IEOR 165 – Lecture 24
Neyman-Pearson Testing

1 Summary of Null Hypothesis Testing

The main idea of null hypothesis testing is that we use the available data to try to invalidate the null hypothesis by identifying situations in which the data is unlikely to have been observed under the situation described by the null hypothesis. Though this is the predominant hypothesis testing framework in medical, biological, and physical sciences, one weakness of this framework is that it does not take into consideration alternative possibilities. There are in fact other hypothesis testing frameworks that consider more refined notions. For instance, statistical decision making (including minimax and Bayesian settings) explicitly considers risk. We will consider a simpler variant in this course.

2 Framework of Neyman-Pearson Testing

In this framework, we define two hypothesis that we would like to choose between. Because there are two possible choices, there are two possible errors. The main idea is that one error is more important, and this determines how we label the hypothesis. The hypothesis for which mistakenly not choosing that hypothesis is more important is called the *null hypothesis*, and the other hypothesis is called the *alternative hypothesis*. An example from scientific fields is where the null hypothesis might represent "no effect". We still talk about accepting (choosing) and rejecting (not choosing) the null hypothesis.

With respect to the two hypothesis, there is certain associated terminology.

- A *type I error* occurs when choosing the alternative hypothesis under the assumption that the null is true, and it is also known as a *false positive*.
- Similarly, a *type II error* occurs when choosing the null hypothesis under the assumption that the alternative is true, and it is also known as a *false negative*.
- A parametric hypothesis class consists of a family of distributions $\mathbb{P}_\theta : \theta \in \Theta$, with the null and alternative hypothesis consisting of a specific set of parameters $H_0 : \theta \in \Theta_0 \subset \Theta$ and $H_1 : \theta \in \Theta_1 \subset \Theta$, where $\Theta_0 \cap \Theta_1 = \emptyset$.

The idea is to begin by defining the level of significance α to be the probability of committing a type I error. Because a test with significance level α also has significance level $\tilde{\alpha} > \alpha$, we also define *size* be the minimum probability of committing a type I error for a specific test. For

instance, if we have a parametric hypothesis test defined by comparing a test statistic $T(X)$, which depends on some data X, to a threshold, then the size is defined by

$$\alpha(c) = \sup\{\mathbb{P}_\theta(T(X) \geq c) : \theta \in \Theta_0\}.$$

Once the size of the test is fixed, the probability of a type II error becomes fixed as well. It is conventional to refer to the *power* of the test, which gives the probability of choosing the alternative hypothesis when it is true, and the power of a test is given by $1 - \mathbb{P}(\text{Type II Error})$. Moreover, the power function for all $\theta \in \Theta$ is defined as

$$\beta(\theta) = \mathbb{P}_\theta(\text{Rejecting } H_0) = \mathbb{P}_\theta(T(X) \geq c).$$

Perhaps confusingly, Neyman-Pearson hypothesis testing still has a concept of a p-value. In fact, the p-value can be defined in the same way as in the case of null hypothesis testing. There is an alternative definition of a p-value that is worth mentioning, and this alternative definition applies to both null hypothesis testing and Neyman-Pearson testing: The p-value for a certain set of data and a specific test statistic is an upper bound on the significance level for which the null hypothesis would be rejected.

3 Example: One-Sided Test for Gaussian with Known Variance

Suppose iid data is from $X_i \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known, and consider the hypothesis

$$H_0 : \mu \leq 0$$
$$H_1 : \mu > 0.$$

We could use the test statistic $T(X) = \sqrt{n}\overline{X}/\sigma$. The power function is given by

$$\beta(\mu) = \mathbb{P}_\mu(T(X) \geq c) = \mathbb{P}_\mu\Big(\sqrt{n}\frac{(\overline{X} - \mu)}{\sigma} \geq c - \sqrt{n}\frac{\mu}{\sigma}\Big).$$

Thus, the size of the test is

$$\alpha(c) = \sup\{\mathbb{P}_\mu(T(X) \geq c) : \mu \leq 0\} = \mathbb{P}_{\mu=0}\Big(\sqrt{n}\frac{\overline{X}}{\sigma} \geq c\Big),$$

where the last equality is because $\beta(\mu)$ is increasing in μ. If $c^* = z(1-a^*)$ (meaning the $(1-\alpha^*)$ quantile of $\mathcal{N}(0,1)$, then $\alpha(c^*) = \alpha^*$.

4 Neyman-Pearson Lemma

One of the benefits of Neyman-Pearson hypothesis testing is that there is powerful theory that can help guide us in designing parametric hypothesis tests. In particular, suppose we have *simple*

hypothesis that consist of a single point

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1.$$

For a given significance level α, we can ask what is the *most powerful* test? By most powerful, we mean the test with highest power over the set of all tests with level α.

We define the simple likelihood ratio to be

$$L(X, \theta_0, \theta_1) = \frac{\prod_{i=1}^n f_{\theta_1}(X_i)}{\prod_{i=1}^n f_{\theta_0}(X_i)},$$

where $f_{\theta_0}(\cdot)$ and $f_{\theta_1}(\cdot)$ denote the density function under the null and alternative. The *likelihood ratio test* (or *Neyman-Pearson test*) is defined by

- Accept the null if $L(X, \theta_0, \theta_1) < k$.
- Reject the null if $L(X, \theta_0, \theta_1) > k$.

When the likelihood ratio is equal to k, theoretically we need to introduce randomization into the test; however, this is typically not an issue in practice.

The Neyman-Pearson lemma has several important consequences regarding the likelihood ratio test:

1. A likelihood ratio test with size α is most powerful.
2. A most powerful size α likelihood ratio test exists (provided randomization is allowed).
3. If a test is most powerful with level α, then it must be a likelihood ratio test with level α.

5 Example: Two Means for Gaussian with Known Variance

Suppose iid data is from $X_i \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known, and consider the hypothesis

$$H_0 : \mu = 0$$
$$H_1 : \mu = \mu_1.$$

Then the likelihood ratio is

$$L(X, 0, \mu_1) = \exp\Big(\frac{\mu_1}{\sigma^2}\sum_{i=1}^n X_i - \frac{n\mu_1^2}{2\sigma^2}\Big).$$

This looks quite different from tests we have considered before. However, observe that any strictly increasing function of the likelihood ratio does not change the basic nature of the test. And so if we define the following strictly increasing function of the likelihood ratio

$$\frac{\sigma}{\mu_1\sqrt{n}}\Big(\log L(X,0,\mu_1) + \frac{n\mu_1^2}{2\sigma^2}\Big),$$

then some algebra gives that this function is equal to $\sqrt{n}\overline{X}/\sigma$. And so for this case, the likelihood ratio test is equivalent to using the sample average as a test statistic.

Using our previous tests, we know that test with size α is given by rejecting the null if $\sqrt{n}\overline{X}/\sigma \geq z(1-\alpha)$. The power of this test is

$$\mathbb{P}\Big(Z \leq z(\alpha) + \frac{\mu_1\sqrt{n}}{\sigma}\Big),$$

where $Z \sim \mathcal{N}(0,1)$.

6 Uniformly Most Powerful Test

Now consider the case of *composite* hypothesis that consist of multiple points

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

For a given significance level α, we can ask what is the *uniformly most powerful* test? By uniformly most powerful, we mean the test with highest power over the set of all tests with level α, for all the possible $\theta \in \Theta_1$ values.

We need to make one technical definition. Define a *monotone likelihood ratio family* to be a parametric model $\mathbb{P}_\theta$ with a test statistic $T(X)$ for which $f_{\theta_b}(x)/f_{\theta_a}(x)$ is an increasing function of $T(X)$, whenever $\theta_a \leq \theta_b$. As an instance, observe that the likelihood ratio from the previous example is a strictly increasing function of the sample average; and so the family of models in the previous example is a monotone likelihood ratio family.

A key result concerns the test that rejects the null hypothesis if and only if $T(x) \geq t(1-\alpha)$, where $t(1-\alpha)$ is the $(1-\alpha)$ quantile of the distribution of $T(X)$ under the distribution $\mathbb{P}_{\theta_0}(\cdot)$. This result says that this test is uniformly most powerful with level α.

7 More Information and References

The material in these notes follows that of the textbook "Mathematical Statistics, Basic Ideas and Selected Topics, Vol. 1, (2nd Edition)" by Peter Bickel and Kjell Doksum.