The Simple Hypothesis Test
Frank Porter
August 24, 2004

Sampling is performed from a probability distribution, $f(x;\theta)$, with unknown parameter θ.[1] For example, often x is the estimator for θ. In our experiment, we observe $x = x^*$. We wish to test the hypothesis:

$$H_0 : \theta = \theta_0,$$

against the alternative:

$$H_1 : \theta = \theta_1.$$

The question being posed is: With what confidence does our data rule out H_0?

Both H_0 and H_1 specify completely the hypothetical sampling distribution, thay are known as "simple" hypotheses. This problem is completely soluble with a most powerful test in classical statistics. The solution involves what is sometimes known as an "ordering principle" for the "likelihood ratio" statistic. We'll formulate things at the start in the way statistics texts do it, by supposing we have a specified confidence level, α, for which we reject the H_0 hypothesis, but then we'll modify slightly to answer the stated question.

We set up the test by saying we are going to "reject" (i.e., consider unlikely) H_0 in favor of H_1 if the observation x lies in some region R of the sample space. This is known as the "critical region" for the test. When we reject one hypothesis in favor of another hypothesis, there are two types of error we could make:

Type I error: Reject H_0 when H_0 is true.
Type II error: Accept H_0 when H_1 is true.

The probability of making a Type I error is:

$$\begin{aligned} \alpha &= \text{Prob}(x \in R | H_0) && (1) \\ &= \int_R f(x;\theta_0)dx. && (2) \end{aligned}$$

The probability α is typically called the "confidence level", or "P-value" in the discussion below.

The probability of making a Type II error is:

$$\begin{aligned} \beta &= \text{Prob}(x \in \bar{R} | H_1) && (3) \\ &= 1 - \int_R f(x;\theta_1)dx, && (4) \end{aligned}$$

where $\bar{R}$ denotes the complement of R in the sample space. The quantity $1-\beta$ is called the "power" of the test; it is the probability that H_0 is correctly rejected.

[1] I'll use language as if x and θ are one-dimensional quantities. However, they may also be multiple dimensional.

Now, there are many possible critical regions R which give the same value for α. We wish to pick the "best critical region", by finding that region for which the power is greatest. Fortunately, this is straightforward for a simple test. We wish to maximize:

$$\begin{aligned} 1-\beta &= \int_R f(x;\theta_1)dx && (5)\\ &= \int_R \frac{f(x;\theta_1)}{f(x;\theta_0)} f(x;\theta_0)dx, && (6)\end{aligned}$$

subject to the constraint:

$$\alpha = \int_R f(x;\theta_0)dx.$$

Notice that

$$\frac{1-\beta}{\alpha} = \left\langle \frac{f(x;\theta_1)}{f(x;\theta_0)} \right\rangle_{(R;H_0)},$$

where the subscript on the $\langle\rangle$ denotes an average over the critical region, under hypothesis H_0. Thus, we wish to build the critical region by including those values of x for which the ratio $\frac{f(x;\theta_1)}{f(x;\theta_0)}$ is largest. This is the "ordering principle". The region R contains all values x for which:

$$\frac{f(x;\theta_1)}{f(x;\theta_0)} \geq \Lambda_\alpha,$$

where Λ_α is determined by the α constraint.

We may re-express this in the context of likelihood functions. Let

$$\lambda(x) \equiv \frac{L(\theta_0;x)}{L(\theta_1;x)}$$

be the "likelihood ratio" for the two hypotheses, for a sampled value x. Note that λ is itself a random variable. The critical region is specified by all values $\lambda \leq \Lambda_\alpha$.

Now, we can turn this around, given a sampling x (and hence λ), and ask what the confidence level, or P-value for H_0 is, according to the value x. That is, what is the probability to get the observed value, or more extreme (where "extreme" is defined in the sense of being in the direction toward favoring H_1), if H_0 is true? This is the probability with which we "rule out" H_0.

Let's try an example: Suppose

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}.$$

We wish to test:

$$H_0 : \theta = \theta_0 = -1,$$

against the alternative:

$$H_1 : \theta = \theta_1 = +1.$$

We sample a value x^* and form the likelihood ratio (we'll also take the logarithm and mulitply by 2 for convenience):

$$\begin{aligned}\ln \lambda^* &= \frac{1}{2}\left[(x^* - \theta_1)^2 - (x^* - \theta_0)^2\right] &(7)\\ &= 2x^*. &(8)\end{aligned}$$

This defines the critical region: $\ln \Lambda_\alpha = 2x^*$. The critical region is thus given by $\ln \lambda \geq 2x^*$. That is, we are trying to determine the probability for λ to exceed the observed value. Since $\ln \lambda = 2x$, we want the probability that $x > x^*$:

$$\alpha = \int_R f(x;\theta_0)dx = \int_{x^*}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta_0)^2} dx.$$

The greater the value of x^*, the smaller is α, and thus the more likely we are to rule out H_0. Note that our result is consistent with the usual intuitive approach in this situation since everything is nicely monotonic.

Consider a specific example, with,

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2},$$

with $\sigma = 0.83$. The sample value is $x^* = 2.72$. The hypotheses being compared are $\theta = \theta_0 = -0.68$ and $\theta = \theta_1 = 0.68$. We already know from our example above that α will be given by the probability that $x > x^*$:

$$\begin{aligned}\alpha &= \int_{x^*}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta_0}{\sigma}\right)^2} dx. &(9)\\ &= 2 \times 10^{-5}. &(10)\end{aligned}$$

Likewise, the power of the test is:

$$\begin{aligned}1-\beta &= \int_{x^*}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta_1}{\sigma}\right)^2} dx. &(11)\\ &= 0.01. &(12)\end{aligned}$$

The rather low power is just telling us that θ_0 and θ_1 are pretty close together on the scale of σ and the value of α. That is, for $\alpha = 2 \times 10^{-5}$, even if H_1 is correct, we will accept H_0 with 99% probability. In a sense, we were "lucky" with our actual sampling, to get such a small value for α (x^* is unlikely even if H_1 is true).

In general, we may not be able to analytically evaluate the sampling distribution for λ, under the hypothesis H_0. In this case, we resort to Monte Carlo simulation to evaluate α. Just be careful that you have really simulated enough experiments to tell you how the tails behave well enough, since that is where the action lies.

Note that the procedure I have described here is not the same as the procedure a Bayesian might apply. The Bayesian would form posterior probabilities:

$$L_0 \equiv \frac{L(\theta_0, x^*)P(\theta_0)}{L(\theta_0, x^*)P(\theta_0) + L(\theta_1, x^*)P(\theta_1)}, \tag{13}$$

$$L_1 \equiv \frac{L(\theta_1, x^*)P(\theta_1)}{L(\theta_0, x^*)P(\theta_0) + L(\theta_1, x^*)P(\theta_1)}, \tag{14}$$

where the prior probabilities are given by $P(\theta_0)$ and $P(\theta_1)$ (with $P(\theta_0)+P(\theta_1) = 1$). Then one may compare the values of the posterior likelihoods ("degrees of belief") at the two hypotheses to get a relative degree of belief in which of the two hypotheses is preferred. This may be contrasted with the strictly frequentist method I have discussed. In my opinion, it is always good to give the frequentist answer, since that describes the data independent of prior beliefs. If you wish to give a Bayesian interpretation in addition, that is fine.