are not, as is commonly stated, the probability that the null hypothesis is correct. They cannot be, for their calculation starts with the assumption that the null hypothesis is true! Neither do they represent the probability that the results are a false positive. The probability that a given result is a false positive depends on the prior probability of the null hypothesis for its calculation, which can be estimated but is generally unknown. And ultimately, systematic bias is the major threat to the validity of most findings, not the effects of random variation. The *P* value assumes that no such bias occurred in the execution of the trial.

Confidence intervals (CIs) have been proposed as an alternative to the use of *P* values for statistical inference. Confidence intervals give a range of outcome values constructed mathematically using the standard deviation as the principal determinant of width and a user-determined coverage percentile (usually 95%).

What can be said about a 95% CI is that, were the study repeated an infinite number of times in the same fashion, an interval so constructed would overlap the value representing the true association 95% of the time. In practice, researchers often use the CI as a *P* value surrogate for hypothesis testing, because it is well known that if the 95% interval overlaps the value representing the null hypothesis, the *P* value will be $> .05$. They also often wrongly assume that there is 95% probability that the reported CI contains the true value when the 95% probability can apply only to a long-run frequency in which many confidence intervals are produced under a repetition of the experiment, which is seldom, if ever, performed. The probability of a single reported CI containing the true value is unknown. As a relative indicator of the precision with which the outcome has been measured, however, the CI can be useful.

Continued on page 71

Returning to the vignette that started this review, the flight nurse calculates a *P*-value of 0.10 with a 95% CI for the OR of 0.72 to 354 for the association between in-flight death from ruptured AAA and ketorolac exposure. Because this value is greater than .05 and the confidence limits cross unity, she realizes these data are unlikely to be considered "proof" of the measured association (nor acceptable to most medical journals). She knows that with more data points the confidence limits will shrink and, if the association holds, her *P* value will also shrink. She concludes that her review of a single flight program is underpowered to make material conclusions about the association, and she will need to collaborate with another flight program to generate a larger N. She approaches a statistician to help her estimate a sample size that should be sufficiently large enough to demonstrate statistical significance.

Despite all the problems and abuses associated with the use of *P* values, they remain the virtual de facto method of assessing data, for better or worse. Without an understanding of what *P* values do or do not mean, the impetus for changing this paradigm is minimal. Alternative techniques, such as the use of Bayes factors or fully Bayesian analysis, exist and are frequently more justifiable methods of induction than the *P* value paradigm. These alternative methods deal directly with the strength of evidence in favor of any hypothesis, not just the null, and can incorporate explicitly into the calculation of probability prior assumptions about the outcome. For now, realize that the *P* value and CIs can serve, at least, as a rough estimate of the precision with which observations have been measured. However, the *P* value's ability to determine the truth about measured effects is usually quite limited.

Reference

1. Goodman SN. Toward evidence-based medical statistics 1: the P value fallacy. *Ann Intern Med* 1999;130:995-1004.

Scott T. Youngquist, MD, MSc, is an assistant professor in the Department of Surgery, Division of Emergency Medicine, at the University of Utah School of Medicine in Salt Lake City, director of the Air Medical Research Institute, and medical director of the Salt Lake City Fire Department. He can be reached at scott.youngquist@utah.edu.

1067-991X/$36.00

doi:10.1016/j.amj.2011.12.007