Table 1 Related Functions Available in Four Popular Software Packages

Names	*Problem 1*	*Problem 2*	*Problem 3*	*Problem 4*	*Problem 5*
Excel	NORMDIST	NORMINV	NORMDIST	N/A	CHART
SPSS	CDF.NORMAL	IDF.NORMAL	PDF.NORMAL	RV.NORMAL	N/A
SAS	CDF	QUANTILE	PDF	RAND	PROC PLOT
Matlab	normcdf	norminv	normpdf	normrnd	plot

For some of the most popular software packages, Table 1 provides the names of functions, modules, and procedures that can be applied to the five very typical problems related to the normal curve and the normal distribution. Note that none of the names is case sensitive, with the exception of Matlab functions and modules, which have to be used in lowercase letters.

4. The family of normal curves that are bell shaped is only for the univariate case, in which only one variable x is involved. However, in the case of multivariate data analysis, the multivariate normal model, which extends the univariate normal distribution model, is commonly used. One example is a bivariate normal distribution model, which applies to two variables. In that case, the bell-shaped normal curve becomes a bell-shaped surface in three dimensions. Accordingly, the probability is indicated by the volume under the bivariate normal distribution surface.

—Hongwei Yang

Further Reading

Aron, A., Aron, E. N., & Coups, E. (2005). *Statistics for psychology* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Johnson, D. E. (1998). *Applied multivariate methods for data analysts*. Pacific Grove, CA: Duxbury.

Martinez, W. L., & Martinez, A. R. (2001). *Computational statistics handbook with Matlab*. New York: Chapman & Hall.

Patel, J. K. (1982). *Handbook of the normal distribution*. New York: Dekker.

Tamhane, A. C., & Dunlop, D. D. (2000). *Statistics and data analysis: From elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall.

Normal curve information:
http://en.wikipedia.org/wiki/Normal_distribution
http://mathworld.wolfram.com/NormalDistribution.html
http://www.answers.com/topic/normal-distribution
http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html
http://www.tushar-mehta.com/excel/charts/normal_distribution
http://www-stat.stanford.edu/~naras/jsm/FindProbability.html

Null Hypothesis Significance Testing

Null hypothesis significance testing (NHST) dominates experimental and correlational methods in psychological research. Investigators are typically concerned with demonstrating the existence of an effect, that is, systematic variation in the data that can be distinguished from random noise, sampling error, or variation due to uncontrolled or nuisance variables. The null hypothesis is often, but does not have to be, identified with chance, and a p value is computed to express how improbable observed empirical data are under the assumption that the null hypothesis is true. When this probability falls below the conventional value of .05, it is concluded that the null hypothesis is false and that it is safe to presume the presence of a systematic source of variation. This inference is not strictly logical because *modus tollens* is not valid when stated probabilistically: From the statement "If the null hypothesis is true, then extreme data are improbable," it does not follow that "If the data are probable, the null hypothesis is false." Because NHST is a method of inductive, not logical, inference, researchers nevertheless believe that the rejection of the null hypothesis indicates the presence of an effect. In the long run, the argument goes, decisions reached

by NHST will generate knowledge faster than would guessing or doing nothing.

Variants of NHST have been developed by various, and sometimes warring, schools of statistical thought. These schools differ in the assumptions they make about the nature of the data and the hypotheses and about how to make inferences. The following illustrations of possible inference strategies begin with information- and assumption-rich scenarios and proceed to the more degraded scenarios typical of most psychological research.

Full-Suite Analysis

Suppose extensive testing has revealed that average self-esteem scores are $\mu = 68$ and 72 for women and men, respectively, and that the standard deviation within each gender is $\sigma = 20$. A sample of 200 scores with a mean of 71 is drawn from one of the two populations. The null hypothesis H_0 is that women were sampled, and the alternative hypothesis H_1 is that men were sampled. Analysis begins with the calculation of the probability of obtaining a mean of 71 or higher if H_0 is true. The z score for the sample mean is

$$(71 - 68)\sqrt{200}/20 = 2.12,$$

and the probability of a score at least this extreme is .017.

Evaluation of the data under the alternative hypothesis H_1 yields $z = .71$, $p = .24$. That is, the data are not improbable under the assumption that men were sampled. The likelihood ratio (LR) of the two p values, $p(\mathrm{D}|H_1)/p(\mathrm{D}|H_0)$, is 14.12, meaning that it is more than 14 times more likely that a sample of men rather than women would yield data of the kind found in the empirical sample. But how likely is it that the sample consisted of men? It is necessary to be explicit about the prior probability of sampling men. A simple intuition is that women and men were equally likely to be sampled, that is, $p(H_0) = p(H_1) = .5$. The summed products of these prior probabilities and their respective p values is the overall probability of the observed data. Here, $p(\mathrm{D}) = p(H_0)p(\mathrm{D}|H_0)+p(H_1)p(\mathrm{D}|H_1) = .13$. This probability is critical for the calculation of the probability of the null hypothesis given the observed data. Bayes' theorem gives $p(H_0|\mathrm{D})$ as $\mathrm{p}(H_0)p(\mathrm{D}|H_0)/p(\mathrm{D}) = .07$. Because the prior probabilities of the two hypotheses are the same, the ratio of the two posterior probabilities is the same as the LR. It can now be said that the sample is more than 14 times more likely to comprise men than women. The assumption of equal priors was just that, an assumption. Suppose the researcher knew that self-esteem scores were collected at four different sites, only one of which comprised men. Now $p(H_0|\mathrm{D}) = .18$, meaning that it is only 4.7 times more likely for the sample data to come from men than from women. Although the prior probability that men were sampled was low, the evidence is still strong enough to reject the null hypothesis that women were sampled and to accept the alternative.

Now consider a study in which 200 women and 200 men are sampled. The null hypothesis is that there is no gender difference in average self-esteem scores (H_0: $\mu_{\text{women}} = \mu_{\text{men}} = 70$, $\sigma = 20$), and the alternative is that there is a 4-point difference (H_1: $\mu_{\text{women}} = 72$, $\mu_{\text{men}} = 68$, $\sigma = 20$). During the early phase of the research program, the two hypotheses may appear to be equally likely to be true. If the gender difference in the sample means is 3.5 points, the revised probability of the null hypothesis is $p(H_0|\mathrm{D}) = .09$. As evidence accumulates, researchers become aware that some hypotheses are riskier than others. Suppose gender differences in self-esteem have become well established, so that $p(H_0) = .1$. Now a 3.5 gender difference still renders the null hypothesis less probable ($p(H_0|\mathrm{D}) = .01$), but there is less room to move.

These examples are idealized: The properties of the two populations (μ and σ) are known, and credible estimates of their prior probabilities are available. Scientific research must often proceed without this full suite of information. Researchers handle the lack of information by suspending certain kinds of inference or by making defensible assumptions where

good information is missing. If the prior probabilities of competing hypotheses are unavailable in a quantifiable and agreed-on format, they can sometimes be estimated on the basis of prior research or derived from theory.

Power

Many researchers are careful to situate their findings within the context of relevant empirical or theoretical work but refrain from making explicit estimates for their hypotheses to be true. Suppose again that a 3.5 gender difference in self-esteem is found. Evaluation of the data under the two hypotheses yields $p(D|H_0)$ = .04 and $p(D|H_1)$ = .401, and thus LR = 10. With prior probabilities barred from quantitative inferences, researchers can still estimate their study's statistical power to detect a 4-point gender difference. The power of the study is the probability that the null hypothesis will be rejected if it is indeed false. To obtain this probability, it is necessary to find the minimum gender difference leading to the rejection of H_0. This difference is given by the product of the z score at which $p(D|H_0)$ = .05 and the standard error of the difference. (The standard error of the difference between means $\overline{X}_1$ and $\overline{X}_2$ is

$$\sigma_{\bar{x}1} - \sigma_{\bar{x}2} = \sqrt{\sigma^2_{\bar{x}1} + \sigma^2_{\bar{x}2}};$$

here, $1.65 \times 2 = 3.3$.)

The power of the study is the complement of the probability of such an effect under the alternative hypothesis. Here, $1 - p(D|H_1)$ = .64. In other words, the prior odds that this study would detect an existing difference of 4 points were about 5 to 3.

In principle, many researchers agree that the null hypothesis should not be rejected when $p(D|H_0) > .05$. In practice, however, they tolerate a good number of exceptions, thus opening the door to the murky world of "marginal significance." Researchers typically care more about limiting the probability that a true null hypothesis is rejected than about increasing the probability that a true effect is detected. Designing a study with a power of .8 is a widely held but seldom attained ideal. One reason for this shortfall is that power consumes resources. In the present example, a total sample of 1,152 individuals would be required to reach the ideal.

Making Decisions

In the decision-theoretic school of hypothesis testing, $p(D|H_0)$ = .05 signifies the probability with which a true null hypothesis is rejected. This decision outcome constitutes a Miss (M). Conversely, $p(D|H_1)$ is the probability that a false null hypothesis is not rejected, a circumstance called a False Positive (FP). The complement of an M is a Hit (H), that is, the retention of a true null hypothesis; the complement of an FP is a Correct Rejection (CR), that is, the rejection of a false null hypothesis. The probability of CR is the power of the study. If there are no resources to increase power, it is tempting to admit more FP. If the null hypothesis is rejected with $p(D|H_0)$ as high as .10, power increases from .64 to .76. The practice of adjusting $p(D|H_0)$ is frowned on, however, when it reflects, not the state of the field and thus appropriate prior probabilities, but rather the researcher's desire to obtain significant results.

Without prior beliefs, there is no way of estimating how probable the four outcomes are. It is only possible to state the conditional probability of p(H) relative to p(M) and of p(FP) relative to p(CR). To illustrate what can be gained from estimating the prior of the null, consider $p(H_0)$ = .75, .5, and .25. The top panel of Figure 1 shows the four conditional probabilities obtained in a high-powered study. Each quadrant of the bottom panel gives three unconditional probabilities that are obtained as products of the conditional probabilities and the prior probabilities of the hypotheses. When, as in the typical empirical case, the probability of rejecting a true null hypothesis (M) is smaller than the probability of accepting a false one (FP), any decrease in the prior probability of H_0 decreases the overall probability of correct decisions

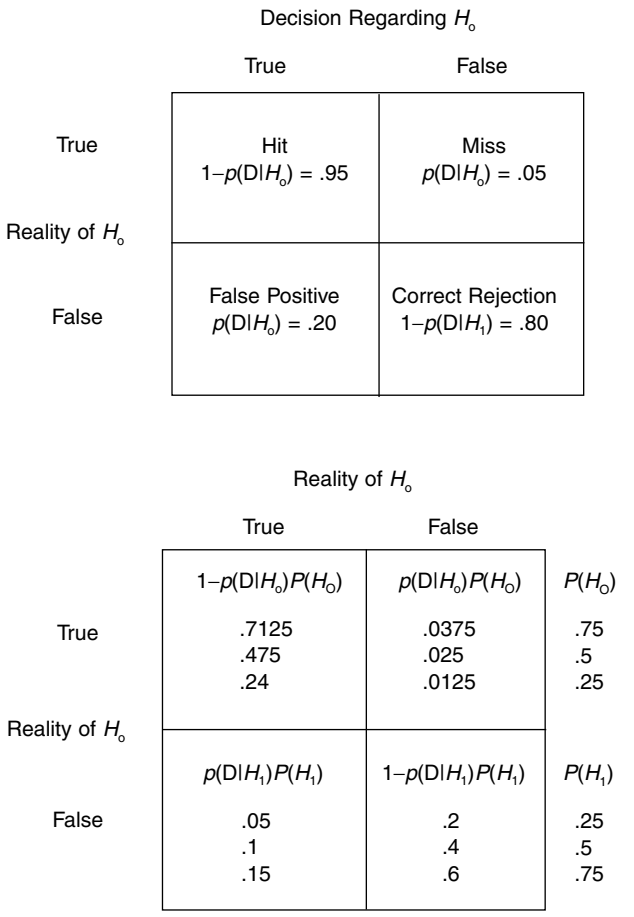

Figure 1 A Decision-Theoretic Scheme for Null Hypothesis Significance Testing

(here, $p(H) + p(CR) = .91$, .88, and .84, respectively, for $p(H_0) = .75$, .5, and .25). This is an odd, but logical, result. As an area of research becomes more mature, null hypotheses become less probable, and the typical conservatism of decision making (i.e., power < 1 − desired significance level) makes it more likely that true effects are missed. Failures to replicate then accumulate, not because previously demonstrated phenomena do not exist, but because studies lack power. Hence, even the principled use of NHST delays scientific progress. The data of solidly designed but underpowered studies are dismissed for ad hoc reasons, or worse, they are seen to add up to a store of anomalies that potentially undermines hard-won knowledge.

Filling In

When a new area of research opens up, it is marked by great uncertainty. The null hypothesis may not have a defensible prior probability, and there may not be a well-formulated alternative hypothesis. Without being able to estimate the posterior probability of the null and with no opportunity to estimate the power of the study, researchers seek to collect only enough data to reject the null. When they do, they can declare only that an effect has been found, and they can report its size (for example with Cohen's *d* or Pearson's *r*). Power analyses can be performed with the obtained effect size, but the wisdom of this practice is a matter of debate. Nevertheless, when enough empirical effect sizes have been reported to justify their aggregation by meta-analysis, these combined effect sizes can serve as point-specific research hypotheses for replication and extension studies.

The life course of a typical research area entails a paradox. In the early stages, NHST can be performed only in its most rudimentary form. At this stage, misinformation and miseducation are most likely to contribute to fallacious conclusions, such as the widespread belief that the *p* value of the data signifies the improbability of the null hypothesis. In the late stages, when specific alternative hypotheses are available, when the power of a study can be determined, and when the probability of hypotheses can be estimated, new data contribute little incremental knowledge. Although NHST can then be used with great precision, its purpose is now to produce judgments about the acceptability of the data and not about the truth or falsity of the hypotheses.

—Joachim I. Krueger

Further Reading

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*, 997–1003.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*, 19–24.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*, 16–26.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, *27*, 313–376.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.