Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA
Spring, 2001

March 27, 2001

Week 11. Testing Statistical Hypotheses

1. Statistical Hypothoses

Up to now we've been looking exclusively at the problem of *estimation*, of trying to identify the value $\theta \in \Theta$ of an uncertain parameter on the basis of an observation $x \in \mathcal{X}$ of a random vector from some probability distribution $x \sim f(x \mid \theta)$ that depends on θ. Today we begin a new quest: again on the basis of an observed value $x \sim f(x \mid \theta)$, we seek to discover whether an assertion about θ is true or false.

We can think about hypotheses (or "assertions" about θ) simply as subsets $H_0 \subset \Theta$, interpreted as the set of $\theta \in \Theta$ for which the assertion or hypothesis is true; thus we would like to discover, on the basis of an observed value $x \in \mathcal{X}$, whether or not $\theta \in H_0$. We denote the complementary set by $H_1 = (H_0)^c = \{\theta \in \Theta : \theta \notin H_0\}$, and call it the *alternate hypothesis*.

There are many approaches to the problem of testing hypotheses; we will consider the Frequentist approach, in both the original fixed-level version and a variant (reporting P-values), and also the Bayesian approach, in both the decision-theoretic and posterior probability versions.

2. Fixed-level Frequentist

If we feel constrained to answer the question "Is $\theta \in H_0$?" with a simple *yes* or *no*, then we may also divide $\mathcal{X}$ up into two sets, the "rejection region" or "critical region" $\mathcal{R} \subset \mathcal{X}$ of those possible outcomes $x \in \mathcal{X}$ for which we will conclude that H_0 is false, and its complement $R^c = \mathcal{X} \backslash R$, the outcomes that will not lead us to reject H_0.

Notice that there are four possibilities:

1. Reject a False Hypothesis: $\theta \in H_1$, and $x \in \mathcal{R}$;
2. Reject a True Hypothesis: $\theta \in H_0$, and $x \in \mathcal{R}$;
3. Accept a False Hypothesis: $\theta \in H_1$, and $x \notin \mathcal{R}$;
4. Accept a True Hypothesis: $\theta \in H_0$, and $x \notin \mathcal{R}$.

Two of these are good (1 and 4) in that we correctly discover the truthity or falsity of the hypothesis; the other two constitute distinct errors, traditionally called a **Type I** error (2: rejecting a true hypotheis) and a **Type II** error (3: failing to reject a false hypotheis).

We can quantify the performance of a test that rejects whenever $x \in \mathcal{R}$ through the function

$$\mathrm{pow}(\theta) \equiv \mathsf{P}[X \in \mathcal{R} \mid \theta],$$

the probabilty of rejection (as a function of $\theta \in \Theta$). Evidently we would like $\mathrm{pow}(\theta)$ to be small for $\theta \in H_0$ and large for $\theta \notin H_0$; with that in mind, define

$$\begin{array}{rclcl} \alpha & = & \sup_{\theta \in H_0} \ \mathsf{P}[X \in \mathcal{R} \mid \theta] & = & \sup_{\theta \in H_0} \ \mathrm{pow}(\theta) \\ \beta & = & \sup_{\theta \notin H_0} \ \mathsf{P}[X \notin \mathcal{R} \mid \theta] & = & \sup_{\theta \notin H_0} \ 1 - \mathrm{pow}(\theta) \end{array}$$

The first of these, α, is the (maximum) probability of a Type-I error; it is called the *size* of the test. The second, β, is the (maximum) probability of a Type-II error. Evidently α will be made small if we let $\mathcal{R}$ be small, but only at the expense of making β larger; designing a hypothesis test involves a compromise.

2.1. Example

Think of a jury's choice in a criminal trial: the two possible errors of conficting an innocent defendant, or of acquitting a guilty one, are quite different, with different consequences for society; the US legal system is designed to minimize the chance for the first of these errors, even at the expense of increasing the second.

The traditional Frequentist approach constructs a suitable rejection region $\mathcal{R}$, as follows. First select some test statistic $T : \mathcal{X} \to \mathbb{R}$ intended to discriminate between H_0 and H_1 by taking larger values typically for $\theta \in H_1$ than for $\theta \in H_0$. For any number $t_c \in \mathbb{R}$ we can then construct a rejection region by $\mathcal{R} = \{x \in \mathcal{X} : T(x) \geq t_c\}$, i.e., we reject H_0 in favor of H_1 when we observe $T(X) \geq t_c$. The *size* of the test (the largest possible probability of a Type-I error) is

$$\alpha = \sup_{\theta \in H_0} \mathsf{P}[x \in \mathcal{R}] = \sup_{\theta \in H_0} \mathsf{P}[T \geq t_c \mid \theta],$$

directly related to t_c. The usual approach is to begin with a value in mind for α (often $\alpha = .05$ or $\alpha = .01$) and then adjust t_c to obtain a critical rejection region satisfying $\mathsf{P}[x \in \mathcal{R} \mid \theta] \leq \alpha$ for each $\theta \in H_0$, i.e., set

$$t_c \equiv \inf_{t<\infty} \{t : \sup_{\theta \in H_0} \mathsf{P}[T \geq t \mid \theta] \leq \alpha\}.$$

2.2. Example

In a digital signal processing example, the signal θ is known to be either plus one or minus one; we observe this signal with noise, however, so $X \sim \mathsf{No}(\theta, 1/4)$. For any $t \in \mathbb{R}$ the rule that rejects whenever $X > t$ will do so with probability

$$\mathrm{pow}(\theta) = \mathsf{P}[X > t \mid X \sim \mathsf{No}(\theta, 1/4)] = \Phi\big(2(\theta - t)\big),$$

so a test at level $\alpha = \mathrm{pow}(-1) = 0.05$ would reject whenever $X > -0.1776$, while a test at level $\alpha = 0.01$ would reject whenever $X > 0.1632$. The symmetric rule would be to reject $H_0 : \theta = -1$ whenever $X \geq 0$; this test has size $\alpha = \Phi(-2) = 0.02275$. An observed value of $X = 0.50$ would be rejected, for example, which seems sensible since $\mathsf{P}[X \leq 0.50 \mid \theta = +1] = \Phi(-1) = 0.1587$, so $\theta = +1$ is perfectly plausible, while $\mathsf{P}[X \geq 0.50 \mid \theta = -1] = \Phi(-3) = 0.0013499$, a minor miracle.

BUT what if we observe $X = 0$? The rule says to reject, at level $\alpha = .02275$, but here the evidence against H_0 is ambivolent ($X = 0$ offers evidence against H_1 just as strong as the evidence against H_0), in stark contrast to $X = 0.5$. For this reason, many classical statisticians prefer:

3. Frequentist p-value

A variation on the fixed-level testing introduced above is the reporting of "p-values," or "observed significance levels." The p-value is defined to be:

> The maximum probability of observing evidence *against* the null hypothesis H_0 at least as strong as that actually observed, *if* the null hypothesis is true.

Evidently the definition begs the question of just what constitutes "strong evidence" against H_0. To implement the definition we once again begin by selecting a test statistic $T : \mathcal{X} \to \mathbb{R}$ intended to discriminate between H_0 and H_1 by taking larger values typically for $\theta \in H_1$ than for $\theta \in H_0$, but now calculate for the observed value $t = T(X)$

$$p = \sup_{\theta \in H_0} \Pr[T \geq t \mid \theta].$$

In the Example above, the p-value is simply

$$p = \Pr[X \geq x \mid \theta = -1] = \Phi(-2 - 2x)$$

which at least distinguishes the level of evidence, although it still does not balance this evidence against that for H_1.

In this example it was clear which outcomes constituted "stronger evidence" against H_0; it is always easy whenever both H_0 and H_1 are "point hypotheses" that specify θ exactly (say, as θ_0 and θ_1, respectively), for the **Neyman-Pearson** lemma then asserts that the "best" statistic T to use is (any monotone function of) the **likelihood ratio**

$$\lambda = \frac{f(x \mid \theta_1)}{f(x \mid \theta_0)}$$

or, for n *i.i.d.* observations from the same density,

$$\lambda_n = \frac{f(x_1 \mid \theta_1) \cdots f(x_n \mid \theta_1)}{f(x_1 \mid \theta_0) \cdots f(x_n \mid \theta_0)}$$

In the Normal example of $H_0 : X \sim \mathsf{No}(\mu_0, \sigma^2)$ *vs.* $H_1 : X \sim \mathsf{No}(\mu_1, \sigma^2)$, for example,

$$\frac{f(x \mid H_1)}{f(x \mid H_0)} = \frac{e^{-(x-\mu_1)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}}{e^{-(x-\mu_0)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}} \propto e^{x(\mu_1-\mu_0)/\sigma^2},$$

so $T(x) = \operatorname{sgn}(\mu_1 - \mu_0) \cdot \sum X_i$ will do (or $T(x) = x$ in our digital signal processing example where $\mu_0 < \mu_1$ and $n = 1$).

The likelihood ratio is "best" in the sense that it maximizes the *power*

$$\operatorname{pow}(\theta) = \mathsf{P}[T \geq t_c \mid \theta]$$

for $\theta \notin H_0$, over all possible statistics T; the Neyman-Pearson lemma expresses formally the simple idea that the best way to separate the space $\mathcal{X}$ of possible outcomes into those $\mathcal{R} \subset \mathcal{X}$ where we *reject* H_0 and those $\mathcal{R}^c \subset \mathcal{X}$ where we *do not reject*, is to sort $\mathcal{X}$ on the basis of the likelihood ratio $\lambda = f(x \mid \theta_1)/f(x \mid \theta_0)$. When either or both of H_0 and H_1 is composite, it is common to use the likelihood ratio defined by

$$\lambda = \frac{\sup_{\theta_1 \in H_1} f(x \mid \theta_1)}{\sup_{\theta_0 \in H_0} f(x \mid \theta_0)}$$

or, if for some statistic $T(x)$ the likelihood ratio $f(x \mid \theta_1)/f(x \mid \theta_0)$ is always a monotone function of $T(x)$ for all $\theta_0 \in H_0$ and $\theta_1 \in H_1$, to use $T(x)$.

3.1. Examples

In the normal-mean testing example of $H_0 : X \sim \mathsf{No}(\mu_0, \sigma^2)$ *vs.* $H_1 : X \sim \mathsf{No}(\mu, \sigma^2)$, with σ^2 and μ unspecified, the likelihood ratio becomes

$$\begin{aligned}
\lambda &= \frac{\sup_{\mu\in\mathbb{R},\ \sigma^2>0} (2\pi\sigma^2)^{-n/2} \exp\big(-n[S^2 + (\bar{x}-\mu)^2]/2\sigma^2\big)}{\sup_{\sigma^2>0} (2\pi\sigma^2)^{-n/2} \exp\big(-n[S^2 + (\bar{x}-\mu_0)^2]/2\sigma^2\big)} \\
&= \frac{(2\pi S^2)^{-n/2} e^{-n/2}}{(2\pi[S^2 + (\bar{x}-\mu_0)^2])^{-n/2} e^{-n/2}} \\
&= [1 + t^2]^{n/2},
\end{aligned}$$

where $t \equiv (\bar{x} - \mu_0)/S$ is Student's t statistic; evidently Gossett's Student t-test is just the likelihood ratio test for H_0.

4. To Test or Estimate?

One way to implement a classical test at level α of the hypothesis $H_0 : \mu = \mu_0$ with two-sided alternative $H_1 : \mu \neq \mu_0$ for normally-distributed data $x_i \sim \mathsf{No}(\mu, \sigma^2)$ would be to construct a $(1-\alpha)$-confidence interval $[L_x, R_x]$ and reject H_0 if $\mu_0 \neq [L_x, R_x]$; this is *exactly* equivalent to the likelihood ratio test. Similarly the usual one-sided tests may be implemented by first constructing one-sided confidence intervals $(-\infty, R_x]$ and $[L_x, \infty]$ and checking to see whether or not the hypothesized μ_0 lies inside them.

This offers an important advantage over both fixed-level and p-value testing—it gives a meaningful quantitative report of how "wrong" a rejected hypothesis might be, when H_0 is rejected, or of how close the test came to rejecting, when H_0 is not rejected.

5. Posterior Probabilities and Bayes Factors

The p-value is small when H_0 is dubious, and is always between zero and one; it is common for naïve users to misinterpret p as "the probability H_0 is true." While that probability is meaningless in the Classical paradigm, it is exactly the measure the Bayesian statistician uses to test hypotheses:

$$\mathsf{P}[H_0 \mid X = x^*]$$

is well-defined whether or not H_0 and H_1 are composite hypotheses. It does require a *prior* probability, however.

In the Normal Means example with prior probabilities π_0 and π_1 for $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$, for example,

$$\begin{aligned}\frac{\mathsf{P}[H_1 \mid X = x]}{\mathsf{P}[H_0 \mid X = x]} &= \frac{\pi_1 (2\pi\sigma^2)^{-1/2} e^{-(x-\mu_1)^2/2\sigma^2}}{\pi_0 (2\pi\sigma^2)^{-1/2} e^{-(x-\mu_0)^2/2\sigma^2}} \\ &= \frac{\pi_1}{\pi_0} e^{(x-\bar{\mu})(\mu_1-\mu_0)/\sigma^2}\end{aligned}$$

where $\bar{\mu} = (\mu_0 + \mu_1)/2$ so, in the Signal Processing example where it may be natural to regard $\theta = \pm 1$ as equally likely so $\pi_0 = \pi_1$ and $\bar{\mu} = 0$,

$$\begin{aligned}\frac{\mathsf{P}[H_1 \mid X = x]}{\mathsf{P}[H_0 \mid X = x]} &= e^{8x} \\ \mathsf{P}[H_0 \mid X = x] &= \frac{1}{1 + e^{8x}}\end{aligned}$$

and, in particular, $\mathsf{P}[H_0 \mid X = 0] = 1/2$, as expected.

Notice that the Bayesian posterior odds $\frac{\mathsf{P}[H_1|X=x]}{\mathsf{P}[H_0|X=x]}$ may be written as the product of two terms, the prior odds $\frac{\pi_1}{\pi_0}$ (which depend on the prior but not on the data) and the likelihood ratio $B = \frac{L(x|\theta_1)}{L(x|\theta_0)}$; this latter factor is sometimes called the *Bayes Factor*. It is easy to compute the posterior probability of either hypothesis from B and π; for example,

$$\mathsf{P}[H_0 \mid x] = \frac{\pi_0}{\pi_0 + \pi_1 B}.$$

5.1. Precise Hypotheses with Composite Alternatives

Bayesian hypothesis testing is simple and appealing when both hypotheses specify $\theta \in \Theta$ precisely, with a suitable (perhaps reference) discrete prior distribution is available; and also when both hypotheses are composite, with a suitable (perhaps reference) diffuse prior distribution. The common situation in which one tests a precise hypothesis $H_0 : \theta = \theta_0$ against a composite alternative (such as $H_1 : \theta \neq \theta_0$) requires more delicacy, however.

A continuous prior distribution $\pi(d\theta)$ with a density $\pi(\theta)$ will give zero probability to any lower-dimensional set like $\{\theta_0\}$ under both the prior *and posterior* distributions, so $\mathsf{P}[H_0 \mid x] = 0$ for all data x with such a prior. In many applications this is a reflection of the real phenomenon that a hypothesis that θ is *exactly* equal to any specific value would be false; no coin falls heads with probability 0.500000000 ± 0, for example, and no treatment is *identical* in effect to the control. Usually a precise hypothesis is merely a concise way of describing that θ is *approximately* equal to θ_0, without belaboring exactly what "approximate" means in this context.

Here are three ways of implementing Bayesian testing of precise hypotheses with imprecise alternatives:

1. Use a mixed prior, with a non-zero point mass at θ_0 and the remainder of the prior mass diffuse. This requires some thought about how that remainder ought to be distributed— in particular, it is seldom possible to use an improper "reference" prior here, and the choice of which prior to use will sometimes affect inference. A common approach is to consider a class of prior distributions (for example, unimodal ones that are symmetric about θ_0), and report the range of possible Bayes factors.

2. Use a continuous prior, replacing the point null hypothesis with a composite one of the form $H_0 : |\theta - \theta_0| < \epsilon$. This requires that the investigator be more specific about what sort of approximation the point hypothesis entails.

3. Use a diffuse prior (proper or not) and perform *estimation*, rather than testing. Report an interval estimate for θ, and note whether or not θ_0 lies in the interval.

6. Interesting Reading

Morris DeGroot found settings in which the misinterpretation of p-values as posterior probabilities is harmless because the two happen to be close together (DeGroot, 1973); (ISDS' own) Jim Berger and Delampady found settings in which the misinterpretation is hopelessly wrong (Berger and Delampady, 1987).

References

Berger, J. O. and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–352.

DeGroot, M. H. (1973), "Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio," *Journal of the American Statistical Association*, 68, 966–969.