of; Neuroendocrinology; Olfactory System; Sex Hormones and their Brain Receptors; Sexual Behavior and Maternal Functions, Neurobiology of; Thirst and Drinking

Bibliography

Bligh J 1966 The thermosensitivity of the hypothalamus and thermoregulation in mammals. *Biological Reviews* **41**: 317–67

Freeman M E 1995 The hypothalamus. In: Conn P M (ed.) *Neuroscience in Medicine*. J. B. Lippincott Company, New York

Gorski R A, Gordon J H, Shryne J E, Southam A M 1978 Evidence for a morphological sex difference within the medial preoptic area of the rat brain. *Brain Research* **148**: 333–46

Hakansson M L, Brown H, Ghilardi N, Skoda R C, Meister B 1998 Leptin receptor immunoreactivity in chemical defined target neurons of the hypothalamus. *Journal of Neuroscience* **18**: 559–72

Hess W R 1928 Stammganglien-Reizversuche. *Berichte der gesamten Physiologie* **42**: 554–5

Numan M 1978 Medial preoptic area and maternal behavior in the female rat. *Journal of Comparative and Physiological Psychology* **87**: 746–59

Risold P Y, Thompson R H, Swanson L W 1997 The structural organization of connections between the hypothalamus and cerebral cortex. *Brain Research Reviews* **24**: 197–254

Satinoff E 1978 Neural organization and evolution of thermal regulation in mammals. *Science* **201**: 16–22

Sladek J R, Sladek C D 1990 Morphology of the endocrine brain, hypothalamus, and neurohypophysis. In: Becker K L (ed.) *Principles and Practice of Endocrinology and Metabolism*. J. B. Lippincott Company, New York

J. Hennig

Hypothesis Testing in Statistics

A *statistical 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. 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 four-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 $H_0: p_h = p_l$ vs. $H_1: p_h > p_l$.

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 $H_0: p_h = p_l$ vs. $H_1: p_h \neq p_l$ in which no direction is specified in the alternative. If we let $\hat{p}_h$ and $\hat{p}_l$ denote the observed success rates in the two SES classes, the one-sided test would reject the null hypothesis if $\hat{p}_h - \hat{p}_l$ is big, while the two-sided test would reject the null hypothesis if $\hat{p}_h - \hat{p}_l$ 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, in testing $H_0: p_h = p_l$ versus $H_1: p_h > p_l$, if we conclude that $p_h > p_l$ but, in fact, $p_h \leqslant p_l$ we have made a Type I error.

1. Constructing Tests

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

1.1 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 $f(x \mid \theta)$. The observed values of these random variables are denoted by $\mathbf{x} = (x_1, \ldots, x_n)$, with likelihood function $L(\theta \mid \mathbf{x})$ given by

$$L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{1}$$

The *likelihood ratio test statistic* for testing $H_0: \theta \in \Theta_0$ vs. $H_1: \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta \mid \mathbf{x})}{\sup_{\Theta_0} L(\theta \mid \mathbf{x})} \tag{2}$$

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

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 with those in the alternative. Small values of the LRT statistic are then interpreted as being evidence against H_0, and lead to rejection.

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 $\mathrm{n}(\theta, 1)$ population. The LRT statistic for testing H_0: $\theta = \theta_0$ versus $H_1: \theta \neq \theta_0$ is

$$\lambda(\mathbf{x}) = \frac{L(\theta_0 | \mathbf{x})}{L(\bar{x} | \mathbf{x})} = \frac{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n} (x_i - \theta_0)^2/2]}{(2\pi)^{-n/2} \exp[-\sum_{i=1}^{n} (x_i - \bar{x})^2/2]} \tag{3}$$

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 $T(\mathbf{x})$.

1.2 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

$$\frac{P(\theta \in \Theta_0 | \mathbf{X})}{P(\theta \in \Theta_0^c | \mathbf{X})} \geqslant k \tag{4}$$

for some constant k, 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 independent identically distributed (i.i.d.) $\mathrm{n}(\theta, \sigma^2)$ and let the prior distribution on θ be $\mathrm{n}(\mu, \tau^2)$ where σ^2, μ, and τ^2 are known. Consider testing $H_0: \theta \leqslant \theta_0$ vs. $H_1: \theta > \theta_0$ where we decide to accept H_0 if $P(\theta \in \Theta_0 | \mathbf{X}) \geqslant P(\theta \in \Theta_0^c | \mathbf{X})$. After some calculation, we find that H_0 will be accepted as true if

$$\bar{X} \leqslant \theta_0 + \frac{\sigma^2(\theta_0 - \mu)}{n\tau^2} \tag{5}$$

1.3 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 $H_0: \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma$, where Γ is an arbitrary index set. If tests are available for each of the problems of testing $H_{0\gamma}: \theta \in \Theta_\gamma$ vs. $H_{1\gamma}: \theta \in \Theta_\gamma^c$ 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

$$\bigcup_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\} \tag{6}$$

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 $H_0: \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma$, and $\{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma}: \theta \in \Theta_\gamma$ vs. $H_{1\gamma}: \theta \in \Theta_\gamma^c$. Then the rejection region for the intersection–union test of H_0 versus H_1 is

$$\bigcap_{\gamma \in \Gamma} \{\mathbf{x}: T_\gamma(\mathbf{x}) \in R_\gamma\} \tag{7}$$

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 four-year schools, 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 by θ_h and θ_l respectively results in the hypothesis test

$$H_0: \theta_\mathrm{h} \leqslant \theta_\mathrm{l} \text{ or } p_\mathrm{h} \leqslant p_\mathrm{l} \quad \text{vs.} \quad H_1: \theta_\mathrm{h} > \theta_\mathrm{l} \text{ and } p_\mathrm{h} > p_\mathrm{l} \tag{8}$$

where the groups are considered different only if H_1 is accepted.

If we have estimators $\hat{\theta}_\mathrm{h}$, $\hat{\theta}_\mathrm{l}$, $\hat{p}_\mathrm{h}$ and $\hat{p}_\mathrm{l}$, a rejection region for the intersection–union test could be given by

$$\{(\hat{\theta}_\mathrm{h}, \hat{\theta}_\mathrm{l}, \hat{p}_\mathrm{h}, \hat{p}_\mathrm{l}): \hat{\theta}_\mathrm{h} - \hat{\theta}_\mathrm{l} > k_1 \quad \text{and} \quad \hat{p}_\mathrm{h} - \hat{p}_\mathrm{l} > k_2\} \tag{9}$$

Thus the intersection–union test decides that the high and low SES groups have different levels of success, that is, H_1 is true, if and only if it decides that each of the individual parameters are different.

The topic of acceptance sampling, in which one tests whether to accept a product based on a collection of

measurements, provides an extremely useful application of an intersection–union test (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).

2. 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

$$P_\theta(\mathbf{X} \in R) = \begin{cases} \text{probability of a Type I error} & \text{if } \theta \in \Theta_0 \\ 1 - \text{the probability of a Type II error} & \text{if } \theta \in \Theta_0^c \end{cases} \tag{10}$$

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 $\mathrm{n}(\theta, \sigma^2)$ population, σ^2 known. The likelihood ratio test of $H_0: \theta \leqslant \theta_0$ vs. $H_1: \theta > \theta_0$ rejects H_0 if $(\bar{X} - \theta_0)/(\sigma/\sqrt{n}) > k$ and has power function

$$P_\theta(\mathbf{X} \in R) = P\left(Z > k + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}} \right) \tag{11}$$

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 (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.

Many other types of evaluations of test 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, Chap. 8). Hypothesis tests can also be evaluated using risk functions (see *Decision Theory: Classical*), as in Hwang et al. (1992).

3. Asymptotics

For the likelihood ratio test statistic (2), the following general theorem allow us to construct a large sample test.

Theorem 1. *Let $X_1, \ldots, X_n$ be a random sample from a probability distribution function or probability mean function $f(x \mid \theta)$. Under some regularity conditions on the model $f(x \mid \theta)$, 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 $\theta \in \Theta$.*

(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, Sect. 10.6) or Lehmann (1986, Sect. 8.8) for precise conditions.)

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

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

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

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

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

where 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 $H_0: \theta \leqslant \theta_0$ at level 0.05 if $(W_n - \theta_0)/\sqrt{\text{Var}\, W_n} > 1.645$. Note that 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 Var W_n also depends on unknown parameters we could look for an estimate S_n^2 of Var W_n with the property that $(\text{Var}\, W_n)/S_n^2$ converges in probability to one. Then, using Slutsky's Theorem (see Casella and Berger 2001, Sect. 5.5), we can deduce that $(W_n - \theta)/S_n$ also converges in distribution to a standard normal distribution. The large-sample test based on this statistic is called a *Wald test*.

4. 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 almost required to also report the p-value of the hypothesis test.

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

See also: Explanation: Conceptions in the Social Sciences; Hypothesis Testing: Methodology and Limitations

Bibliography

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

Casella G, Berger R L 2001 *Statistical Inference*, 2nd edn. 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*, 2nd edn. Springer, New York

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

G. Casella and R. L. Berger

Hypothesis Testing: Methodology and Limitations

Hypothesis tests are part of the basic methodological toolkit of social and behavioral scientists. The philosophical and practical debates underlying their application are, however, often neglected. The fruitful application of hypothesis testing can benefit from a clear insight into, the underlying concepts and their limitations.

1. The Idea of Hypothesis Testing

A test is a statistical procedure to obtain a statement on the truth of falsity of a proposition, on the basis of empirical evidence. This is done within the context of a model, in which the fallibility or variability of this empirical evidence is represented by probability. In this model, the evidence is summarized in observed data, which is assumed to be the outcome of a stochastic, i.e., probabilistic, process; the tested proposition is represented as a property of the probability distribution of the observed data.

1.1 Some History

The first published statistical test was by John Arbuthnot in 1710, who wondered about the fact that in human births, the fraction of boys born year after year appears to be slightly larger than the fraction of girls (cf. Hacking 1965). He calculated that this empirical fact would be exceedingly unlikely (he obtained a probability of 1/483 600 000 000 000 000 000 000) if the probability of a male birth were exactly 0.5, and argued that this was a proof of divine providence, since boys—some of whom will be soldiers—have a higher risk of an early death, so that a higher ratio of male births is needed to obtain an equal ratio of males among young adults. We see here the basic elements of a test: the proposition

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7