36-705: Intermediate Statistics **Fall 2019**

Lecture 20: October 14

Lecturer: Siva Balakrishnan

The classical statistical hypothesis testing framework (as with much of statistics) originated with Fisher.

Example 1: The story goes that a colleague of Fisher claimed to be able to distinguish if in an English tea, milk was added before water (or the other way around).

Fisher proposed to give her 8 cups of tea, 4 of which had milk first, and 4 of which had tea first in a random order. The point was roughly, that if she was "labeling" at random then she would have a small chance (1 in 70) of getting every cup right.

In his description, the null hypothesis was that she had no ability to distinguish. She actually got them all correct, which would have happened by chance with probability 0.014. He concluded that since this probability was less that 0.05 that it was "statistically significant".

Notice the asymmetries/arbitrary-ness in this description: only a null hypothesis is actually specified (i.e. there is no alternative hypothesis – it is in some sense implicit), i.e. the null hypothesis is often special. Furthermore, there is an arbitrary choice of a cut-off 0.05 below which we declare something is significant.

Hypothesis testing is really everywhere. It would probably alarm you to know how many policy decisions, nutrition decisions, scientific results live or die on the basis of hypothesis tests.

Example 2: A couple of typical examples to emphasize again why the null might really be special. A common example is in forensics. Things like fingerprint matches, DNA matches, deciding whether pieces of glass match in their chemical composition etc. are actually problems of a statistical nature. Here perhaps following the "innocent till proven guilty" adage, the null hypothesis is that the defendant is innocent. We then need to review evidence and choose to either reject or fail to reject (i.e. acquit) the defendant. It is perhaps clear that there in many cases is a heavier price for false convictions and so it makes sense to control this error. Indeed, deciding how to choose a significance level in this context is a huge debate.

Example 3: Another common example is in epidemiology. A drug company has new drug and wishes to compare it with current standard treatment. Federal regulators tell the company that they must demonstrate that new drug is better than current treatment to receive approval. The firm runs clinical trial where some patients receive new drug, and others receive standard treatment. There is some numeric response of the "treatment effect" (say higher is better). The null hypothesis is that the new drug is no better than the current standard. Again false positives (where we falsely declare the new drug as better when it is actually harmful) are potentially much worse than false negatives, so we would like to protect against this.

20.1 The formal framework

Let $X_1, \ldots, X_n \sim p(x; \theta)$. Suppose we we want to know if $\theta = \theta_0$ or not, where θ_0 is a specific value of θ. For example, if we are flipping a coin, we may want to know if the coin is fair; this corresponds to $p = 1/2$. If we are testing the effect of two drugs — whose means effects are θ_1 and θ_2 — we may be interested to know if there is no difference, which corresponds to $\theta_1 - \theta_2 = 0$.

We formalize this by stating a *null hypothesis* H_0 and an alternative hypothesis H_1. For example:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad \theta \neq \theta_0.$$

More generally, consider a parameter space Θ. We consider

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$. If Θ_0 consists of a single point, we call this a *simple null hypothesis.* If Θ_0 consists of more than one point, we call this a *composite null hypothesis.*

Example 20.1 $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}. \quad \Box$$

The question is not whether H_0 is true or false. The question is whether there is sufficient evidence to reject H_0, much like a court case. Our possible actions are: reject H_0 or retain (don't reject) H_0.

	Decision	
	Retain H_0	Reject H_0
H_0 true	$\surd$	Type I error (false positive)
H_1 true	Type II error (false negative)	$\surd$

20.1.1 Normal Quantiles

I almost always confuse these so this is mostly for my own reference. Let Φ be the cdf of a standard Normal random variable Z. For $0 < \alpha < 1$, let

$$z_\alpha = \Phi^{-1}(1-\alpha).$$

Hence,

$$P(Z > z_\alpha) = \alpha.$$

Also, $P(Z < -z_\alpha) = \alpha$.

20.2 Constructing Tests

Hypothesis testing involves the following steps:

1. Choose a *test statistic* $T_n = T_n(X_1, \ldots, X_n)$.
2. Choose a rejection region $R \subset \mathcal{X}^n$.
3. If $(X_1, \ldots, X_n) \in R$ we reject H_0 otherwise we retain H_0.

Although strictly speaking you can define the rejection region without an associated test statistic, often it will be the case that R will be defined in terms of the test statistic, i.e. we simply reject if the test statistic takes an "extreme value".

Example 20.2 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Suppose we test*

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}.$$

Let $T_n = n^{-1} \sum_{i=1}^n X_i$ *and* $R = \{x_1, \ldots, x_n : \ |T_n(x_1, \ldots, x_n) - 1/2| > \delta\}$. *So we reject* H_0 *if* $|T_n - 1/2| > \delta$.

We need to choose T and R so that the test has good statistical properties. We will consider the following tests:

1. The Neyman-Pearson Test
2. The Wald test
3. The Likelihood Ratio Test (LRT)
4. The permutation test.

Before we discuss these methods, we first need to talk about how we evaluate tests.

20.3 Error Rates and Power

Suppose we reject H_0 when $(X_1, \ldots, X_n) \in R$. Define the *power function* by

$$\beta(\theta) = P_\theta(X_1, \ldots, X_n \in R).$$

We want $\beta(\theta)$ to be small when $\theta \in \Theta_0$ and we want $\beta(\theta)$ to be large when $\theta \in \Theta_1$. The general strategy is:

1. Fix $\alpha \in [0, 1]$.
2. Now try to maximize $\beta(\theta)$ for $\theta \in \Theta_1$ subject to $\beta(\theta) \leq \alpha$ for $\theta \in \Theta_0$.

Notice the asymmetry that we always favor the null hypothesis and only consider tests that control the Type-I error.

We need the following definitions. A test is *size* α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

A test is *level* α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

A size α test and a level α test are almost the same thing. The distinction is made bcause sometimes we want a size α test and we cannot construct a test with exact size α but we can construct one with a smaller error rate.

Example 20.3 *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Suppose we test*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta > \theta_0.$$

This is called a **one-sided alternative**. *Suppose we reject H_0 if $T_n > c$ where*

$$T_n = \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}.$$

Then

$$\begin{aligned}
\beta(\theta) &= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) = P_\theta\left(\frac{\overline{X}_n - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right),
\end{aligned}$$

where Φ is the cdf of a standard Normal and $Z \sim \Phi$. Now

$$\sup_{\theta\in\Theta_0} \beta(\theta) = \beta(\theta_0) = 1 - \Phi(c).$$

To get a size α test, set $1 - \Phi(c) = \alpha$ so that

$$c = z_\alpha$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$. Our test is: reject H_0 when

$$T_n = \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > z_\alpha.$$

Example 20.4 *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Suppose*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0.$$

This is called a **two-sided** *alternative. We will reject H_0 if $|T_n| > c$ where T_n is defined as before. Now*

$$\begin{aligned}
\beta(\theta) &= P_\theta(T_n < -c) + P_\theta(T_n > c) \\
&= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} < -c\right) + P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) \\
&= P\left(Z < -c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + \Phi\left(-c - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

since $\Phi(-x) = 1 - \Phi(x)$. The size is

$$\beta(\theta_0) = 2\Phi(-c).$$

To get a size α test we set $2\Phi(-c) = \alpha$ so that $c = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1-\alpha/2) = z_{\alpha/2}$. The test is: reject H_0 when

$$|T| = \left|\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}.$$

20.4 The Neyman-Pearson Test

Let $\mathcal{C}_\alpha$ denote all level α tests. A test in $\mathcal{C}_\alpha$ with power function β is **uniformly most powerful (UMP)** if the following holds: if β' is the power function of any other test in $\mathcal{C}_\alpha$ then $\beta(\theta) \geq \beta'(\theta)$ for all $\theta \in \Theta_1$.

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. (Simple null and simple alternative.)

Theorem 20.5 *Let $L(\theta) = p(X_1, \ldots, X_n; \theta)$ and*

$$T_n = \frac{L(\theta_1)}{L(\theta_0)}.$$

Suppose we reject H_0 if $T_n > k$ where k is chosen so that

$$P_{\theta_0}(X^n \in R) = \alpha.$$

This test is a UMP level α test.

One nice thing about this is that it is a "general recipe" for doing a hypothesis test. The drawback of course is that it only applies to the restricted class of simple versus simple tests.

The Neyman-Pearson test, despite its restricted applicability is a very important conceptual contribution. When it is applicable it is an optimal test. This is often called the Neyman-Pearson Lemma, and we will prove this today.

20.4.1 The Neyman-Pearson Lemma

Proof: Let us denote the test function of the NP test as ϕ_{NP} and the test function of any other test we want to compare against as ϕ_A. The test function simply takes the value 1 if the test rejects and 0 otherwise. Since the parameters θ_0 and θ_1 are fixed we will be thinking of the likelihood as X^n is varied. To ease notation we will assume that one sample is observed (nothing changes more generally) and denote $f_0(x) = L(\theta_0; x)$ and $f_1(x) = L(\theta_1; x)$. So with this notation we simply reject if:

$$\frac{f_1(x)}{f_0(x)} \geq k.$$

To prove the NP Lemma, we will first argue that the following is true:

$$\int_x \underbrace{(\phi_{NP}(x) - \phi_A(x))}_{T_1} \underbrace{(f_1(x) - k f_0(x))}_{T_2} dx \geq 0.$$

To see this we can just consider some cases:

1. If both tests reject or if both tests accept then the inequality is clearly true since the LHS is 0.

2. If NP rejects, and the test A accepts then $\phi_{NP}(x) = 1$, and $\phi_A(x) = 0$, so $T_1 \geq 0$. Since the NP test rejected the null we know that:

   $$\frac{f_1(x)}{f_0(x)} \geq k,$$

   so that $T_2 \geq 0$. So the inequality is true in this case.

3. If NP accepts and the test A rejects then both T_1 and T_2 are negative so the inequality is also true in this case.

So we can see that for every x, $T_1 \times T_2 \geq 0$ so it is true when we integrate over x. Now, we can rearrange this inequality to see that:

$$\begin{aligned}\int_x (\phi_{NP}(x) - \phi_A(x)) f_1(x) dx &\geq k \int_x (\phi_{NP}(x) - \phi_A(x)) f_0(x) dx \\ &= k \left(\underbrace{\int_x \phi_{NP}(x) f_0(x) dx}_{=\alpha} - \underbrace{\int_x \phi_A(x) f_0(x) dx}_{\leq \alpha} \right) \\ &\geq 0.\end{aligned}$$

This proves the NP lemma, i.e. that the power of the NP test is larger than the power of any other test.