"Hypothesis Testing"
An Article for the International Encyclopedia of the Social and Behavioral Sciences

George Casella*
Cornell University

Roger L. Berger
North Carolina State University

September 23, 1999

Introduction

A *hypothesis* is a statement about a population parameter, 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. Typically, a hypothesis test is specified in terms of a *test statistic* $W(X_1, \ldots, X_n) = W(\mathbf{X})$, a function of the sample. For example, a test might specify that H_0 is to be rejected if $\overline{X}$, the sample mean, is greater than 3. In this case $W(\mathbf{X}) = \overline{X}$ is the test statistic and the rejection region is $\{(x_1, \ldots, x_n) : \overline{x} > 3\}$.

Constructing Tests

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

1. *Likelihood Ratio Tests*

   The likelihood ratio method of hypothesis testing is related to the maximum likelihood estimators discussed in the article on point and

*Supported by National Science Foundation Grant DMS-9971586. Email: gc15@cornell.edu. This is technical report BU-1454-M in the Department of Biometrics, Cornell University, Ithaca, NY 14853.

interval estimation. Given a likelihood function $L(\theta|\mathbf{x})$, the *likelihood ratio test statistic* for testing $H_0: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}. \tag{1}$$

A *likelihood ratio test* (LRT) is any test that has a rejection region of the form $\{\mathbf{x}: \lambda(\mathbf{x}) \square c\}$, where c is any number satisfying $0 \square c \square 1$.

If we interpret the likelihood function as measuring how likely the values of θ are, then we see that the LRT is comparing the plausibility of the θ values in the null hypothesis to those in the alternative. Small values of the LRT statistic are 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 $\lambda = L(\theta_0|\mathbf{x})/L(\hat{\theta}|\mathbf{x})$, where $\hat{\theta}$ is the MLE of θ.

Example Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\theta, 1)$ population. The LRT statistic for testing $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ is

$$\lambda(\mathbf{x}) = \frac{L(\theta_0|\mathbf{x})}{L(\bar{x}|\mathbf{x})} = \frac{(2\pi)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \theta_0)^2/2\right]}{(2\pi)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \bar{x})^2/2\right]}.$$

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

2. *Bayesian Tests*

The Bayesian paradigm prescribes that the sample information be combined with the prior information using Bayes' Theorem to obtain the posterior distribution $\pi(\theta|\mathbf{x})$. 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 $\frac{P(\theta \in \Theta_0|\mathbf{X})}{P(\theta \in \Theta_0^c|\mathbf{X})} \geq c$ for some constant c, and to reject H_0 otherwise. Equivalently, we can reject H_0 if $P(\theta \in \Theta_0^c|\mathbf{X})$ is greater than a specified number.

Example Let $X_1, \ldots, X_n$ be iid $\mathcal{N}(\theta, \sigma^2)$ and let the prior distribution on θ be $\mathcal{N}(\mu, \tau^2)$ where σ^2, μ, and τ^2 are known. Consider testing $H_0 : \theta \square \theta_0$ versus $H_1 : \theta > \theta_0$ where we decide to accept H_0 if $P(\theta \in \Theta_0 | \mathbf{X}) \geq P(\theta \in \Theta_0^c | \mathbf{X})$. After some calculation, we find that H_0 will be accepted as true if

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

3. *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 $H_0 : \quad \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma$, 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 $\{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}$, then the rejection region for the union–intersection test is

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

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

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 $H_0 : \quad \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma$, and $\{\mathbf{x} : T_\gamma(\mathbf{x}) \in R_\gamma\}$ 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

$$\bigcap_{\gamma \in \Gamma} \{x : T_\gamma(x) \in R_\gamma\}. \tag{2}$$

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 The topic of acceptance sampling provides an extremely useful application of an intersection–union test (see Berger 1982).

In assessing the quality of upholstery fabric, standards dictate that parameters relating to strength and flammability must satisfy $\theta_1 > 50$

pounds and $\theta_2 > .95$, respectively. This results in the hypothesis test

$$H_0 : \{\theta_1 \square 50 \text{ or } \theta_2 \square .95\} \quad \text{versus} \quad H_1 : \{\theta_1 > 50 \text{ and } \theta_2 > .95\},$$

where a batch of material is acceptable only if H_1 is accepted.

If $X_1, \ldots, X_n$ are iid $\mathcal{N}(\theta_1, \sigma^2)$ and $Y_1, \ldots, Y_m$ are iid Bernoulli(θ_2), where $Y_i = 1$ if the ith sample passes the flammability test, the rejection region for the intersection–union test is given by

$$\left\{(\mathbf{x}, \mathbf{y}) : \frac{\overline{x} - 50}{s/\sqrt{n}} > t \text{ and } \sum_{i=1}^{m} y_i > b\right\}.$$

Thus the intersection–union test decides the product is acceptable, that is, H_1 is true, if and only if it decides that each of the individual parameters meets its standard.

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

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.

If R denotes the rejection region for a test, the *power function* is

$$P_\theta(\mathbf{X} \in R) = \begin{cases} \text{probability of a Type I Error,} & \text{if } \theta \in \Theta_0, \\ 1- \text{ the probability of a Type II Error,} & \text{if } \theta \in \Theta_0^c. \end{cases}$$

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 a $\mathcal{N}(\theta, \sigma^2)$ population, σ^2 known. The likelihood ratio test of $H_0 : \theta \square \theta_0$ versus $H_1 : \theta > \theta_0$ rejects H_0 if $(\overline{X} - \theta_0)/(\sigma/\sqrt{n}) > c$ and has power function

$$P_\theta(\mathbf{X} \in R) = P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right),$$

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 ($\sup_{\theta\in\Theta_0} P_\theta(X \in R)$), α, of the test used 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.

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 are constructed, a different test being defined for each value of α. The p-value for the sample point $\mathbf{x}$ is the smallest value of α for which this sample point will lead to rejection of H_0.

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.

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 1990, Chapter 8). Hypothesis tests can also be evaluated using risk functions, as in Hwang *et al.* (1992).

Asymptotics

For the *LRT* statistic (1), the following general theorem allows us to ensure construct a large sample test.

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

Rejection of $H_0\colon \theta \in \Theta_0$ for small values of $\lambda(\mathbf{X})$ is equivalent to rejection for large values of $-2\log\lambda(\mathbf{X})$. Thus,

$$H_0 \text{ is rejected if and only if } -2\log\lambda(\mathbf{X}) \geq \chi^2_{\nu,\alpha},$$

[1]The "regularity conditions" are mainly concerned 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 Lehmann (1986, Section 8.8) for precise conditions.

where ν is the degrees of freedom specified in Theorem 1.

Another large-sample test construction is based on asymptotic normality of a point estimator. 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

$$\frac{W_n - \theta}{\sigma_n} \to Z,$$

where σ_n^2 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 could reject $H_0 : \theta \square \theta_0$ at level .05 if $(W_n - \theta_0)/\sigma_n > 1.645$.

In some instances, σ_n also depends on unknown parameters. In such a case, we look for an estimate S_n of σ_n with the property that σ_n/S_n converges in probability to one. Then, using Slutsky's Theorem (see Casella and Berger 1990, Section 5.3), we can deduce that $(W_n - \theta)/S_n$ also converges in distribution to a standard normal distribution. A large-sample test may be based on this fact. Whether σ_n is estimated assuming $\theta = \theta_0$, or not, can lead to *score* and *Wald tests*, respectively.

References

1. Berger, R. L. (1982). Multiparameter Hypothesis Testing and Acceptance Sampling. *Technometrics* 24, 295–300.

2. Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Pacific Grove, CA: Wadsworth/Brooks Cole.

3. Hwang, J. T., Casella, G., Robert, C., Wells, M. T. and Farrell, R. H. (1992). Estimation of accuracy in testing. *Ann. Statist.* **20**, 490–509.

4. Lehmann, E. L. (1986). *Testing Statistical Hypotheses, Second Edition.* New York: Springer-Verlag