From Berger, R.L., 2015. Hypothesis Testing in Statistics. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 11. Oxford: Elsevier. pp. 491–493.
ISBN: 9780080970868

Hypothesis Testing in Statistics

Roger L Berger, School of Mathematical and Natural Sciences, Arizona State University, Phoenix, AZ, USA

This article is a revision of the previous edition article by G. Casella, R.L. Berger, volume 10, pp. 7118–7121, © 2001, Elsevier Ltd.

Abstract

A statistical hypothesis is a statement about one or more population parameter(s). A hypothesis test is a rule for deciding which of two complementary hypotheses is true, based on some data. Various methods for constructing hypothesis tests (e.g., likelihood, Bayesian, intersection–union) are available. Various measures for evaluating hypothesis tests, most of which are related to the tests' probabilities of making correct or incorrect decisions, are available, also. Construction and evaluation of tests, if a large sample is available, is sometimes simplified by asymptotic results.

Introduction

A *statistical hypothesis* is a statement about one or more population parameter(s), and the two complementary hypotheses in a hypothesis-testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by H_0 and H_1, respectively.

If θ denotes a population parameter, the general format of the null and alternative hypotheses is $H_0: \theta \in \Theta_0$ and $H_1: \theta \in \Theta_0^c$, where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. A hypothesis test is a rule for deciding whether H_0 or H_1 is true, based on some data. Typically, a hypothesis test is specified in terms of a *test statistic*, $W(X_1, \ldots, X_n) = W(\boldsymbol{X})$, a function of the data, $\boldsymbol{X}$. The sample space for W is partitioned into the *rejection region*, R, and its complement, the *acceptance region*. If $W \in R$ is observed, the null hypothesis H_0 is rejected, and the decision is made that H_1 is true. If $W \notin R$, H_0 is accepted as true.

For example, in a study to examine the factors that influence a student's success at completing a 4-year college, one factor of interest might be socioeconomic status (SES). If, for simplicity, we just have two groups (high SES and low SES), and the success probabilities of the respective groups are p_h and p_l, then a hypothesis test of interest may be $H_0: p_h \leq p_l$ versus $H_1: p_h > p_l$.

This is an example of a one-sided test, in which the alternative hypothesis specifies a direction. In contrast, we may consider the two-sided test, $H_0: p_h = p_l$ versus $H_1: p_h \neq p_l$, in which no direction is specified in the alternative. If we let $\widehat{p}_h$ and $\widehat{p}_l$ denote the observed success rates in the two SES classes, the one-sided test would reject the null hypothesis if $\widehat{p}_h - \widehat{p}_l$ is big, while the two-sided test would reject the null hypothesis if $\widehat{p}_h - \widehat{p}_l$ is either big or small.

A hypothesis test of $H_0: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta_0^c$ might make one of two types of errors. If $\theta \in \Theta_0$, but the hypothesis test incorrectly decides to reject H_0, then the test has made a *Type I error*. If, on the other hand, $\theta \in \Theta_0^c$, but the test decides to accept H_0, a *Type II error* has been made. In our example of testing $H_0: p_h \leq p_l$ versus $H_1: p_h > p_l$, if we conclude that $p_h > p_l$, but in fact $p_h \leq p_l$, we have made a Type I error.

Constructing Tests

There are many methods of deriving test statistics for a hypothesis test, a few of which follow.

Likelihood Ratio Tests

The likelihood ratio method of hypothesis testing is related to maximum likelihood estimators (MLE) (discussed in *Estimation: Point and Interval*). Suppose that we have a sample, $X_1, \ldots, X_n$, where the X_i are independent, each with distribution given by the probability density or mass function, $f(x|\theta)$. The observed values of these random variables are denoted by $\boldsymbol{x} = (x_1, \ldots, x_n)$, with likelihood function $L(\theta|\boldsymbol{x})$ given by

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

The *likelihood ratio test (LRT) statistic* for testing $H_0: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta_0^c$ is

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\Theta} L(\theta|\boldsymbol{x})}$$

An LRT is a test that has a rejection region of the form $\{\boldsymbol{x}: \lambda(\boldsymbol{x}) \leq k\}$, where k is any number satisfying $0 \leq k \leq 1$.

If we interpret the likelihood function as measuring how likely the values of θ are, then we see that the LRT statistic is comparing the plausibility of the θ values in the null hypothesis with those in the alternative. Small values of the LRT statistic are then interpreted as being evidence against H_0 and lead to rejection of H_0.

If the null hypothesis consists of a single value θ_0, and the alternative is everything else, then the LRT statistic is simply $\lambda(\boldsymbol{x}) = \frac{L(\theta_0|\boldsymbol{x})}{L(\widehat{\theta}|\boldsymbol{x})}$, where $\widehat{\theta}$ is the MLE of θ.

Example. Let $X_1, \ldots, X_n$ be a random sample from a normal population with mean θ and variance 1 ($n(\theta, 1)$ population). The LRT statistic for testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ is

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_0|\boldsymbol{x})}{L(\overline{x}|\boldsymbol{x})} = \frac{(2\pi)^{-\frac{n}{2}} \exp\left[-\sum_{i=1}^{n} \frac{(x_i-\theta_0)^2}{2}\right]}{(2\pi)^{-\frac{n}{2}} \exp\left[-\sum_{i=1}^{n} \frac{(x_i-\overline{x})^2}{2}\right]},$$

where $\overline{x}$ is the sample mean calculated from $\boldsymbol{x}$.

If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then, as with MLE, the LRT statistic is a function of T. That is, $\lambda(\boldsymbol{x})$ depends on $\boldsymbol{x}$ only

International Encyclopedia of the Social & Behavioral Sciences, Second Edition, 2015, 491–493

through $T(\boldsymbol{x})$. In the previous example $T(\boldsymbol{X}) = \overline{X}$ is a sufficient statistic, and the expression for $\lambda(\boldsymbol{x})$ can be simplified to

$$\lambda(\boldsymbol{x}) = \exp\left[-\frac{n(\overline{x}-\theta_0)^2}{2}\right].$$

Bayesian Tests

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' theorem to obtain the posterior distribution, $\pi(\theta|\boldsymbol{x})$. All inferences about θ are now based on the posterior distribution. In a hypothesis-testing problem, the posterior distribution may be used to calculate the probabilities that H_0 and H_1 are true.

One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if

$$\frac{P(\theta\in\Theta_0|\boldsymbol{X})}{P(\theta\in\Theta_0^c|\boldsymbol{X})} > k,$$

for some constant k, and to reject H_0 otherwise. Equivalently, she can reject H_0 if $P(\theta\in\Theta_0^c|\boldsymbol{X}) \geq 1/(1+k)$.

Example. Let $X_1, \ldots, X_n$ be a random sample from an $n(\theta, \sigma^2)$ population, and let the prior distribution on θ be $n(\mu, \tau^2)$, where σ^2, μ, and τ^2 are known. Consider testing $H_0{:}\theta \leq \theta_0$ versus $H_1{:}\ \theta > \theta_0$, and suppose we decide to accept H_0 if $P(\theta\in\Theta_0|\boldsymbol{X}) > P(\theta\in\Theta_0^c|\boldsymbol{X})$. After some calculation, we find that H_0 will be accepted as true if

$$\overline{X} < \theta_0 + \frac{\sigma^2(\theta_0-\mu)}{n\tau^2}.$$

Union–Intersection and Intersection–Union Tests

In some situations, tests for complicated null hypotheses can be developed from tests for simpler null hypotheses. The union–intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, say $H_0{:}\theta\in\bigcap_{\gamma\in\Gamma}\Theta_\gamma$, where Γ is an arbitrary index set. If tests are available for each of the problems of testing $H_{0\gamma}{:}\theta\in\Theta_\gamma$ versus $H_{1\gamma}{:}\theta\in\Theta_\gamma^c$, where the rejection region for the test of $H_{0\gamma}$ is $\{\boldsymbol{x}{:}\ W_\gamma(\boldsymbol{x})\in R_\gamma\}$, then the rejection region for the union–intersection test is

$$\bigcup_{\gamma\in\Gamma}\{x{:}W_\gamma(x)\in R_\gamma\}.$$

The rationale is that if any one of the hypotheses $H_{0\gamma}$ is rejected, then H_0 must be rejected, also.

A complementary method, the intersection–union method, may be useful if the null hypothesis is conveniently expressed as a union. Suppose we wish to test the null hypothesis $H_0{:}\theta\in\bigcup_{\gamma\in\Gamma}\Theta_\gamma$, and $\{\boldsymbol{x}{:}W_\gamma(\boldsymbol{x})\in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma}{:}\theta\in\Theta_\gamma$ versus $H_{1\gamma}{:}\theta\in\Theta_\gamma^c$. Then the rejection region for the intersection–union test of H_0 versus H_1 is

$$\bigcap_{\gamma\in\Gamma}\{\boldsymbol{x}{:}W_\gamma(\boldsymbol{x})\in R_\gamma\},$$

and H_0 is false if and only if all of the $H_{0\gamma}$ are false, so H_0 can be rejected if and only if each of the individual hypotheses $H_{0\gamma}$ can be rejected.

Example. Returning to our example about success in 4-year colleges, we might argue that there are measures of success other than finishing. For example, yearly income could be used to measure success, and denoting mean incomes in the high and low SES groups θ_h and θ_l, respectively, results in the hypothesis test

$$H_0{:}\theta_h \leq \theta_l \text{ or } p_h \leq p_l \quad \text{versus} \quad H_1{:}\theta_h > \theta_l \text{ and } p_h > p_l,$$

where the high group is considered superior only if H_1 is accepted.

If we have estimators $\widehat{\theta}_h$, $\widehat{\theta}_l$, $\widehat{p}_h$, and $\widehat{p}_l$, a rejection region for the intersection–union test could be given by

$$\left\{\left(\widehat{\theta}_h, \widehat{\theta}_l, \widehat{p}_h, \widehat{p}_l\right){:}\widehat{\theta}_h - \widehat{\theta}_l > k_1 \text{ and } \widehat{p}_h - \widehat{p}_l > k_2\right\}.$$

Thus the intersection–union test decides that the high SES group is superior to the low SES group, that is, H_1 is true, if and only if it decides that each of the individual parameters is better in the high group.

Intersection–union tests provide important tools in acceptance sampling, in which one decides whether to accept a product based on a collection of measurements (see Berger, 1982).

There are many other methods available for constructing hypothesis tests, methods based on invariance, pivots, robust, or large-sample arguments, to name a few. For more on hypothesis testing see Lehmann (1986).

Evaluating Tests

As mentioned previously, a hypothesis test might make one of two different kinds of errors. The probabilities of making these errors can be calculated using the *power function*. If R denotes the rejection region for a test, the *power function* is $\beta(\theta) = P_\theta(\boldsymbol{X}\in R)$. If $\theta\in\Theta_0$, $\beta(\theta)$ is the probability of a Type I error. If $\theta\in\Theta_0^c$, $\beta(\theta)$ is one minus the probability of a Type II error. Hence, a good test has power function near one for most $\theta\in\Theta_0^c$ and near zero for most $\theta\in\Theta_0$.

Example. Let $X_1, \ldots, X_n$ be a random sample from an $n(\theta, \sigma^2)$ population, σ^2 known. The LRT of $H_0{:}\theta \leq \theta_0$ versus $H_1{:}\theta > \theta_0$ rejects H_0 if $\sqrt{n}(\overline{X}-\theta_0)/\sigma \geq k$ and has power function

$$\beta(\theta) = P\left(Z \geq k + \frac{\theta_0-\theta}{\sigma/\sqrt{n}}\right),$$

where Z is a standard normal random variable.

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the *size* $\left(\alpha = \sup_{\theta\in\Theta_0} P_\theta(\boldsymbol{X}\in R)\right)$ of the test and the decision to reject H_0 or accept H_0. The size of the test carries important information. If α is small, the decision to reject H_0 is fairly convincing, but if α is large, the decision to reject H_0 is not very convincing, because the test has a large probability of incorrectly making that decision. A common practice is to choose the desired small value of α (often 0.05 or 0.01) and then to choose the critical value, k in the previous example, so the resulting test has the specified size, α.

Another way of reporting the results of a hypothesis test, one that is data dependent, is to report the *p-value*. Typically, not one but an entire class of tests is constructed, a different test being defined for each value of α. The p-value for the sample point $\boldsymbol{x}$ is the smallest value of α for which this sample point will lead to rejection of H_0 (*see* Significance, Tests of). Because rejection of

H_0 using a test with small size is more convincing evidence that H_1 is true than rejection of H_0 with a test with large size, the interpretation of p-values goes in the same way. The smaller the p-value, the stronger the sample evidence that H_1 is true.

In the previous example, if data $\boldsymbol{x}$ with sample mean $\bar{x}$ is observed, the p-value is

$$p(\boldsymbol{x}) = P\left(Z \geq \frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}\right).$$

Many other types of evaluations of tests can be done. The theory of most powerful tests shows how to construct best tests under a variety of conditions (see Lehmann, 1986 or Casella and Berger, 2001: Chapter 8). Hypothesis tests can also be evaluated using risk functions as in Hwang et al. (1992).

Asymptotics

For the LRT statistic, the following general theorem allows the construction of a large-sample test.

Theorem 1. *Let $X_1, \ldots, X_n$ be a random sample from a probability density function or probability mass function $f(\boldsymbol{x}|\theta)$. Under some regularity conditions on $f(\boldsymbol{x}|\theta)$, if $\theta \in \Theta_0$ then the distribution of the statistic $-2\log\lambda(\boldsymbol{x})$ converges to a chi-squared distribution as the sample size $n \to \infty$. The degrees of freedom of the limiting chi-squared distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$.*

The 'regularity conditions' are concerned mainly with the existence and behavior of the derivatives (with respect to the parameter) of the likelihood function, and the support of the distribution (it cannot depend on the parameter). See Casella and Berger (2001: Section 10.6) or Lehmann (1986: Section 8.8) for precise conditions.

Rejection of H_0 for small values of $\lambda(\boldsymbol{x})$ is equivalent to rejection for large values of $-2\log\lambda(\boldsymbol{x})$. Thus, by Theorem 1 the test that rejects H_0 if and only if $-2\log\lambda(\boldsymbol{x}) \geq \chi^2_{\nu,\alpha}$, where ν is the degrees of freedom specified in the theorem, is a test with size approximately equal to α, if the sample size is large.

As might be expected, Theorem 1 has wide applicability. In particular, it is extremely useful in categorical data analysis (*see* Multivariate Analysis: Discrete Variables (Loglinear Models)).

Other large-sample test constructions are based on asymptotic normality of a point estimator (*see* Estimation: Point and Interval). Suppose we wish to test a hypothesis about a real-valued parameter θ, and $W_n = W(X_1, \ldots, X_n)$ is a point estimator of θ based on a sample of size n that satisfies

$$\frac{W_n - \theta}{\sqrt{\mathrm{Var}\, W_n}} \to Z$$

as $n \to \infty$, where $\mathrm{Var}W_n$ is the variance of W_n and Z is a standard normal random variable. We now have the basis for an approximate test, for example, we would reject $H_0{:}\theta \leq \theta_0$ at level 0.05 if $(W_n - \theta_0)/\sqrt{\mathrm{Var}\, W_n} \geq 1.645$. Note, $\mathrm{Var}\, W_n$ could depend on θ_0, and we can still use it in the test statistic. This type of test, where we use the actual variance of W_n, is called a *score test*.

If $\mathrm{Var}\, W_n$ also depends on unknown parameters, we could look for an estimate S_n^2 of $\mathrm{Var}\, W_n$ with the property that $(\mathrm{Var}\, W_n)/S_n^2$ converges in probability to one. Then, using Slutsky's theorem (see Casella and Berger, 2001: Section 5.5), we can deduce that $(W_n - \theta)/S_n$ converges in distribution to a standard normal distribution, also. The large-sample test based on this statistic is called a *Wald test*.

More details about asymptotic tests may be found in Boos and Stefanski (2013) and Lehmann (1999).

Conclusions

Hypothesis testing is one of the most widely used, and some may say abused, methodologies in statistics. Formally, the hypotheses are specified, an α-level is chosen, a test statistic is calculated, and it is reported whether H_0 or H_1 is accepted. In practice, it may happen that hypotheses are suggested by the data, the choice of α-level may be ignored, more than one test statistic is calculated, and many modifications to the formal procedure may be made. Most of these modifications cause bias and can invalidate the method. For example, a hypothesis suggested by the data is likely to be one that has 'stood out' for some reason, and, hence, H_1 is likely to be accepted, unless the bias is corrected for (using something like Scheffe's method – see Hsu, 1996).

Perhaps the most serious criticism of hypothesis testing is the fact that, formally, it can only be reported that either H_0 or H_1 is accepted at the prechosen α-level. Thus, the same conclusion is reached if the test statistic only barely rejects H_0 and if it rejects H_0 resoundingly. Many feel that this is important information that should be reported, and thus it is commonplace to report the p-value of the hypothesis test, also.

For further details on hypothesis testing see the classic book by Lehmann (1986). Introductions are provided by Casella and Berger (2001) or Schervish (1995), also, and a good introduction to multiple comparisons is Hsu (1996); *see* Hypothesis Tests, Multiplicity of.

Acknowledgment

This article is a revised version of an article on the same topic in the first edition of the *International Encyclopedia of the Social and Behavioral Sciences*, which was coauthored with the late George Casella, University of Florida.

See also: Bayesian Statistics; Hypothesis Testing: Methodology and Limitations; Hypothesis Tests, Multiplicity of; Likelihood in Statistics; Linear Hypothesis; Model Testing and Selection, Theory of; Multiple Comparisons, Statistics of; Significance, Tests of; Statistical Sufficiency.

Bibliography

Berger, R.L., 1982. Multiparameter hypothesis testing and acceptance sampling. Technometrics 24, 295–300.

Boos, D.D., Stefanski, L.A., 2013. Essential Statistical Inference, Theory and Methods. Springer, New York.

Casella, G., Berger, R.L., 2001. Statistical Inference, second ed. Wordsworth/Brooks Cole, Pacific Grove, CA.

Hsu, J.C., 1996. Multiple Comparisons, Theory and Methods. Chapman & Hall, London.

Hwang, J.T., Casella, G., Robert, C., Wells, M.T., Farrell, R.H., 1992. Estimation of accuracy in testing. Annals of Statistics 20, 490–509.

Lehmann, E.L., 1986. Testing Statistical Hypotheses, second ed. Springer, New York.

Lehmann, E.L., 1999. Elements of Large-Sample Theory. Springer-Verlag, New York.

Schervish, M., 1995. Theory of Statistics. Springer, New York.