LECTURE 8: TESTING

Michal Kolesár*

September 23, 2024

REFERENCE CB, Chapter 8.1, 8.3.1, and 8.3.4–5, or HMC, Chapter 4.5–4.6, H22, Chapter 13.

1. HYPOTHESES

A hypothesis is a statement, which is either true or false, about the population distribution that generated the data X. If we have a statistical model, $X \sim F(\cdot \mid \theta)$, $\theta \in \Theta$, then the hypothesis partitions the parameter space into two parts, $\Theta = \Theta_0 \cup \Theta_1$, such that the hypothesis is true if $\theta \in \Theta_0$, and false otherwise. Typically, the hypothesis is called the *null hypothesis*, and we write $H_0 \colon \theta \in \Theta_0$. The class of alternatives (also called the *alternative hypothesis*) is written as $H_1 \colon \theta \in \Theta_1$. If a hypothesis uniquely identifies the distribution of the data (the parameter space under that hypothesis only has one element), it is called *simple*. Otherwise, the hypothesis is called *composite*.

Example 1. Let $X_1, \ldots, X_n$ be a random sample from distribution $\mathcal{N}(\theta, \sigma^2)$ with σ^2 known, and $\theta \in \Theta$. If $\Theta = \mathbb{R}$, and $\Theta_0 = \{\theta \colon \theta \leq \theta_0\}$, then $H_0 \colon \theta \leq \theta_0$ and $H_1 \colon \theta > \theta_0$. Both hypotheses are composite. Another example: if $\Theta_0 = \{\theta_0\}$, then $H_0 \colon \theta = \theta_0$ and $H_1 \colon \theta \neq \theta_0$. Here the null is simple. ⊠

It is customary to mention both the null and the alternative hypotheses since the full parameter space Θ is often not explicitly specified.

2. TESTING

Testing a hypothesis means that we need to decide, based on the value of a random variable X, whether some hypothesis H_0 that has been formulated is correct. The choice here lies between only two decisions: *accept* or *reject* the hypothesis, so that, viewed as a decision problem, the action space is $\mathcal{A} = \{\text{accept}, \text{reject}\}$. (Some people argue that "do not reject" is perhaps more accurate).

*Email: mkolesar@princeton.edu.

A non-randomized decision rule splits the sample space $\mathcal{X}$ into two parts, S_0 and S_1, the acceptance and *critical* regions. If we observe $X \in S_1$, then we reject H_0, otherwise we accept it. If we denote the decision to reject the test as "1", then we can view the decision rule as a function $\phi\colon \mathcal{X} \mapsto \{0,1\}$. If S_1 is the critical region, then

$$\phi(x) = \mathbb{1}\{x \in S_1\}.$$

Typically, if our data is $X = (X_1, \ldots, X_n)$, then the critical region can be written as $S_1 = \{T(X_1, \ldots, X_n) < c\}$ or $S_1 = \{T(X_1, \ldots, X_n) > c\}$ for some statistic T and a cutoff, also called a *critical value*, $c \in \mathbb{R}$. The critical value c might depend both on the null and the alternative.

Since we can take two actions, and the hypothesis is either true or false, a testing problem has four possible outcomes:

	H_0 true	H_1 true
Accept	Correct decision	Type II error
Reject	Type I error	Correct decision

For any $\theta \in \Theta$, let

$$\beta_\phi(\theta) = E_\theta[\phi(X)],$$

denote the *power function* of the test ϕ. Often we drop the subscript, and just write $\beta(\theta)$. The probability of making a Type I error when $\theta_0 \in \Theta_0$ is the true value is $\beta(\theta_0)$. The probability of making a Type II error when $\theta_1 \in \Theta_1$ is the true value is $1 - \beta(\theta_1)$.

Ideally, we'd like to keep both types of error small. We can always make the probability of type I error smaller by making $\phi(x)$ smaller, and make the probability of type II error smaller by making $\phi(x)$ bigger. For instance, if our test is given by $\phi(x) = \mathbb{1}\{\bar{x}_n < c\}$, then making c small reduces type I error, and increases type II error. So there is a trade-off.

The consequences of type I and type II errors are often quite different, so it makes sense to treat them asymmetrically. One way of doing this is to set up the problem as a decision problem, and assign loss 1 to a type I error and loss λ to a type II error. Then the risk is

$$R(\theta, \phi) = \mathbb{1}\{\theta \in \Theta_0\}\beta_\phi(\theta) + \lambda\, \mathbb{1}\{\theta \in \Theta_1\}(1 - \beta_\phi(\theta)).$$

In practice, rather than choosing λ, it is customary to instead impose a bound $\alpha \in (0,1)$, called the *level of significance*, on the risk we're willing to accept under the null. That is, we require that

$$E_\theta[\phi(X)] = \beta_\phi(\theta) = P_\theta(X \in S_1) \leq \alpha \quad \text{for all } \theta \in \Theta_0. \tag{1}$$

Subject to the level constraint, we then seek to minimize the probability of making a type-II error, or, equivalently, maximize the *power* of the test over Θ_1. Although usually

the level constraint implies that $\sup_{\theta \in \Theta_0} P_\theta(X \in S_1) = \alpha$, it is useful to introduce a separate term for the left-hand side; the maximum null rejection probability is called the *size* of the test, so that the requirement is that the *size of the test may not exceed a given level of significance*. Many people use the terms "size" and "level" loosely and sometimes interchangeably.

Since the power of the test depends on the true parameter value, it is possible that one test has maximal power among all tests with a given level at one parameter value while another test has maximal power at some other parameter value (i.e. the risk functions cross). So it is possible that there is no *uniformly most powerful test*. In this situation the researcher should use some additional criteria to choose a test. You may recall from our earlier decision theory discussion that this amounts to

1. impose unbiasedness (which for tests we will define in the next lecture note) or some invariance requirements; or
2. use weighted average power, weighted by some function $w(\theta)$ on Θ_1, or seek to maximize the minimum power under the alternative.

This explains why, in any particular model, a wide variety of tests have typically been proposed in the statistical and econometric literature. However, unlike in estimation problems, there is an important class of problems in which uniformly most powerful tests do exist. We will discuss how to find such as test, if it exists, in the next lecture note. For now, we will use the *ad hoc* method to come up with a test.

Example 2. Suppose we observe $X_i \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 known, and we wish to test $H_0 \colon \mu = \mu_0$ against $H_1 \colon \mu > \mu_0$, so that the parameter space is $\{\mu \in \mathbb{R} \colon \mu \geq \mu_0\}$. Suppose we want a test with level 5%. By a sufficiency argument, the test should depend on the data only through the sufficient statistic $\overline{X}_n$. Furthermore, it seems reasonable to consider tests of the form $\phi(x) = \mathbb{1}\{\bar{x}_n > c\}$ for some c. We need to choose c such that

$$P_{\mu_0}(\overline{X}_n \geq c) = P_{\mu_0}\left(\frac{\overline{X}_n - \mu_0}{\sigma} \geq \frac{c - \mu_0}{\sigma}\right) = 1 - \Phi(\sqrt{n}(c - \mu_0)/\sigma) \leq 0.05.$$

Since the power is a decreasing function of c in this example, one should choose c such that the above display holds with equality, which gives

$$c = \sigma \Phi^{-1}(0.95)/\sqrt{n} + \mu_0.$$

So, our test will reject the null for large values of the z-statistic,

$$\phi(X) = \mathbb{1}\left\{\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq \Phi^{-1}(0.95)\right\}. \qquad \boxtimes$$

The result of any test is either acceptance or rejection of the null hypothesis. At the same time, it would be interesting to know to what extent our decision relies on the particular choice of the significance level. The concept of a *p-value* gives us such a measure: it is the smallest significance level for which we would still reject the null.

Remark 1. This definition is a little different from the definition on p. 252 in HMC, or Section 8.3.2 in CB, but it is more precise in that the common interpretations of p-values rely on this definition, rather definitions that involve statements about "probability of observing data even more extreme than what we actually observe" (p stands for probability, I think due to this definition).

Notice that the p-value is a random variable, and for it to make sense, the critical regions $S_{1,\alpha}$ need to be nested for different significance levels α (this is typically, but not always, the case).

Example 2 (continued). By generalizing the argument to a generic significance level α, our test would reject the null if

$$\phi(X) = \mathbb{1}\left\{ \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \geq \Phi^{-1}(1-\alpha) \right\}.$$

Since Φ^{-1} is monotone in α, the smallest significance level α for which we reject is

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} = \Phi^{-1}(1-\alpha) \implies p = 1 - \Phi\left(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \right).$$

Note again that the p-value is a function of $\overline{X}_n$, and thus is a random variable. What's its distribution under the null? ⊠

Remark 2. One can show more generally that if the rejection regions $S_{1,\alpha}$ are nested and if $P_\theta(X \in S_{1,\alpha}) = \alpha$ for all $\alpha \in (0,1)$ and $\theta \in \Theta_0$, then $P_\theta(p \leq u) = u$ for all $\theta \in \Theta_0$ and all $u \in [0,1]$. You can think about how this result changes if $P_\theta(X \in S_{1,\alpha}) < \alpha$ for some α and some $\theta \in \Theta_0$.

If the p-value is much smaller than 0.05, then our decision is not particularly sensitive to having chosen 0.05 as the significance level. Moreover, reporting the p-value rather than just the decision to reject has an advantage that, once the p-value is reported, any researcher can decide for himself whether he or she accepts or rejects the null hypothesis depending on their own favorite level of the test.

Let us now review some frequent misunderstandings associated with p-values:

1. p-value is not the probability that the null is true. There is no such probability at all since parameters are not random according to the frequentist (classical) approach.
2. p-value is not the probability of falsely rejecting the null. This probability is measured by the size of the test.

3. One minus p-value is not the probability of the alternative being true. Again, there is no such probability since parameters are not random.

4. The level of the test is not determined by a p-value. Instead, once we know the p-value of the test, the level of the test determines whether we accept or reject the null hypothesis.

Example 3. Let $X_1, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. We want to test $H_0 \colon \sigma^2 = \sigma_0^2$ against the alternative $H_1 \colon \sigma^2 < \sigma_0^2$. Note that both hypotheses are composite since both of them contain all possible values of μ. Let us construct a test based on sample variance S_n^2. We know that $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$. Since small values of $(n-1)S_n^2/\sigma_0^2$ are a sign in favor of the alternative, we should reject for small values of this statistic. Under the null, $(n-1)S_n^2/\sigma_0^2 \sim \chi^2(n-1)$. Then a test with level, say, 5%, rejects the null hypothesis if $(n-1)S_n^2/\sigma_0^2 < \chi^{-2}_{0.05}(n-1)$ where $\chi^{-2}_{0.05}(n-1)$ denotes the 5%-quantile of $\chi^2(n-1)$. What is the power of this test? Fix $\sigma_1^2 < \sigma_0^2$. Then

$$P_{\sigma_1^2}\left((n-1)S_n^2/\sigma_0^2 \le \chi^{-2}_{0.05}(n-1)\right) = P_{\sigma_1^2}\left((n-1)S_n^2/\sigma_1^2 \le (\sigma_0^2/\sigma_1^2)\chi^{-2}_{0.05}(n-1)\right)$$
$$= F_{\chi^2(n-1)}((\sigma_0^2/\sigma_1^2)\chi^{-2}_{0.05}(n-1)),$$

where $F_{\chi^2(n-1)}$ denotes the cdf of $\chi^2(n-1)$. So the power of the test increases as σ^2 decreases. Notice that it doesn't depend on μ, only σ^2: in general this may not be the case.

Suppose $n = 101$, $\sigma_0^2 = 1$, and we observe $S_n^2 = 0.9$. What is the p-value of our test? By previous arguments, the smallest significance level α for which we'd reject is given by $(n-1)S_n^2/\sigma_0^2 = \chi_\alpha^{-2}(n-1)$, so the p-value equals

$$p = F_{\chi^2(n-1)}((n-1)S_n^2/\sigma_0^2) = F_{\chi^2(100)}(100 \cdot 0.9/1) = F_{\chi^2(100)}(90) \approx 0.25.$$

Thus, the test with level 5% does not reject the null hypothesis. ⊠

3. PIVOTAL STATISTICS

A statistic is called *pivotal* if its distribution is independent of unknown parameters. Pivotal statistics are useful in testing because one can calculate quantiles of their distributions and, thus, critical values for tests based on these statistics. For instance, $(n-1)S_n^2/\sigma_0^2$ from Example 3 is pivotal under the null since its distribution does not depend on the *nuisance parameter* μ.

Example 4. Suppose we observe $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and wish to test the null $H_0 \colon \mu = \mu_0$ against the alternative $H_1 \colon \mu \neq \mu_0$. Then the distribution of the t-statistic $T_n = (\overline{X}_n - \mu_0)/\sqrt{S_n^2/n}$ is $t(n-1)$ under the null, and it doesn't in particular depend on σ^2: it's pivotal. Since, intuitively, large values of $|T_n|$ are evidence against the null, this suggests

a test of the form

$$\phi(X) = \mathbb{1}\{|T_n| \geq c\}.$$

Since the probability density function (PDF) of t-distribution is symmetric around, the appropriate critical value c is given by $t^{-1}_{0.975}(n-1)$, the 0.975 quantile of the t-distribution. Indeed,

$$\begin{aligned} P_{\mu_0}\left(|T_n| \geq t^{-1}_{0.975}(n-1)\right) &= P_{\mu_0}(T_n < -t^{-1}_{0.975}) + P_{\mu_0}(T_n > t^{-1}_{0.975}) \\ &= P_{\mu_0}(T_n < t^{-1}_{0.025}) + P_{\mu_0}(T_n > t^{-1}_{0.975}) \\ &= 0.025 + 0.025 = 0.05. \end{aligned}$$

⊠

Example 5. Let $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ be independent random samples from distributions $\mathcal{N}(\mu_X, \sigma^2_X)$ and $\mathcal{N}(\mu_Y, \sigma^2_Y)$, with unknown σ^2_X and σ^2_Y. We want to test the null $H_0\colon \mu_X = \mu_Y$ against $H_1\colon \mu_X > \mu_Y$.

A natural place to start is to note that if the null hypothesis is true, then $\overline{X}_n$ should be close to $\overline{Y}_n$ with high probability. But $\overline{X}_n - \overline{Y}_n \sim \mathcal{N}(0, \sigma^2_X/n_X + \sigma^2_Y/n_Y)$ with σ^2_X and σ^2_Y unknown. Another possibility is to consider

$$t = \frac{\overline{X}_n - \overline{Y}_n}{\sqrt{S^2_X/n_X + S^2_Y/n_Y}},$$

but this doesn't work either as the exact distribution of $S^2_X/n_X + S^2_Y/n_Y$ is not pleasant:

$$S^2_X/n_X + S^2_Y/n_Y \sim \frac{\sigma^2_X}{n_X}\frac{\chi^2(n_X-1)}{n_X-1} \frac{\sigma^2_Y}{n_Y}\frac{\chi^2(n_Y-1)}{n_Y-1}$$

Indeed, there are no pivotal statistics here, and finding a test with good finite-sample properties is hard here: this problem is known as the *Behrens-Fisher problem*, and people thought about this problem since Behrens (1929) and Fisher (1939). People usually appeal to asymptotics here and argue that, by the law of large numbers, as $n_X, n_Y \to \infty$, $S^2_X/\sigma^2_X \xrightarrow{p} 1$ and $S^2_Y/\sigma^2_Y \xrightarrow{p} 1$, so that

$$\begin{aligned} \frac{S^2_X/n_X + S^2_Y/n_Y}{\sigma^2_X/n_X + \sigma^2_Y/n_Y} &= \frac{(\sigma^2_X/n_X)(S^2_X/\sigma^2_X) + (\sigma^2_Y/n_Y)(S^2_Y/\sigma^2_y)}{\sigma^2_X/n_X + \sigma^2_Y/n_Y} \\ &= S_X/\sigma^2_X + \frac{\sigma^2_Y/n_Y}{\sigma^2_X/n_X + \sigma^2_Y/n_Y}(S^2_Y/\sigma^2_Y - S^2_X/\sigma^2_X) \\ &\xrightarrow{p} 1. \end{aligned}$$

Thus, by the Slutsky's theorem, $t \Rightarrow \mathcal{N}(0,1)$. So we can use quantiles of standard normal distribution to form a test with size approximately equal to the required level of the test. This gives us a test with "asymptotically correct size".

If n_X or n_Y is small, then the asymptotics may not be accurate. This problem resurfaced in econometrics recently due to small sample sizes in randomized trials, and a

generalization of this problem appears when the data are "clustered", which is again very common in empirical work. There is a growing econometric literature considering various solutions. As an aside, two classic approaches are to use critical values from a t-distribution with $\min(n_X - 1, n_Y - 1)$ degrees of freedom, which yields valid, but conservative tests (Mickey and Brown 1966), or to use a degrees of freedom adjustment due to Welch (1951). ⊠

REFERENCES

Behrens, Walter Ulrich. 1929. "Ein Beitrag Zur Fehlerberechnung Bei Wenigen Beobachtungen." *Landwirtschaftliche Jahrbücher* 68:807–837.

Fisher, Ronald Aylmer. 1939. "The Comparison of Samples with Possibly Unequal Variances." *Annals of Eugenics* 9, no. 2 (June): 174–180. https://doi.org/10.1111/j.1469-1809.1939.tb02205.x.

Mickey, M. Ray, and Morton B. Brown. 1966. "Bounds on the Distribution Functions of the Behrens-Fisher Statistic." *The Annals of Mathematical Statistics* 37, no. 3 (June): 639–642. https://doi.org/10.1214/aoms/1177699457.

Welch, Bernard Lewis. 1951. "On the Comparison of Several Mean Values: An Alternative Approach." *Biometrika* 38, no. 3 (December): 330–336. https://doi.org/10.1093/biomet/38.3-4.330.