Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 8: Testing hypotheses continued

Lecturer: Dominik Rothenhäusler *February 6*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

8.1 Neyman-Pearson lemma

Here we look at an optimality property of the likelihood ratio test for simple versus simple hypothesis tests. The likelihood ratio for $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_1$ is

$$\lambda = \frac{f_0(x_1, \ldots, x_n)}{f_A(x_1, \ldots, x_n)}.$$

For IID data we get

$$\lambda = \frac{\prod_{i=1}^n f(x_i; \theta_0)}{\prod_{i=1}^n f(x_i; \theta_1)}.$$

Because both hypotheses are simple we have an expression for both numerator and denominator. If one of them would be composite then that probability would ordinarily depend on which θ from that composite hypothesis was true.

The likelihood ratio test rejects H_0 if $\lambda \leqslant \lambda^*$ where λ^* is a threshold that we choose. If we use $\lambda^* = 1$ we reject H_0 if H_A has higher likelihood. In a Bayesian analysis the threshold could be adjusted to account for different prior probabilities of θ_0 and θ_1 or to account for different losses. In the Neyman-Pearson setup we choose λ^* to get a desired type I error probability.

With our chosen λ^*, we get type I error probability

$$\alpha^* = \Pr(\lambda(X) \leqslant \lambda^*; H_0)$$

and type II error probability

$$\beta^* = \Pr(\lambda(X) > \lambda^*; H_A)$$

Here we follow Rice by writing $\lambda(X)$ letting X be "all our data".

The Neyman-Pearson lemma is as follows. If any other test based on the same X has $\alpha \leqslant \alpha^*$ then that test has $\beta \geqslant \beta^*$. The power of a test is $1 - \beta$ so the alternative test then has power no better than the likelihood ratio test has. Lets assume that $0 < \lambda^* < \infty$ to rule out unimportant corner cases.

Proof. We prove it for a continuously distributed X. Let the other test reject when $x \in C$ and fail to reject when $x \in \bar{C}$ (the complement of C). By the definition of the LR test, we have

$$\alpha^* = \int_{x:\lambda(x)\leqslant\lambda^*} f_0(x)\,\mathrm{d}x \quad \text{and} \quad \beta^* = \int_{x:\lambda(x)>\lambda^*} f_A(x)\,\mathrm{d}x.$$

Now the difference in type II error probabilities is

$$\beta - \beta^* = \int_{\bar{C}} f_A(x)\,\mathrm{d}x - \int_{\lambda>\lambda^*} f_A(x)\,\mathrm{d}x.$$

Watch the small changes in the next steps. They are standard manipulations to drive the expression towards something we know. First, we can remove $\bar{C} \cap \lambda > \lambda^*$ from both integral domains. That is where both tests fail to reject under H_A. This brings

$$\begin{aligned}
\beta - \beta^* &= \int_{\bar{C}\cap\lambda\leqslant\lambda^*} f_A(x)\,\mathrm{d}x - \int_{C\cap\lambda>\lambda^*} f_A(x)\,\mathrm{d}x \\
&\geqslant \frac{1}{\lambda_*}\left(\int_{\bar{C}\cap\lambda\leqslant\lambda^*} f_0(x)\,\mathrm{d}x - \int_{C\cap\lambda>\lambda^*} f_0(x)\,\mathrm{d}x\right) \\
&= \frac{1}{\lambda_*}\left(\int_{\bar{C}\cap\lambda\leqslant\lambda^*} f_0(x)\,\mathrm{d}x + \int_{C\cap\lambda\leqslant\lambda^*} f_0(x)\,\mathrm{d}x - \int_{C\cap\lambda\leqslant\lambda^*} f_0(x)\,\mathrm{d}x - \int_{C\cap\lambda>\lambda^*} f_0(x)\,\mathrm{d}x\right) \\
&= \frac{1}{\lambda_*}\left(\int_{\lambda\leqslant\lambda^*} f_0(x)\,\mathrm{d}x - \int_{C} f_0(x)\,\mathrm{d}x\right) \\
&= \frac{1}{\lambda_*}(\alpha - \alpha^*) \\
&\geqslant 0.
\end{aligned}$$

□

No other test can beat the LR test on both α and β.

8.2 Example and a uniformly most powerful test

Suppose we have an exponential random variable with PDF $f(x;\theta) = (1/\theta)e^{-x/\theta}$. Consider $H_0 : \theta = 1$ and $H_A : \theta = 3$. The LR test rejects if

$$\frac{e^{-x}}{(1/3)e^{-x/3}} < \lambda^*. \tag{8.1}$$

We can rearrange that to say we reject H_0 if

$$x > x^* = -\frac{3}{2}\log(\lambda^*/3).$$

That is, whatever λ^* we choose corresponds to some x^* where we will reject H_0 if $x > x^*$. Using the exponential distribution we find that $\Pr(X > x^*; H_0) = \exp(-x^*)$. So to get a desired level α we set $\alpha = \exp(-x^*)$, that is $x^* = -\log(\alpha)$.

Knowing x^* we could if we wanted work backwards, plugging $x^* = -\log(\alpha)$ into (8.1) to find the critical value λ^*. However we already have the test in terms of x^* and that is more convenient to use.

The LR test is our most powerful α-level test for $H_0 : \theta = 1$ versus $H_A : \theta = 3$ by the Neyman-Pearson lemma. It is also the most powerful α-level test for $H_0 : \theta = 1$ versus $H_A : \theta = 7$. The same holds for any H_A

with a θ value larger than 1. We never used the alternative value of θ when constructing the threshold x^*. This test rejecting $H_0 : \theta = 1$ in favor of $H_A : \theta = \theta_1$ rejects when $x > -\log(\alpha)$ no matter what $\theta_1 > 1$ we use. It is a ***uniformly most powerful*** (UMP) test because it has that property over a range of θ values.

A different test is uniformly most powerful over $H_A : \theta = \theta_1$ for $\theta_1 < \theta_0$. That test rejects for small x.

UMP tests are rare and special. We don't ordinarily have them.

8.3 A few facts about p-values

There is a lot of important work on p-values lately. It is worth adding a small segment about them. First, if the test statistic $T(X)$ has a continuous distribution then the p-value is uniformly distributed on $(0, 1)$. It was constructed to have a 5% chance of being below 0.05 a 1% chance of being below 0.01 and so on.

To see formally that it is uniformly distributed, the observed value of p is $p = \Pr(T(X) \geqslant T(x); H_0)$. Let F be the CDF of $T(X)$ under H_0. Then we can write the observed p-value as $p = 1 - F(T(x))$. It is then the observed value of a random variable $P = 1 - F(T(X))$. Now

$$\begin{aligned}
\Pr(P \leqslant \alpha; H_0) &= \Pr(1 - F(T(X)) \leqslant \alpha; H_0) \\
&= \Pr(F(T(X)) \geqslant 1 - \alpha; H_0) \\
&= 1 - \Pr(F(T(X)) \leqslant 1 - \alpha; H_0) \\
&= 1 - \Pr(T(X) \leqslant F^{-1}(1 - \alpha); H_0) \\
&= 1 - F(F^{-1}(1 - \alpha)) \\
&= 1 - (1 - \alpha) \\
&= \alpha,
\end{aligned}$$

and so $P \sim U(0, 1)$.

Suppose that we test some hypothesis H_0 k times getting k independent p-values, p_j for $j = 1, \dots, k$. This is the same hypothesis being tested k times. It might be about whether a dietary intervention reduces blood pressure. If we want to combine those k p-values into a single p-value for H_0 there are several ways to go about it. Doing this is called ***meta-analysis***. It is a study of studies. We could look for a small value of

$$\min_{1 \leqslant j \leqslant k} p_j, \quad \max_{1 \leqslant j \leqslant k} p_j, \quad \text{or} \quad \prod_{1 \leqslant j \leqslant k} p_j$$

among many others. The latter choice was studied by Fisher and is usually favored. It has a nice distributional property. Under H_0 $-2\log(\prod_j p_j) = -2\sum_j \log(p_j) \sim \chi^2_{(2k)}$. Under the alternative this product is even smaller than a $\chi^2_{(2k)}$ random variable. So we reject if $\prod_j p_j$ is small, or equivalently if $-2\sum_j \log(p_j)$ is larger than the $1 - \alpha$ quantile of the $\chi^2_{(2k)}$ distribution. Here is a sketch of how that distribution is derived. First show directly that $-\log(U(0, 1))$ has a standard exponential distribution. Then realize that is a special case of a Gamma distribution. Then recall that summing IID Gammas gives you another Gamma distribution. Then note that χ^2 is a special case of the Gamma distribution. The whole reason for multiplying the logarithm by 2 is to turn the Gamma distribution into one of the χ^2 special cases.

There is a second multiple p-value issue. Suppose now that we have k different null hypotheses H_{0j} for $j = 1, \dots, k$. The example from class was based on a famous XKCD cartoon `https://xkcd.com/882/`. Somebody tests for a link between jelly beans and acne and finds there is not one at $\alpha = 0.05$. They then run separate tests on purple, brown, pink, blue, $\cdots$, orange, doing p-values 20 in all. Of those, one is statistically significant at the 0.05 level (green). That one makes headlines and the others are ignored. If

all k null hypotheses are true and independent and we test them at the α level then the number of false discoveries we get is $\mathrm{Bin}(k, \alpha)$. The expected number of false discoveries is $k \times \alpha$ and the probability of one or more false discoveries is

$$1 - (1 - \alpha)^k.$$

This can be very close to 1 if k is large. There are many ways to contend with the multiple testing issue. The simplest is to do k tests at level α/k each. That makes the probabilty of any false discovery be at most α. To see this let $Z_j = 1$ if the j'th test brings a false discovery and 0 otherwise. Then under all k null hypotheses

$$\begin{aligned}
\Pr(\text{any false discoveries}) &= \Pr(\max_{1 \leqslant j \leqslant k} Z_j = 1) \\
&= \mathbb{E}(\max_{1 \leqslant j \leqslant k} Z_j) \\
&\leqslant \mathbb{E}(\sum_{1 \leqslant j \leqslant k} Z_j) \\
&= \sum_{1 \leqslant j \leqslant k} \mathbb{E}(Z_j) \\
&= \sum_{1 \leqslant j \leqslant k} \alpha/k \\
&= \alpha.
\end{aligned}$$

This approach of dividing the signifance level α by the number k of tests being performed is known as the ***Bonferroni*** method.

8.4 Confidence intervals

Let X be all our data and let θ be the parameter of interest. A $100(1-\alpha)\%$ ***confidence interval*** for θ is an interval of the form $[L(X), U(X)]$ where

$$\Pr(L(X) \leqslant \theta \leqslant U(X); \theta) = 1 - \alpha$$

holds for all θ. For an IID sample this is

$$\Pr(L(X_1, \ldots, X_n) \leqslant \theta \leqslant U(X_1, \ldots, X_n); \theta) = 1 - \alpha.$$

It is important to note that it is L and U that are random while θ is a fixed non-random value. So $\Pr(\cdots)$ refers to the X's inside.

We saw in class that a confidence interval can be turned into a test. If θ_0 is outside the confidence interval, reject $H_0 : \theta = \theta_0$ in favor of $H_A : \theta \neq \theta_0$. Similarly, a test of $H_0 : \theta = \theta_0$ can be turned into a ***confidence set*** made up of all the 'unrejected' θ values. We say confidence 'set' because the result might not be an interval. It would still have the coverage properties. We proved one of these equivalences in class; both are in the text.

To show you that confidence intervals exist, consider $X_i \sim N(\mu, \sigma^2)$ with μ unknown and σ somehow known. Then

$$\Pr\Big(\bar{X} - 2.58\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + 2.58\frac{\sigma}{\sqrt{n}}\Big) = 0.99,$$

so we get a confidence interval. If we don't know σ then we can plug in an estimate $\hat{\sigma}$ and (if it is a good estimate) get

$$\Pr\Big(\bar{X} - 2.58\frac{\hat{\sigma}}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + 2.58\frac{\hat{\sigma}}{\sqrt{n}}\Big) \doteq 0.99.$$

Later we will get something even more precise using the t distribution.

In the method of moments, the central limit theorem gives

$$\Pr\Big(\bar{X} - 2.58\frac{\hat{\sigma}}{\sqrt{n}} \leqslant \mathbb{E}(X) \leqslant \bar{X} + 2.58\frac{\hat{\sigma}}{\sqrt{n}}\Big) \doteq 0.99$$

if $\hat{\sigma}$ is good enough to consistently estimate $\sqrt{\mathrm{Var}(X)}$.

Using Fisher information we get

$$\Pr\Big(\hat{\theta} - \frac{2.58}{\sqrt{nI(\theta_0)}} \leqslant \theta \leqslant \hat{\theta} + \frac{2.58}{\sqrt{nI(\theta_0)}}\Big) \doteq 0.99.$$

If we don't know $I(\theta_0)$ we can usually plug in $I(\hat{\theta})$ (more later in the course). We can reject $H_0 : \theta = \theta_0$ at the $\alpha = 0.01$ level if

$$|\hat{\theta} - \theta_0| > 2.58/\sqrt{nI(\theta_0)}.$$

8.5 Generalized likelihood ratio test

For composite hypotheses we generalize the likelihood ratio test. We reject H_0 in favor of H_A if

$$\Lambda = \frac{\max_{\theta\in H_0} L(\theta)}{\max_{\theta\in H_0\cup H_A} L(\theta)} \leqslant \lambda^*.$$

The numerator is the likelihood under H_0 and the denominator is the likelihood under an even larger set of possibilities. So clearly $\Lambda \leqslant 1$. It has to be enough less than 1 for rejection to be meaningful. We will see how much less.

Suppose that X_i describe the number of copies of a gene that somebody gets. The Hardy-Weinberg model has $\Pr(X_i = 0) = \theta^2$, $\Pr(X_i = 1) = 2\theta(1-\theta)$ and $\Pr(X_i = 2) = \theta^2$. It is a $\mathrm{Bin}(2, \theta)$ model as if their genes were independent random Bernoulli variables one from each parent. In the Hardy-Weinberg model we might want to test $\theta = \theta_0$ for some hypothesized value of θ_0 (maybe $1/2$) versus $\theta \neq \theta_0$. We might also want to test the Hardy-Weinberg model itself against the more general model

$$\Pr(X = x) = \begin{cases} \theta_0, & x = 0 \\ \theta_1, & x = 1 \\ 1 - \theta_0 - \theta_1, & x = 2. \end{cases}$$