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

Lecture 7: Testing hypotheses

Lecturer: Dominik Rothenhäusler *February 1*

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.

7.1 Context

Now we switch from estimation to testing. Until now the data X_i followed a parametric distribution $f(x;\theta)$ where f was perfectly well known and θ was an unknown parameter or vector of parameters. We then looked at ways to pick a value $\hat{\theta}$ to use for θ from the data: moments, likelihood and Bayes.

The family f of distributions could be based on science or experience or perhaps even convenience. Now we switch to testing. We will test whether a given value of θ seems right. We can even test whether that family f of distributions is right.

7.2 Hypothesis test types

We will be testing one hypothesis versus another, H_0 versus H_1. Suppose that we know $\theta = \theta_0$ or $\theta = \theta_1$. Then we can test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. These are both ***simple hypotheses*** because, together with f, they completely describe the distribution of our data $X_1, \ldots, X_n$. We might also test $H_0 : \theta = \theta_0$ versus an hypothesis like $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$. These latter hypotheses are ***composite hypotheses*** because they allow more than one possible value of θ and hence more than one distribution for $X_1, \ldots, X_n$.

Later we will test whether $X_i \sim f(x;\theta)$ is true for any θ. That is, instead of testing whether some θ is right, we test whether f is right. This is a ***goodness of fit test***.

7.3 Bayes testing

Rice begins by working out Bayes hypothesis tests for a discrete data setting. Here we use a continuous data model. The likelihood ratio

$$\mathrm{LR} = \frac{\Pr(\text{all our data} \mid H_0)}{\Pr(\text{all our data} \mid H_1)}$$

plays a critical role. Notice that we use a ratio not a difference. If the numerator and denominator were 0.51 and 0.50 respectively the difference would be the same as 0.02 and 0.01 or even 0.01 and 0. It makes

intuitive sense that 0.51 versus 0.50 is far less conclusive than the others, and we will see that the ratio comes out naturally in our formulas.

In Bayesian testing we compute the posterior probability

$$\frac{\Pr(H_0 \mid \text{data})}{\Pr(H_1 \mid \text{data})} = \frac{\Pr(H_0)}{\Pr(H_1)} \times \frac{\Pr(\text{data} \mid H_0)}{\Pr(\text{data} \mid H_1)}. \tag{7.1}$$

Of course this means our model has to specify the prior probabilties $\Pr(H_0)$ and $\Pr(H_1)$. When H_0 and H_1 are the only possibilities then $\Pr(H_0) = 1 - \Pr(H_1)$ and we get

$$\frac{\Pr(H_0 \mid \text{data})}{1 - \Pr(H_0 \mid \text{data})} = \frac{\Pr(H_0)}{1 - \Pr(H_0)} \times \frac{\Pr(\text{data} \mid H_0)}{\Pr(\text{data} \mid H_1)}$$

that is

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}.$$

If there are more than 2 possibilities then (7.1) still holds for any two of them, but the probability ratios are not odds. [Recall that the odds for event E are $\Pr(E)/(1 - \Pr(E))$.]

Now suppose that H_0 is that an egg has one yolk while under H_1 the egg has two yolks. We weight the egg getting $W \sim N(\mu_1, \sigma^2)$ if there is one yolk and $W \sim N(\mu_2, \sigma^2)$ if there are two. Here $\mu_1 = 2$, $\mu_2 = 2.6$ and $\sigma = 0.4$. After some algebra

$$\text{LR} = \frac{\Pr(W \mid H_0)}{\Pr(W \mid H_1)} = \exp\Big(-\frac{1}{2\sigma^2}[2W(\mu_2 - \mu_1) + \mu_1^2 - \mu_2^2]\Big).$$

Because $\mu_2 - \mu_1 > 0$ the likelihood ratio ***decreases*** exponentially with W. We can find the value W giving $\text{LR} = 1$. It is $(\mu_1 + \mu_2)/2 = 2.3$. If the weight lands half way between the two means then the likelihood ratio equals one and the data fit the two distributions equally well.

If we saw $W = 2.3$ (the midpoint) we would not ordinarily think that the egg is equally probable to be a single or double yolker. In our experience double yolk eggs are rare. Suppose we model that with a prior distribution $\Pr(H_1) = 0.001$ and $\Pr(H_0) = 0.999$. Now

$$\frac{\Pr(H_0 \mid W)}{\Pr(H_1 \mid W)} = 999 \times \exp\Big(-\frac{1}{2\sigma^2}[2W(\mu_2 - \mu_1) + \mu_1^2 - \mu_2^2]\Big). \tag{7.2}$$

Taking account of this prior information there is some value $w_0 > 2.3$ with $\Pr(H_1 \mid W = w_0) = \Pr(H_0 \mid W = w_0)$. The double yolk is more probable if $W > w_0$, and less probable if $W < w_0$.

7.4 Loss function

If one of H_0 or H_1 is true, and we pick one of them we are either right or wrong. Being right is good, but there are two error types: picking H_1 when the truth was H_0 and picking H_0 when the truth was H_1. These two errors might not be equally severe.

Suppose that we are running a recycling plant and a chunk of material going by our scanner is either glass or ceramic. What we think is glass, we melt to recycle. What we think is ceramic we reject. With one error, we might miss a gram or two of usable glass, with another error we might get something that explodes and causes damage. Similar asymmetric losses come up if we are classifying an email as spam or not, a customer as a good loan prospect or not, and so on.

If we think of picking H_1 as a positive discovery then we can make up a loss table like the following:

LOSS	Pick H_0	Pick H_1
H_0 true	0	FP
H_1 true	FN	0

where FN and FP are possibly unequal positive damages attributed to our mistake types.

If we pick H_0 then our expected loss is

$$0 \times \Pr(H_0 \mid \text{data}) + \text{FN} \times \Pr(H_1 \mid \text{data}) = \text{FN} \times \Pr(H_1 \mid \text{data}).$$

If we pick H_1 then our expected loss is

$$\text{FP} \times \Pr(H_0 \mid \text{data}).$$

We should pick H_1 if it has a smaller expected loss. That is, if

$$\text{FP} \times \Pr(H_0 \mid \text{data}) < \text{FN} \times \Pr(H_1 \mid \text{data}).$$

Rearranging this, we pick H_1 if

$$\frac{\text{FN} \times \Pr(H_1 \mid \text{data})}{\text{FP} \times \Pr(H_0 \mid \text{data})} > 1$$

that is if

$$\frac{\Pr(H_1)}{\Pr(H_0)} \times \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)} \times \frac{\text{FN}}{\text{FP}} > 1.$$

We pick H_0 if

$$\frac{\Pr(H_0)}{\Pr(H_1)} \times \frac{\Pr(\text{data} \mid H_0)}{\Pr(\text{data} \mid H_1)} \times \frac{\text{FP}}{\text{FN}} > 1.$$

Our decision criterion is now the product of the prior probability ratio times the likelihood ratio times a loss ratio. Note that only the ratio of the losses enters. Also, if those losses are equal then we pick H_1 if it has higher posterior probability.

It could have been pretty bad to ignore the prior odds of 999:1 for the double yolk egg. Sometimes it is hard to get a good prior ratio to use.

7.5 Neyman-Pearson setup

In the Neyman-Pearson setup we retain H_0 but start calling the other hypothesis H_A. Here H_0 is a ***null hypothesis*** describing a situation considered uninteresting. It might be that a coin has 50% probability of coming up heads, or that a psychic's predictions are complete chance, or that our newly developed drug is exactly as effective as a placebo (e.g., sugar pill) or that two populations of Barramundi have the same average weight, or that something in your diet has exactly no effect on something else in your health. Then H_A is an alternative hypothesis describing one or more ways that H_0 could be wrong.

Very often H_0 is a simple hypothesis such as $H_0 : \theta = \theta_0$ for some special value θ_0. When $\theta \in \mathbb{R}$ (i.e., not a vector) we might have any of the following alternative hypotheses

$$\begin{aligned}
&H_A : \theta = \theta_1 && \text{(simple)}\\
&H_A : \theta \neq \theta_0 && \text{(two sided)}\\
&H_A : \theta > \theta_0 && \text{(one sided)}\\
&H_A : \theta < \theta_0 && \text{(one sided).}
\end{aligned}$$

The choice depends on our scientific goals and understanding. Examples later. Very commonly the null value θ_0 is 0.

In this setup we construct a statistic T designed to take small values if H_A is true than if H_0 is true. In general this is T(all our data). For IID data it would be $T(X_1, \ldots, X_n)$. Here are some examples. If $X_i \sim N(\mu, \sigma^2)$ and H_0 is $\mu = 0$ and H_A is $\mu < 0$ then we might take $T(X_1, \ldots, X_n) = \bar{X}$. If we chose instead to have H_A be $\mu > 0$ then we might take $T(X_1, \ldots, X_n) = -\bar{X}$.

Given this statistic we then decide to ***reject*** H_0 if and only if

$$T(X_1, \ldots, X_n) < t_0.$$

If we don't reject H_0 then we ***accept*** H_0. This does not mean that we have proved H_0 true, just that we tried and failed to reject it. It is also possible that the data are just not informative enough to do it. We might reject H_0 later when we have more or better data. Acceptance is only made grudgingly.

There are now two types of error that we could make. Rejecting H_0 when it is true is called a ***Type I*** error. It frequently corresponds to a ***false discovery*** because H_0 was meant to describe "nothing interesting". Failing to reject H_0 when it is false is called a ***Type II*** error. Here are our possible errors:

Error	Accept H_0	Reject H_0
H_0 true	None	Type I
H_A true	Type II	None

The probability of a type I error is denoted α and the probability of a type II error is denoted β. We have

$$\alpha = \Pr(T(X_1, \ldots, X_n) < t_0 \mid H_0).$$

If H_A is simple, then we can write

$$\beta = \Pr(T(X_1, \ldots, X_n) \geqslant t_0 \mid H_A).$$

If H_A is composite then β can depend on which θ in H_A is the true one. We reject H_0 if we observe a value $T(x_1, \ldots, x_n)$ in the set $(-\infty, t_0)$. This set of T values for which we reject H_0 is called the ***rejection region***. Similarly $[t_0, \infty)$ is the ***acceptance region*** for $T(X_1, \ldots, X_n)$. It is also interesting to think of the set of data values that would lead to rejection. That is

$$\{(x_1, \ldots, x_n) \mid T(x_1, \ldots, x_n) < t_0\}.$$

If $(X_1, \ldots, X_n)$ is in that set, then we reject H_0.

Up to now, any statistic T and any threshold t_0 give us error rates α and β. If we decrease t_0 we generally make α smaller (and never larger) while generally making β larger (and never smaller). For any statistic T that we decide to use there is then a tradeoff in picking t_0. Later we will see good ways to pick the statistic T itself.

The customary way to work this tradeoff is to fix a small level for α. Lots of people use 0.05 but this is a very lenient default that generates lots of false discoveries. One might also use 0.01 or even smaller values. Given that value of α we solve

$$\alpha = \Pr(T(X_1, \ldots, X_n) < t_0 \mid H_0)$$

to find the threshold t_0. Then under a simple alternative hypothesis the type II error probability is

$$\beta = \Pr(T(X_1, \ldots, X_n) \geqslant t_0 \mid H_A).$$

When H_A is composite the type II error probability β typically depends on which θ in the alternative H_A is the true one.

The ***p*-value** is the quantity

$$p = \Pr(T(X_1, \ldots, X_n) < T(x_1, \ldots, x_n) \mid H_0).$$

This is the chance of getting a value of T as large as the one we got or larger. Our test rejects H_0 when $p < \alpha$. For instance, a common choice is to reject H_0 if $p < 0.05$ though this is probably a poor default in many settings. The logic of the p-value is as follows

> "If $p < \alpha$ then either H_0 is wrong or a very rare event has been observed."

A threshold of 0.05 does not seem very rare any more.

The p-value is ***certainly not*** $\Pr(H_A$ is true $\mid$ data) though this is a common misinterpretation. One could reasonably prefer this posterior probability statement to having a p-value. However in order to get that posterior probability statement out of Bayes rule, one would need to specify a prior probability for H_0 and H_A.

Example

Let $X \sim N(\mu, 1)$ with $H_0 : \mu = 0$ and $H_A : \mu = 2$. Here all our data is just the one X. The alternative H_A makes X larger than it would be under H_0 so we decide to reject H_0 if $X > t_0$ for some t_0 (heads up: before we rejected if $T < t_0$). That is $T(X) = X$. To find t_0 we solve

$$\alpha = \int_{t_0}^{\infty} \varphi(x)\,\mathrm{d}x$$

where φ is the $N(0,1)$ pdf. It has CDF commonly denoted Φ (and available in R as `pnorm`). Here we are solving

$$\alpha = 1 - \Phi(t_0).$$

Now $N(0,1)$ is a symmetric distribution. This means that $\varphi(x) = \varphi(-x)$. It also implies that $\Phi(x) = 1 - \Phi(-x)$. (Prove this to yourself if it is not familiar.) So now we know that $\alpha = \Phi(-t_0)$ from which $t_0 = -\Phi^{-1}(\alpha)$. The R function `qnorm` produces Φ^{-1}, the ***quantile function*** of $N(0,1)$. Now

$$\beta = \Pr(N(2,1) < t_0) = \Pr(N(0,1) < t_0 - 2) = \Phi(t_0 - 2) = \Phi(-\Phi^{-1}(\alpha) - 2).$$

Figure 7.1 shows how t_0 and β change with α. Making α small makes β high. Higher than we would like in this model. If $\bar{X}$ would be the average of n IID $N(0,1)$ or $N(2,1)$ observations then it would have the $N(0, 1/n)$ distribution under H_0 and $N(2, 1/n)$ under H_A. For large n we could get both a small α and a small β.

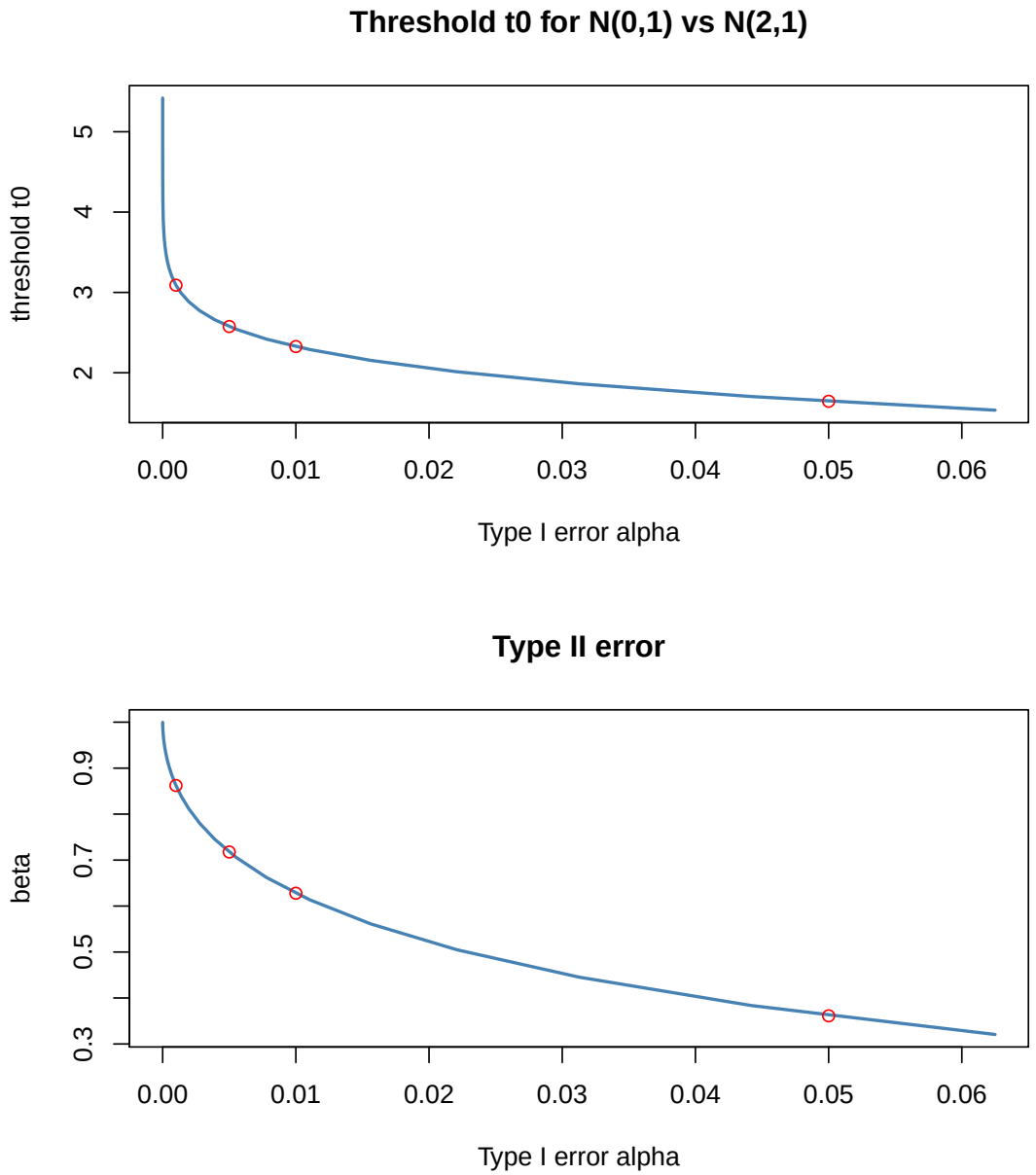

Figure 7.1: The top curve shows how t_0 depends on α for the example test. The bottom shows β versus α. There are circles to mark $\alpha \in \{.05, .01, .005, .001\}$.