Lecture 12 — November 3

Lecturer: Lester Mackey *Scribe: Jae Hyuck Park, Christian Fong*

Warning: *These notes may contain factual and/or typographic errors.*

12.1 Summary

In this lecture, we are going to review briefly the concepts developed for tackling the problem of optimal inference in the context of point estimation. Then, we are going to consider a different kind of inference setting, that of hypothesis testing. We will develop the Neyman-Pearson paradigm and show how to find optimal tests for so called "simple" problems.

12.2 Point Estimation Recap

In the first part of the course, we focused on optimal inference in the setting of point estimation (see Figure 12.2). We formulated this problem in the framework of decision theory and focused on finite sample criteria of optimality. We immediately discovered that uniform optimality was seldom attainable in practice, and thus, we developed our theory of optimality along two lines: constraining and collapsing.

To restrict ourselves to interesting subclasses of estimators, we first introduced the notion of unbiasedness, which lead us to UMRUEs/UMVUEs. Then we considered certain symmetry constraints in the context of location invariant decision problems—this was formalized via the concept of equivariance, which led us to MREs. The situation is similar in hypothesis testing: to develop a useful notion of optimality, we will need to impose constraints on tests. These constraints arise in the form of risk bounds, unbiasedness, and equivariance.

Another way to achieve optimality was to collapse the risk function. We introduced the notions of average risk (optimized by Bayes estimators) and worst case risk (optimized by minimax estimators). We saw that Bayes estimators have many desirable properties and provide tools to reason about both concepts.

We also considered a few different model families such as exponential family and location and scale family. We came out with certain notions of optimality, optimal unbiased estimator was easy to find for exponential family with complete sufficient statistics and for location-scale family, we defined the best equivariant estimators.

12.3 Hypothesis Testing

We now shift our attention to the problem of *hypothesis testing*. As discussed above, many of the same issues we encountered in defining and achieving optimality for point estimation will reemerge here. However, we do encounter two entirely new objects: the null, H_0, and

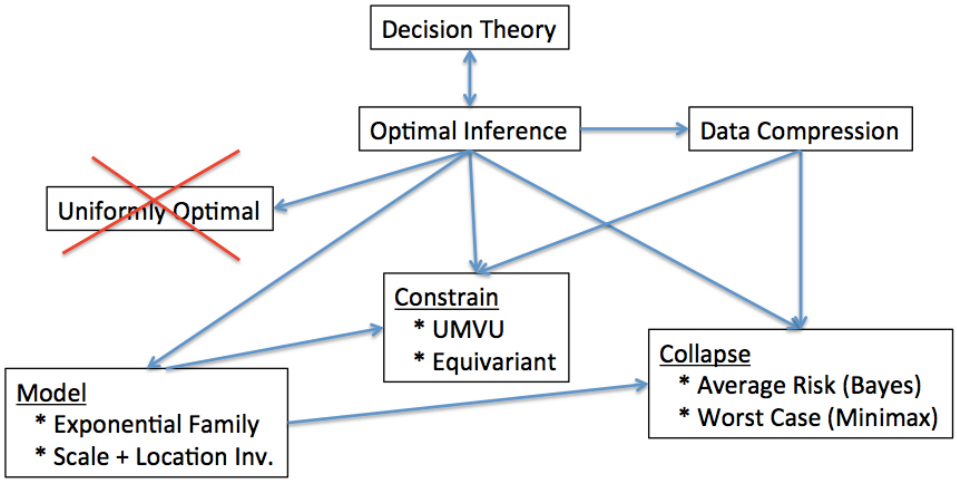

Figure 12.1. Review of point estimation.

alternative, H_1, hypotheses. It is this development which will lead us to introduce the notion of risk bounds.

12.3.1 Model Setup

Hypothesis testing is just a particular type of decision problem. As usual, we assume that the data is sampled according to $X \sim \mathbb{P}_\theta$ and that $\mathbb{P}_\theta$ belongs to the model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$. In addition to the standard setup, we divide the models in $\mathcal{P}$ into two disjoint subclasses known as "hypotheses":

$$\begin{aligned} H_0 &: \theta \in \Omega_0 \subset \Omega \text{ (null hypothesis)} \\ H_1 &: \theta \in \Omega_1 = \Omega \setminus \Omega_0 \text{ (alternative hypothesis)} \end{aligned}$$

Our goal is to infer which hypothesis is correct. This can be cast as classification, so our decision space is

$$\mathcal{D} = \{\text{accept } H_0,\ \text{reject } H_0\}\,.$$

Example 1. You have developed a new anti-itch cream, and you suspect that it may have grave side effects. We can test this hypothesis on a population of mice by applying anti-itch cream to some and no cream to others. In particular, we would like to see if there is a change in life expectancy when the cream is used. Thus, the two hypotheses are H_0 : No change in life expectancy and H_1: Change in life expectancy after application of cream. Based on the results of the experiments we can choose to accept or reject the hypothesis.

While one could imagine equipping this decision problem with a variety of loss functions, there is a canonical loss function $L(\theta, d)$ that gives rise to the optimality goals classically espoused in hypothesis testing. We specify this loss in matrix form in Table 12.3.1. The columns represent the true state of nature, i.e., whether $\theta \in \Omega_0$ or $\theta \in \Omega_1$, while the rows indicate the decision that was made.

	$\theta \in \Omega_0$	$\theta \in \Omega_1$
Reject H_0	1 (Type I Error)	0 (Good)
Accept H_0	0 (Good)	1 (Type II Error)

Table 12.1. Canonical loss function $L(\theta, d)$.

We have two types of error which induce a loss. A **Type I error** or false positive occurs when we reject H_0 when it is in fact true. Similarly a **Type II error** or false negative occurs when we accept H_0 when it is false. In practice, one might assign different loss values to the two types of error, as they could have significantly different consequences. This would lead us back towards point estimation where the target is now a binary parameter indicating whether H_0 or H_1 is true. What we do below is different from this formulation. We in fact induce a more stringent form of asymmetry between Type I and Type II errors through the introduction of risk bounds in the next section. For now, we will forgo discussing the advantages/disadvantages of the respective formulations.

Because our loss function is this kind of indicator function, throughout our discussion of hypothesis testing we will allow for randomized decision procedures δ_ϕ which we specify in terms of a **test function** $\phi(X) \in [0, 1]$ (also known as the **critical function**). The test function ϕ indicates that $\delta_\phi(X, U)$ rejects H_0 w.p. $\phi(X)$.[1] In other words,

$$\phi(X) = \mathbb{P}\left(\delta_\phi(X, U) = \text{Reject } H_0 \mid X\right).$$

where U is as usual a uniform random variable independent of X. The function $\phi(X)$ completely specifies the behavior of δ_ϕ, and it will be convenient to work directly with $\phi(X)$ in place of δ_ϕ. While we could safely ignore randomized procedures when considering optimality under convex loss functions (due to the Rao-Blackwell theorem), we must account for improvements due to randomization under our non-convex loss L.

In order to start reasoning about optimality in this context we are going to introduce definition.

Definition 1. The **power function** of a test ϕ is $\beta(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta\left(\text{Reject } H_0\right)$.

Indeed we can describe our risk function wholly in terms of this power function.

Note: If $\theta_0 \in \Omega_0$, then $\beta(\theta_0) = R(\theta_0, \delta_\phi) =$ Type I Error rate. For $\theta_1 \in \Omega_1$, then $\beta(\theta_1) = 1 - R(\theta_1, \delta_\phi) = 1 -$ Type II Error rate.

Our ideal optimality goal is to minimize $\beta(\theta_0)$ uniformly for all $\theta_0 \in \Omega_0$ and maximize $\beta(\theta_1)$ uniformly for all $\theta_1 \in \Omega_1$. Unfortunately, it is typically impossible to minimize all errors across all parameters. So what can we do? We can constrain the form of the procedure that we are allowed to use, leading to the Neyman-Pearson paradigm of hypothesis testing.

[1] Recall that randomized decision procedures are functions of the data and an independent source of randomness $U \sim \text{Unif}[0, 1]$.

12.4 The Neyman-Pearson Paradigm

We are going to start by fixing a value $\alpha \in (0,1)$. We will call it the *level of significance*. We will require that our procedures satisfy the following risk bound:

$$\sup_{\theta_0 \in \Omega_0} \mathbb{E}_{\theta_0} \phi(X) = \sup_{\theta_0 \in \Omega_0} \beta(\theta_0) \leq \alpha.$$

The quantity $\sup_{\theta_0 \in \Omega_0} \beta(\theta_0)$ is called the *size* of the test. If the size of the test ϕ is bounded by α, ϕ is called a *level α test*. The level of the test represents the tolerance we have for falsely rejecting the null hypothesis. Essentially, instead of trying to minimize the Type I error, the Neyman-Pearson paradigm simply bounds it and focuses on minimizing the Type II error. Our new optimality goal can be summarized as follows.

Optimality Goal: Find a level α test that maximizes the power $\beta(\theta_1) = \mathbb{E}_{\theta_1}[\phi(X)]$ for each $\theta_1 \in \Omega_1$. Such a test is called *uniformly most powerful* (UMP).

We still need to maximize power uniformly for all alternatives θ_1, so what have we gained by working under this risk bound? It turns out that in special cases, UMP (i.e., optimal) tests, formulated in this way, exist. The intuition is as follows. A test is determined by the region of the sample space for which it rejects the null hypothesis (the rejection region). The differences in data generated from $\theta \in \Omega_0$ versus data generated from $\theta \in \Omega_1$ determine the "shape" of effective versus ineffective rejection regions. If it turns out that a certain shape is optimal for all pairs $\theta_0 \in \Omega_0$ and $\theta_1 \in \Omega_1$, then we will have a UMP. When UMP tests do not exist, we will introduce additional constraints such as unbiasedness and equivariance, as we did with point estimation.

12.5 MP for the "simple" case

Definition 2. A hypothesis H_0 is called **simple** if $|\Omega_0| = 1$, otherwise it is called **composite**. The same is true for H_1.

The "simple" case in hypothesis testing is the case when both H_0 and H_1 are simple. In this case, we will use the notation

$$H_0 : X \sim p_0$$
$$H_1 : X \sim p_1$$

where p_0, p_1 denote the densities of $\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1}$ with respect to some common measure μ, and we call $\mathbb{E}_{p_1}[\phi(X)]$ the **power of the test** ϕ. Our goal in the simple case can be compactly described as:

$$\begin{aligned} \max_{\phi} \quad & \mathbb{E}_{p_1}[\phi(X)] \\ \text{s.t.} \quad & \mathbb{E}_{p_0}[\phi(X)] \leq \alpha. \end{aligned}$$

Any maximizing test ϕ is called *most powerful* (MP). (We drop the word "uniformly" as we are only interested in maximizing the power under a single alternative distribution.) In this setting, it turns out that it is not too difficult to find such a test; one is given by

Lemma 1 (Neyman-Pearson)**.**

(i) **Existence.** *For testing $H_0 : p_0$ vs. $H_1 : p_1$, there is a test $\phi(X)$ and a constant k such that:*

(a) $\mathbb{E}_{p_0}\phi(X) = \alpha$ *(size = level).*

(b) $\phi(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} > k \text{ (always reject if likelihood ratio is } > k). \\ 0 & \text{if } \frac{p_1(x)}{p_0(x)} < k \text{ (always accept if likelihood ratio is } < k). \end{cases}$

Such a test is called a likelihood ratio test *(LRT).*

(ii) **Sufficient.** *If a test satisfies (a),(b) for some constant k, it is* most powerful *for testing $H_0 : p_0$ vs. $H_1 : p_1$ at level α. (Hence, the LRT from part (i) is most powerful.)*

(iii) **Necessary.** *If a test ϕ is MP at level α then it satisfies (b) for some k, and it also satisfies (a) unless there exists a test of size $< \alpha$ with power 1. (In the latter case, we did not need to expend all of budgeted Type I error.)*

This is a comprehensive lemma that essentially "solves" the hypothesis testing problem for the simple case. Note that the lemma makes no explicit mention of how the test behaves when $\frac{p_1(x)}{p_0(x)} = k$. When p_0 and p_1 are continuous with respect to Lebesgue measure, this region has measure 0 and hence is of no consequence to us. Otherwise, the behavior in this region is usually determined by the desire to satisfy the size = level constraint of part (a), as we will see next time when we prove the NP lemma.

Example 2. As a first example consider a situation where you observe the first symptom X of an illness and the goal is to distinguish between two possible illnesses (different distributions over X). The problem parameters are given in the following table:

X	sneezing	fever	fainting	sore throat	runny nose
$H_0 : p_0$ (cold)	1/4	1/100	1/100	3/100	70/100
$H_1 : p_1$ (flu)	1/2	10/100	2/100	5/100	33/100
$r(x) = p_1(x)/p_0(x)$	2	10	2	5/3	33/70

.

We want to come up with a most powerful test for this model. According to the NP lemma, we need to compute the likelihood ratio: $r(x) = \frac{p_1(x)}{p_0(x)}$. Now, suppose that the test rejects the cold hypothesis iff $X \in \{\text{fever}\}$. Is this MP for some level α? This test satisfies part (b) of the NP lemma for $k \in (2, 10)$, let's say $k = 5$. The size is given by the probability that $\mathbb{P}_{p_0}(X = \text{fever}) = \frac{1}{100}$, which implies this is a most powerful test at $\alpha = \frac{1}{100}$.

The flexibility to specify the outcome of the test when $r(x) = k$ is important. To see this, suppose ϕ rejects if $X \in \{\text{fever}, \text{fainting}\}$. This satisfies (b) of the NP lemma for $k = 2$; simply specify that $\phi(\text{fainting}) = 1$ and $\phi(\text{sneezing}) = 0$. The size is given by $\mathbb{P}_{p_0}(X \in \{\text{fever}, \text{fainting}\}) = \frac{2}{100}$, so it is most powerful at $\alpha = \frac{2}{100}$.

Next, we will look at a setting involving continuous density functions.

Example 3. Let $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with σ known. Consider the following two hypotheses:

$$H_0 : \mu = 0 \text{ and } H_1 : \mu = \mu_1$$

where μ_1 is known. Since this is a simple case (only two distributions), we calculate the likelihood ratio:

$$\begin{aligned} r(x) = \frac{p_1(x)}{p_0(x)} &= \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(x_i - \mu_1)^2/2\sigma^2\right)}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-x_i^2/2\sigma^2\right)} \\ &= \exp\left(\frac{1}{\sigma^2}\mu_1 \sum_{i=1}^n x_i - \frac{n\mu_1^2}{2\sigma^2}\right). \end{aligned}$$

We observe that the likelihood ratio is only a function of the sufficient statistic (this is in general true by the Factorization Criterion). We have the following equivalences:

$$\begin{aligned} r(x) > k &\Leftrightarrow \mu_1 \frac{\sum x_i}{\sigma^2} - \frac{n\mu_1^2}{2\sigma^2} > \log k \\ &\Leftrightarrow \mu_1 \sum x_i > k' \\ &\Leftrightarrow \begin{cases} \sum x_i > k'' & \text{if } \mu_1 > 0 \\ \sum x_i < k''' & \text{if } \mu_1 < 0 \end{cases} \end{aligned}$$

Let us focus on the first case where $\mu_1 > 0$ (but it is important to note that the two cases $\mu_1 > 0$ versus $\mu_1 < 0$ induce different rejection regions). We can rewrite the test in a different form so that the left hand side of the inequality has standard normal distribution under the null:

$$\Leftrightarrow \frac{\sqrt{n}\bar{x}}{\sigma} > k''''$$

Note: Here the sufficient statistic essentially determines a MP test. Also, observe that for a given level α, the constant involved in the LRT is uniquely determined by the constraint $\mathbb{E}_{p_0}\phi(X) = \alpha$ and thus depends only on the distribution of the null hypothesis and not at all on μ_1 (provided that $\mu_1 > 0$).

Thus, we have by the NP lemma that the test: reject H_0 iff $\sqrt{n}\bar{x}/\sigma > k(\alpha)$ is MP where we pick $k'''' = k(\alpha)$ (where $k(\alpha) \equiv z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal) so that it equals the target level. This value is uniquely determined by the size constraint:

$$\mathbb{E}_{p_0}\phi(X) = \alpha = \mathbb{P}_{\mu=0}\left(\frac{\sqrt{n}\bar{X}}{\sigma} > k''''\right).$$

Now an important observation: for any $\mu_1 > 0$ the MP test is the same, which means that it is actually a UMP at level α for testing:

$$H_0 : \mu = 0 \text{ and } H_1 : \mu > 0.$$

So we see no μ_1 dependence here and it's really nice that I can test against any $\mu_1 > 0$ at once. Similarly, we could derive a distinct UMP test for testing against $H_1 : \mu < 0$. Unfortunately, no UMP test exists for testing

$$H_0 : \mu = 0 \text{ and } H_1 : \mu \neq 0$$

because the $\mu > 0$ test dominates the $\mu < 0$ test in the $\mu > 0$ scenario and vice versa i.e. the shape of the most powerful rejection rejection (whether we reject for $\bar{x} > k$ or whether we reject for $\bar{x} < k$) is dependent on the sign of μ_1.