Lecture 22: October 24

Lecturer: Siva Balakrishnan

22.1 Review and Outline

In the last class we discussed cross-validation:

1. Model-selection via the train-validation split.
2. Cross-validation in regression, and density estimation.

Today we will begin discussing hypothesis testing. We will again follow Wasserman's book for this portion of the course.

22.2 Hypothesis Testing

The typical (and most basic) setting is that we observe:

$$X_1, \ldots, X_n \sim f_\theta$$

and want to test if $\theta = \theta_0$ or not. A typical example is where we have a coin and would like to know if the coin is fair or not. In a clinical trial we might have a control group and a group taking the drug, and we would like to know if the difference in some health outcome is 0 or not.

The way we formalize this is by defining a *null hypothesis* H_0 and an *alternative hypothesis* H_1.

So we would say:

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

The more general case is that we have two sets of parameters Θ_0 and Θ_1 which are non-overlapping, i.e. $\Theta_0 \cap \Theta_1 = \emptyset$ and would like to test the hypothesis:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1.$$

We will refer to the case when Θ_0 is a single point as a *simple* null versus the more general case of a *composite* null.

Example 1: In the example above of testing if a coin is fair or not. We have

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

Our null and alternate hypotheses are:

$$H_0 : p = 1/2$$
$$H_1 : p \neq 1/2.$$

In this case we have a simple null.

In hypothesis testing, the question is never if the null hypothesis is true or not. Rather the question of interest is whether we have sufficient evidence to reject the null hypothesis or not. So in hypothesis testing, there are two possibilities you reject the null hypothesis or you retain it. To reiterate, retaining the null hypothesis is not a statement about whether it is true or not.

There are two types of errors one might make in hypothesis testing: a *Type I* error is when the null hypothesis is true but was incorrectly rejected, and a *Type II* error is when the alternate hypothesis was true but we failed to reject the null.

22.3 Construction of Tests

The typical way we construct tests is:

1. We choose a *test statistic* $T_n = T_n(X_1, \ldots, X_n)$.
2. We choose a *refection region* R.
3. If $T_n \in R$ we reject H_0 otherwise we retain H_0.

Example 2: Suppose again $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, and we test:

$$H_0 : p = 1/2$$
$$H_1 : p \neq 1/2.$$

A natural test statistic would be:

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and a natural rejection region would be:

$$R = \{(X_1, \ldots, X_n) : |T_n(X_1, \ldots, X_n) - 1/2| \geq \delta\}.$$

Effectively we reject H_0 if T_n is far from $1/2$. We need to choose δ to ensure that the test has good properties.

More generally, we need to choose both the test statistic T and the rejection region R to ensure our tests are good. Let us now discuss how we evaluate tests.

22.4 Evaluating Tests

Suppose that we reject the null hypothesis when $(X_1, \ldots, X_n) \in R$. We can define the *power function* as:

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

We would like that $\beta(\theta)$ to be small over Θ_0 and large over Θ_1. The Neyman-Pearson paradigm is the following:

1. Pick an $\alpha \in [0, 1]$.
2. Then try to maximize $\beta(\theta)$ over Θ_1 subject to
$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

Tests of this form are called level α tests, i.e. level α tests are ones for which: $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$. Lets look at a couple of examples:

Example 3: Suppose $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. We want to test:

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

The alternative here is called a *one-sided alternative.*

A natural test statistic here would again be the average but we will re-scale it for convenience:

$$T_n(X_1, \ldots, X_n) = \frac{\frac{1}{n}\sum_{i=1}^n X_i - \theta_0}{\sigma/\sqrt{n}}.$$

Again, a natural strategy would be to reject if $T_n > t$ for some threshold t. We would like to compute the power function:

$$\beta(\theta) = P_\theta(T_n > t) = P_\theta\left(\frac{\frac{1}{n}\sum_{i=1}^n X_i - \theta}{\sigma/\sqrt{n}} > t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).$$

Now, we can see that when the true mean is θ, the quantity:

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - \theta}{\sigma/\sqrt{n}} \sim N(0,1),$$

so that the power function is simply:

$$\beta(\theta) = P\left(Z > t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).$$

So now we can try to implement the Neyman-Pearson paradigm. We want to pick the threshold t so that:

$$\sup_{\theta \in \Theta_0} 1 - \Phi\left(t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \leq \alpha,$$

which is the same as:

$$1 - \Phi(t) \leq \alpha,$$

We want to maximize $\beta(\theta)$ when $\theta > \theta_0$ so we use the threshold:

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

Example 4: Suppose $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. We want to test:

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

This is now a *two-sided alternative*. A natural idea, would be to reject if the magnitude $|T_n| > t$ for some threshold t. In this case, the power function:

$$\beta(\theta) = P_\theta(T_n < -t) + P_\theta(T_n > t),$$

which as before we can expand as:

$$\begin{aligned}
\beta(\theta) &= P_\theta\left(\frac{\frac{1}{n}\sum_{i=1}^n X_i - \theta}{\sigma/\sqrt{n}} < -t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + P_\theta\left(\frac{\frac{1}{n}\sum_{i=1}^n X_i - \theta}{\sigma/\sqrt{n}} > t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(t + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

Again to implement the NP paradigm we notice that under the null we have that:

$$\beta(\theta_0) = \Phi(-t) + 1 - \Phi(t) = 2\Phi(-t) \leq \alpha,$$

so we set:

$$t = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1 - \alpha/2).$$

To summarize our progress so far: we have seen how to set up a hypothesis testing problem formally. We have discussed the Neyman-Pearson paradigm which gives us a way to set a test threshold (or set a rejection region): for a given test statistic we set the threshold to ensure that the test has level α, while giving maximum power.

This however, pre-supposes that we know how to come up with a good/reasonable test statistic. In our next lecture we will discuss general principles that will guide us towards good test statistics.