Econ 481-3
Topics in Econometrics

Spring 2021
Ver. April 22, 2021

Northwestern University

Lecture Notes by

IVAN A. CANAY

Department of Economics
Northwestern University

Contents

Part I

Instrumental Variables 101

Lecture 1

Selection on Observables

There are no lecture notes for this topic. You are supposed to read two papers and the slides we used in class.

Bibliography

Lecture 2

Roy Models and LATE

There are no lecture notes for this topic. You are supposed to read two papers and the slides we used in class.

Bibliography

Lecture 3

Marginal Treatment Effects

There are no lecture notes for this topic. You are supposed to read two papers and the slides we used in class.

Bibliography

Lecture 4

Extrapolation

There are no lecture notes for this topic. You are supposed to read two papers and the slides we used in class.

Bibliography

Lecture 5

Outcome Tests

There are no lecture notes for this topic. You are supposed to read two papers and the slides we used in class.

Bibliography

Part II

Understanding Asymptotic Approximations

Lecture 6

Asymptotic Comparisons of Tests I[1]

Consider the following generic version of a testing problem. One observes data $X_i, i = 1, \ldots, n$ i.i.d. with distribution $P \in \mathbf{P} = \{P_\theta : \theta \in \Theta\}$ and wishes to test the null hypothesis $H_0 : \theta \in \Theta_0$ versus the alternative $H_1 : \theta \in \Theta_1$. A test is simply a function $\phi_n = \phi_n(X_1, \ldots, X_n)$ that returns the probability of rejecting the null hypothesis after observing $X_1, \ldots, X_n$. For example, ϕ_n might be the indicator function of a certain test statistic $T_n = T_n(X_1, \ldots, X_n)$ being greater than some critical value $c_n(1-\alpha)$. The test is said to be (pointwise) asymptotically of level α if,

$$\limsup_{n\to\infty} E_\theta [\phi_n] \leq \alpha, \quad \forall \theta \in \Theta_0 \ .$$

Such tests include: Wald tests, quasi-likelihood ratio tests, and Lagrange multiplier tests. Suppose one is given two different tests of the same null hypothesis, $\phi_{1,n}$ and $\phi_{2,n}$, and both tests are (pointwise) asymptotically of level α. How can one choose between these two competing tests of the same null hypothesis? We will now explore the answer to this question in the context of a specific example.

6.1 A Symmetric Location Model

Suppose P_θ is the distribution with density $f(x-\theta)$ on the real line (w.r.t. Lebesgue measure). Suppose further that f is symmetric about 0 and that it's median, 0, is unique. Because f is symmetric about 0, $f(x-\theta)$ is symmetric about θ. We also have that $E_\theta[X] = \theta$ and $\text{med}_\theta[X] = \theta$. Finally, suppose that the variance of P_0 is positive and finite; that is, $\sigma_0^2 = \int x^2 f(x) dx \in (0, \infty)$.

[1]Today's notes are based on Azeem Shaikh's notes. I want to thank him for kindly sharing them.

Notice that we could take f to be the density of a normal distribution and satisfy all of our assumptions. But many other choices of f satisfy these assumptions. For example, we could take f to be the uniform density on $[-1, 1]$, the logistic density, or the Laplace density.

Suppose $\Theta_0 = \{0\}$ and $\Theta_1 = \{\theta \in \mathbf{R} : \theta > 0\}$; i.e., we wish to test the null hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta > 0$. How could we test this null hypothesis?

One such test is of course based on the familiar t-statistic:

$$\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} \ ,$$

where

$$\bar{X}_n = \frac{1}{n} \sum_{1 \le i \le n} X_i \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{1 \le i \le n} (X_i - \bar{X}_n)^2 \ .$$

Under the assumptions above, the CLT, and the CMT, we get

$$\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} \xrightarrow{d} N(0, 1)$$

under P_0. Thus, we may take

$$\phi_{1,n} = I\left\{ \frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} > z_{1-\alpha} \right\}$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution. Obviously, this test is asymptotically of level α (because $z_{1-\alpha}$ is a continuity point of the standard normal distribution).

A second test is based off of the following observation. Since f has median 0 under the null hypothesis, the number of positive and negative observations should be roughly equal (at least asymptotically). This suggests a test based on the test statistic:

$$\frac{1}{n} \sum_{1 \le i \le n} I\{X_i > 0\} \ .$$

How does this statistic behave under the null hypothesis? We can compute that

$$E_0[I\{X_i > 0\}] = P_0\{X_i > 0\} = 1 - F(0) = \frac{1}{2}$$

and thus

$$V_0[I\{X_i > 0\}] = F(0)(1 - F(0)) = \frac{1}{4} \ .$$

Thus, by the CLT we have that

$$\frac{2}{\sqrt{n}} \sum_{1 \le i \le n} (I\{X_i > 0\} - \frac{1}{2}) \xrightarrow{d} N(0, 1) \ .$$

So, we could take

$$\phi_{2,n} = I\left\{ \frac{2}{\sqrt{n}} \sum_{1 \le i \le n} \left(I\{X_i > 0\} - \frac{1}{2} \right) > z_{1-\alpha} \right\} .$$

This test is known as the sign test. Obviously, this test is also asymptotically of level α.

6.2 A Naive Approach

It is natural to base comparisons of two different tests on their power functions. The power function of a test is the function $\pi_n(\theta) = E_\theta[\phi_n]$; i.e., it is the probability of rejecting the null hypothesis as a function of the unknown parameter θ. In this problem it will be difficult to compare the finite-sample power functions of the two tests, but we may try to do so in an asymptotic sense. To this end, let's compute the power functions of each of the above two tests at a fixed $\theta > 0$.

Let's start with the t-test. The key trick is to realize that

$$\begin{aligned} \pi_{1,n}(\theta) &= P_\theta \left\{ \frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} > z_{1-\alpha} \right\} \\ &= P_0 \left\{ \frac{\sqrt{n}\bar{Y}_n + \sqrt{n}\theta}{\hat{\sigma}_n} > z_{1-\alpha} \right\} \\ &= P_0 \left\{ \frac{\sqrt{n}\bar{Y}_n}{\hat{\sigma}_n} > z_{1-\alpha} - \frac{\sqrt{n}\theta}{\hat{\sigma}_n} \right\} , \end{aligned}$$

where $Y_i = X_i - \theta$ is distributed according to P_0. Importantly, we have done this in the denominator, too, using the fact that

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{1 \le i \le n} (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{1 \le i \le n} (Y_i - \bar{Y}_n)^2 .$$

Since $\frac{\sqrt{n}\bar{Y}_n}{\hat{\sigma}_n}$ converges in distribution to a standard normal under P_0 and $z_{1-\alpha} - \frac{\sqrt{n}\theta}{\hat{\sigma}_n}$ diverges in probability to $-\infty$ under P_0, it follows that

$$\pi_{1,n}(\theta) \to 1$$

for every $\theta > 0$.

Now let's consider the sign test. Begin by considering the behavior of

$$\frac{1}{n} \sum_{1 \le i \le n} I\{X_i > 0\}$$

under P_θ. Using the same trick as above, it is easy to compute that

$$\begin{aligned} E_\theta[I\{X_i > 0\}] &= P_\theta \{X_i > 0\} \\ &= P_0 \{Y_i > -\theta\} \\ &= 1 - F(-\theta) \ , \end{aligned}$$

which implies that

$$V_\theta[I\{X_i > 0\}] = F(-\theta)(1 - F(-\theta)) \ .$$

Thus, by the central limit theorem for i.i.d. observations, we have that

$$S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (I\{X_i > 0\} - (1 - F(-\theta)))$$

converges in distribution to $N(0, F(-\theta)(1 - F(-\theta)))$. We can now see that

$$\begin{aligned} \pi_{2,n}(\theta) &= P_\theta \left\{ \frac{2}{\sqrt{n}} \sum_{1 \le i \le n} \left(I\{X_i > 0\} - \frac{1}{2} \right) > z_{1-\alpha} \right\} \\ &= P_\theta \left\{ 2S_n(\theta) > z_{1-\alpha} - 2\sqrt{n} \left(\frac{1}{2} - F(-\theta) \right) \right\} \ . \end{aligned}$$

Because f is symmetric about 0, $F(-\theta) < \frac{1}{2}$. We can now conclude as before that

$$\pi_{2,n}(\theta) \to 1$$

for every $\theta > 0$.

So, we see that a pointwise comparison of power functions of the two tests is completely uninformative. Both tests have power tending to 1 against any fixed alternative $\theta > 0$. In general, tests that have power tending to 1 against any fixed $\theta \in \Theta_1$ are said to be consistent. Any reasonable test will be consistent, so consistency is too weak of a requirement to be of use when trying to choose among different tests.

6.3 Local Asymptotic Power

Here, as always, there are an innumerable number of ways of embedding our situation with a sample of size n in a sequence of hypothetical situations with sample sizes larger than n. When choosing among these different asymptotic frameworks, it is important to keep in mind that what we are really interested in is the finite-sample behavior of the power function; that is, the behavior of the power function for our sample of size n. In the preceding section, we have shown that the power tends to 1 at any fixed $\theta > 0$ as n tends to infinity. Of course, in our sample of size n we know that the

power is not 1 uniformly for $\theta > 0$. It may be very close to 1 for θ "far" from 0, but for θ "close" to 0 we would expect the finite-sample power function to be < 1. Of course, what we mean by "far" and "close" will change with our sample size n. Our asymptotic framework should reflect this fact. The above framework in which the alternative $\theta > 0$ is fixed does not. This suggests that we should consider the behavior of the power function evaluated at a sequence of alternatives θ_n, where θ_n tends to 0 at some rate. One can think of this as providing a locally asymptotic approximation to the power function.

It turns out that if θ_n tends to 0 slowly enough, then the power function will still tend to 1 as n tends to infinity. This follows from the following useful fact: If for every $\epsilon > 0$, $E_n(\epsilon) \to 1$, then there exists a sequence ϵ_n tending to 0 slowly enough so that $E_n(\epsilon_n) \to 1$. I won't prove this fact, but it isn't too hard to do it yourself. You can also find a proof in David Pollard's *A User's Guide to Measure-Theoretic Probability.*

Likewise, if θ_n tends to 0 quickly enough, then for asymptotic purposes it's as if $\theta_n = 0$. For any such sequence, the power function tends to α as n tends to infinity in each of the above two examples.

There is a delicate rate in between the two extremes above such that if θ_n tends to 0 at this rate, then the power will tend to a limit in $(\alpha, 1)$. This rate may be different in different problems, but in problems such as this one in which the distribution depends on θ in a "smooth" way it must be that $\theta_n = O(\frac{1}{\sqrt{n}})$. So, we will consider sequences $\theta_n = \frac{h}{\sqrt{n}}$, where $h \in \mathbf{R}$.

Let's again consider the t-test first. The calculation will be very similar to the one in the preceding section for the t-test. An important distinction is that now we must consider a triangular array of random variables because the distribution of the data is changing with each n. For each n, let $X_{i,n}, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution P_{θ_n}. The trick, as before, will be to write the power in terms of $Y_{i,n} = X_{i,n} - \theta_n$, which is distributed according to P_0. We can now see that,

$$\begin{aligned}
\pi_{1,n}(\theta_n) &= P_{\theta_n}\left\{\frac{\sqrt{n}\bar{X}_{n,n}}{\hat{\sigma}_{n,n}} > z_{1-\alpha}\right\} \\
&= P_0\left\{\frac{\sqrt{n}\bar{Y}_{n,n} + \sqrt{n}\theta_n}{\hat{\sigma}_{n,n}} > z_{1-\alpha}\right\} \\
&= P_0\left\{\frac{\sqrt{n}\bar{Y}_{n,n}}{\hat{\sigma}_{n,n}} > z_{1-\alpha} - \frac{h}{\hat{\sigma}_{n,n}}\right\} .
\end{aligned}$$

Since the distribution of $Y_{i,n}$ is no longer changing with n, our analysis from before applies and we see that

$$\frac{\sqrt{n}\bar{Y}_{n,n}}{\hat{\sigma}_{n,n}} \xrightarrow{d} N(0,1)$$

under P_0. Since $\hat{\sigma}_{n,n}$ converges in probability under P_0 to σ_0, we have that

$$\pi_{1,n}(\theta_n) \to 1 - \Phi\left(z_{1-\alpha} - \frac{h}{\sigma_0}\right) .$$

This limit is called the local asymptotic power function of the t-test. Notice that it depends on the so-called local parameter h.

A remark on interpretation is warranted here. We are really only interested in the power of the test at a single $\theta > 0$, not a sequence θ_n. So, how should we use the above approximation in practice? Given a sample of size n and a $\theta > 0$, we can solve for the corresponding value of h by equating θ and θ_n. By doing so, we find that $h = \sqrt{n}\theta$. Plugging this value of h into the above expression, we get our approximation to the power of the test at θ.

Now let's consider the sign test. Begin as before by considering the behavior of

$$\frac{1}{n}\sum_{1\leq i\leq n} I\{X_{i,n} > 0\}$$

under P_{θ_n}. Our earlier analysis shows that

$$E_{\theta_n}[I\{X_{i,n} > 0\}] = 1 - F(-\theta_n) ,$$

and

$$V_{\theta_n}[I\{X_{i,n} > 0\}] = F(-\theta_n)(1 - F(-\theta_n)) .$$

We'd like to assert that

$$S_n(\theta_n) = \frac{1}{\sqrt{n}}\sum_{1\leq i\leq n} (I\{X_{i,n} > 0\} - (1 - F(-\theta_n)))$$

converges in distribution under P_{θ_n} to a normal distribution. To do this, we will need a central limit theorem for a triangular array. The most general such theorem is the Lindeberg-Feller central limit theorem. Here's a special case of it:

Theorem 6.1 *For each n, let $Z_{n,i}, i = 1, ..., n$ be i.i.d. with distribution P_n. Suppose $E_n[Z_{n,i}] = 0$ and $V_n[Z_{n,i}] = \sigma_n^2 < \infty$. If for each $\epsilon > 0$*

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} E_n[Z_{n,i}^2 I\{|Z_{n,i}| > \epsilon\sqrt{n}\sigma_n\}] = 0$$

then

$$\frac{\sqrt{n}\bar{Z}_{n,n}}{\sigma_n} \xrightarrow{d} N(0,1)$$

under P_n.

For the general version of the Lindeberg-Feller central limit theorem see, for example, Theorem 11.2.5 of Romano and Lehman (2005). For a proof see Theorem 27.2 of Billingsley (1995).

So let's apply the theorem with

$$\begin{aligned} Z_{n,i} &= I\{X_{i,n} > 0\} - (1 - F(-\theta_n)) \\ \sigma_n^2 &= F(-\theta_n)(1 - F(-\theta_n)) \ . \end{aligned}$$

For any fixed h, σ_n is also bounded away from 0 because $F(0) = \frac{1}{2}$, F is continuous by assumption (it's the integral of f), and $\theta_n \approx 0$ for large n. We also have that σ_n is bounded from above because F is bounded. Finally, we have that $Z_{n,i}$ is bounded because I and F are both bounded. Therefore, the condition required in the theorem holds trivially in this case and then,

$$\frac{S_n(\theta_n)}{\sigma_n} \xrightarrow{d} N(0,1) \quad \text{or} \quad 2S_n(\theta_n) \xrightarrow{d} N(0,1)$$

under P_{θ_n}, since $\sigma_n^2 \to F(0)(1 - F(0)) = \frac{1}{4}$.

We can now finish our analysis for the sign test. We have that

$$\begin{aligned} \pi_{2,n}(\theta_n) &= P_{\theta_n}\left\{ \frac{2}{\sqrt{n}} \sum_{1 \le i \le n} \left(I\{X_{i,n} > 0\} - \frac{1}{2} \right) > z_{1-\alpha} \right\} \\ &= P_{\theta_n}\left\{ 2S_n(\theta_n) > z_{1-\alpha} - 2\sqrt{n}\left(\frac{1}{2} - F(-\theta_n) \right) \right\} \ . \end{aligned}$$

Since F is differentiable by assumption (with derivative equal to f), we see that

$$F(-\theta_n) = F(0) - f(0)\theta_n + o(\theta_n) \ ,$$

and so

$$\sqrt{n}\left(\frac{1}{2} - F(-\theta_n) \right) = \sqrt{n}\left(F(0) - F(-\theta_n) \right) = \sqrt{n}\theta_n f(0) + \sqrt{n}o(\theta_n) \to hf(0)$$

assuming f is continuous at 0. Together with the result about the asymptotic normality of $S_n(\theta_n)$ above, we find that

$$\pi_{2,n}(\theta_n) \to 1 - \Phi\left(z_{1-\alpha} - 2hf(0) \right) \ .$$

We are now (finally) in a position to compare these two tests based on their local asymptotic power functions. It is easy to see that if $2f(0) > \frac{1}{\sigma_0}$, then the sign test will be preferred to the t-test in a local asymptotic power sense; otherwise, the t-test will be preferred to the sign test.

If f is the normal density, then we know that the t-test should be uniformly most powerful for testing the null hypothesis. Reassuringly, if we

plug in the standard normal density for f, we find that the above analysis bears this out. Likewise, if f is the density of a logistic or a uniform distribution, then the t-test is preferred to the sign test.

If, on the other hand, we consider distributions with "fatter" tails, we find that the situation is reversed. For example, if we take f to be the density of a Laplace distribution, the above analysis implies that the sign test is preferred to the t-test in a local asymptotic power sense. In fact, we can make the ratio of $2f(0)$ to $1/\sigma_0$ arbitrarily large by considering densities f with more and more mass in the tails. Thus, the moral of this story is that if the underlying distribution is symmetric, then, the t-test, while preferred for many distributions, is not as robust as the sign test to "fat" tails (and can in fact be arbitrarily worse than the sign test!).

The square of the ratio of $2f(0)$ to $1/\sigma_0$ is sometimes referred to as the asymptotic relative efficiency of the sign test w.r.t. the t-test, i.e.,

$$ARE_{2,1} = 4f(0)^2\sigma_0^2 \ .$$

Asymptotic relative efficiency is defined analogously for other pairs of tests.

Bibliography

Billingsley, P. (1995): *Probability and Measure*, Wiley-Interscience.

Pollard, D. (2002): *A User's Guide to Measure Theoretic Probability*, Cambrigde University Press, New York.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Lecture 7

Contiguity

Today's lecture is about a technique to obtain the limit distribution of a sequence of statistics under underlying laws Q_n from a limiting distribution under laws P_n. This will be particularly useful to compute local asymptotic power of different statistics. We already did this in the first lecture when we studied the asymptotic local power of the t-test and the sign test for testing the location of a statistical model. However, extending such analysis to more complicated models is typically too complicated and becomes intractable. Today we will introduce an alternative way of computing local asymptotic power that is based on the idea of contiguity.

7.1 Absolute continuity and likelihood ratios

Let P and Q be measures on a measurable space $(\Omega, \mathcal{A})$. Then Q is *absolutely continuous* with respect to P if for every measurable set A we have that $P\{A\} = 0$ implies $Q\{A\} = 0$; this is denoted by $Q << P$. Furthermore, P and Q are *orthogonal* if Ω can be partitioned as $\Omega = \Omega_P \cup \Omega_Q$ with $\Omega_P \cap \Omega_Q = \emptyset$ and $P\{\Omega_Q\} = Q\{\Omega_P\} = 0$. Orthogonality is denoted by $P \perp Q$.

Theorem 7.1 (Radon-Nikodym) *Suppose Q and P are probability measures on $(\Omega, \mathcal{A})$. Then $Q << P$ if and only if there exists a measurable function $L(x)$ such that,*

$$Q\{A\} = \int_A L(x)dP, \text{ for all } A \in \mathcal{A} .$$

The function $L(x) \equiv dQ(x)/dP(x)$ is called the Radon-Nikodym derivative (or density) or likelihood ratio.

In general two measures P and Q need be neither absolutely continuous nor orthogonal. Suppose these two measures possess densities p and q with

respect to a measure μ. In this case, $\Omega_P = \{p > 0\}$ and $\Omega_Q = \{q > 0\}$. The measure Q can be written as the sum $Q = Q^a + Q^\perp$ of the measures,

$$Q^a\{A\} = Q\{A \cap \{p > 0\}\}; \quad Q^\perp\{A\} = Q\{A \cap \{p = 0\}\} \ .$$

This decomposition is called the *Lebesgue decomposition of* Q with respect to P. In what follows we shall think of the likelihood ratio as a random variable $dQ/dP : \Omega \mapsto [0, \infty)$ and study its law under P.

Lemma 7.1 *Let P and Q be probability measures with densities p and q with respect to a measure μ. Then for the measures Q^a and $Q^\perp$,*

1. *$Q = Q^a + Q^\perp$, $Q^a << P$, $Q^\perp \perp P$.*
2. *$Q^a\{A\} = \int_A (q/p)dP$ for every measurable set A*
3. *$Q << P$ if and only if $Q\{p = 0\} = 0$ if and only if $\int (q/p)dP = 1$*

PROOF. See van der Vaart (1998, Lemma 6.2). ■

Note that the function q/p is a density of Q^a with respect to P. It is denoted dQ/dP (not dQ^a/dP), so that $dQ/dP = q/p$, P-a.s.

Suppose that $T = f(X)$ is an estimator or test statistic. How can we compute the distribution of T under Q if we know how to compute probabilities under P? If the probability measure Q is absolutely continuous with respect to a probability measure P, then the Q-law of a random variable X can be calculated from the P-law of the pair $(X, q/p)$ though the formula,

$$E_Q[f(X)] = \int_{\mathcal{X}} f(x)dQ(x) = \int_{\mathcal{X}} f(x)\frac{q(x)}{p(x)}dP(x) = E_P[f(X)\frac{dQ}{dP}] \ .$$

The validity of this formula depends essentially on the absolute continuity of Q with respect to P, because a part of Q that is orthogonal to P cannot be recovered from any P-law. Thus, under absolute continuity of Q with respect to P, the problem of finding the distribution of $f(X)$ under Q can be in principle obtained form the joint distribution of $f(X)$ and dQ/dP under P.

7.2 Contiguity

Consider now an asymptotic version of the problem. Let $(\Omega_n, \mathcal{A}_n)$ be measurable spaces, each equipped with a pair of probabilities P_n and Q_n. Let T_n be some random vector and suppose the asymptotic distribution of T_n under P_n is easily obtained, but the behavior of T_n under Q_n is also required. For example, if T_n represents a test function for testing P_n versus Q_n, the power of T_n is the expectation under Q_n. Under what conditions

can a Q_n-limit law of random vectors T_n be obtained from suitable P_n-limit laws? The concept is called *contiguity* and essentially denotes a notion of "asymptotic absolute continuity".

At a first glance one could think that asking that the sequences are such that $Q_n << P_n$ for all n would be enough. This is however not true, as the following example suggests.

Example 7.1 *Let $P_n = N(0,1)$, $Q_n = N(\xi_n, 1)$, $\xi_n \to \infty$. It is immediate to see that $Q_n << P_n$ since $P_n\{E_n\} = 0$ implies $Q_n\{E_n\} = 0$ for all n. However, the asymptotic problem is not about probabilities at each n but rather about limit probabilities. This is, we can perfectly have a situation where $P_n\{E_n\} > 0$ for all n and $P_n\{E_n\} \to 0$. Suppose $P_n\{E_n\} \to 0$. Does it follow that $Q_n\{E_n\} \to 0$? The answer in this case is no. Let $E_n = \{x : |x - \xi_n| < 1\}$. We have $Q_n\{E_n\} \approx 0.68$ for all n, but $P_n\{E_n\} \to 0$.*

Definition 7.1 (Contiguity) *Let Q_n and P_n be sequences of measures. We say that Q_n is contiguous w.r.t. to P_n, denoted $Q_n \triangleleft P_n$, if for each sequence of measurable sets A_n, we have that*

$$P_n\{A_n\} \to 0 \Rightarrow Q_n\{A_n\} \to 0 \ .$$

Note that Example 7.1 shows that absolute continuity does not imply contiguity. The following example provides an extension.

Example 7.2 *Suppose P_n is the joint distribution of n i.i.d. observations $X_1, \ldots, X_n$ from $N(0,1)$ and Q_n is the joint distribution of n i.i.d. observations from $N(\xi_n, 1)$. Unless $\xi_n \to 0$, P_n and Q_n cannot be contiguous. For example, suppose $\xi_n > \epsilon > 0$ for all large n and consider $E_n = \{\bar{X}_n > \epsilon/2\}$. By the LLN, $P_n\{E_n\} \to 0$ but $Q_n\{E_n\} \to 1$. In fact, in order for P_n and Q_n to be contiguous, $n^{1/2}\xi_n$ needs to be bounded.*

For probability measures P and Q, Lemma (7.1) implies that the following three statements are equivalent,

$$Q << P, \quad Q\left(\frac{dP}{dQ} = 0\right) = 0, \quad E_P \frac{dQ}{dP} = 1 \ .$$

This equivalence persists if the three statements are replaced by their asymptotic counterparts, as proved by Le Cam. In what follows, we use the notation $\overset{P_n}{\rightsquigarrow}$ to denote $\overset{d}{\to}$ under P_n.

Lemma 7.2 (Le Cam's first lemma) *Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$. Then the following statements are equivalent:*

1. $Q_n \triangleleft P_n$.

2. *If* $dP_n/dQ_n \overset{Q_n}{\rightsquigarrow} U$ *along a subsequence, then* $\Pr(U > 0) = 1$.

3. *If* $dQ_n/dP_n \overset{P_n}{\rightsquigarrow} V$ *along a subsequence, then* $E[V] = 1$.

4. *For any statistic* $T_n : \Omega_n \to \mathbf{R}^k$*: If* $T_n \overset{P_n}{\to} 0$*, then* $T_n \overset{Q_n}{\to} 0$.

PROOF. See van der Vaart (1998, Lemma 6.4). ■

Corollary 7.1 *Let* $dQ_n/dP_n \overset{P_n}{\rightsquigarrow} V$ *and suppose* $\log(V) \sim N(\mu, \sigma^2)$ *(this is,* V *has a log normal distribution). Then* Q_n *and* P_n *are mutually contiguous if and only if* $\mu = -\frac{1}{2}\sigma^2$*, which follows from* $E[V] = \exp(\mu + \frac{1}{2}\sigma^2)$.

Example 7.3 *Contiguity does not imply absolute continuity. Let* $P_n = U[0,1]$*,* $Q_n = U[0, \theta_n]$*,* $\theta_n \to 1$*,* $\theta_n > 1$*. Note that by Le Cam first Lemma,* $dQ_n/dP_n = 1/\theta_n \overset{P_n}{\rightsquigarrow} V = 1$ *so* $Q_n \triangleleft P_n$*. However, it is not true that* $Q_n << P_n$ *since* $P_n\{[1, \theta_n]\} = 0$ *while* $Q_n\{[1, \theta_n]\} > 0$.

Example 7.4 *Let* $P_n = N(0,1)$*,* $Q_n = N(\xi_n, 1)$*. Then,*

$$\log(L_n(X)) = \log\left(\frac{dQ_n}{dP_n}\right) = \xi_n X - \frac{1}{2}\xi_n^2 \ .$$

This converges if $\xi_n \to \xi$ *with* $|\xi| < \infty$ *which yields* $\xi X - \frac{1}{2}\xi^2$ *in the limit. Hence,* $\log L_n \overset{P_n}{\rightsquigarrow} N(-\frac{1}{2}\xi^2, \xi^2)$ *and the relationship between the mean and the variance is satisfied. We then get contiguity for* $|\xi| < \infty$.

Example 7.5 *Suppose* P_n *is the joint distribution of* n *i.i.d. observations* $X_1, \ldots, X_n$ *from* $N(0,1)$ *and* Q_n *is the joint distribution of* n *i.i.d. observations from* $N(\xi_n, 1)$*. Then,*

$$\log(L_n(X_1, \ldots, X_n)) = \xi_n \sum X_i - \frac{n\xi_n^2}{2} \ ,$$

and so

$$\log(L_n(X_1, \ldots, X_n)) \sim N\left(-\frac{1}{2}n\xi_n^2, n\xi_n^2\right) \ .$$

By the arguments similar to that of the previous example, Q_n *is contiguous to* P_n *if and only if* $n\xi_n^2$ *remains bounded, i.e.* $\xi_n = O(n^{-\frac{1}{2}})$*. Think about the case* $\sqrt{n}\xi_n \to \infty$ *(which violates 3 in Lemma 7.2).*

Note that contiguity implies that the sequences of measures P_n and Q_n do not separate asymptotically: given data from P_n or Q_n it is impossible to tell with certainty from which of the two sequences the data is generated, at least in an asymptotic sense, as $n \to \infty$. Indeed, if P_n and Q_n are contiguous, and ϕ_n is a sequence of tests with error probabilities $E_{P_n}[\phi_n]$

for testing the null hypothesis P_n satisfying $E_{P_n}[\phi_n] \to 0$, then the power $E_{Q_n}[\phi_n]$ at the alternative Q_n satisfies $E_{Q_n}[\phi_n] \to 0$ as well.

Actually, contiguity implies more: contiguity is "asymptotic absolute continuity", meaning that it is possible to derive asymptotic probabilities computed under Q_n from those computed under P_n. This is the content of Le Cam's third lemma, which we discuss below.

Contiguity has turned out to be a wonderful tool in many proofs, where one is given a choice to prove convergence in probability to zero under the measure of interest, or under any other convenient, contiguous sequence. However, the application of contiguity that has made it popular is in the comparison of statistical tests. Here one is given a sequence of tests ϕ_n concerning a parameter θ attached to a statistical model $(P_{n,\theta} : \theta \in \Theta)$ and corresponding power functions

$$\pi_n(\theta) = E_{P_{n,\theta}}[\phi_n] \ .$$

If P_{n,θ_0} and P_{n,θ_1} are asymptotically separated, then any "good" sequence of tests of the null hypothesis θ_0 versus the alternative θ_1 will have $\pi_n(\theta_0) \to 0$ and $\pi_n(\theta_1) \to 1$. Such alternatives are not of much interest to compare the quality of two sequences of tests. On the other hand, contiguous alternatives will not allow this type of degeneracy, and hence may be used to pick a best test, or compute a relative efficiency of two given sequences of tests. Such contiguous alternatives may be given through the context, for instance of a parametric model.

To prevent asymptotic separation of alternative hypotheses P_n and Q_n the full force of contiguity is not needed. Contiguity has a further use, which is to alleviate the problem of computing the limiting distribution of a test statistic under a (contiguous) alternative. This technique is skillfully applied to rank procedures in Hájek and Sidák (1967), and has since become a standard tool in the asymptotic analysis of tests. The basic procedure, known as Le Cam's third lemma, is stated below.

Lemma 7.3 (Le Cam's third lemma) *Suppose that*

$$\left(X_n, \log \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} N\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau' & \sigma^2 \end{pmatrix}\right) \ .$$

Then,

$$X_n \overset{Q_n}{\rightsquigarrow} N(\mu + \tau, \Sigma) \ .$$

This Lemma shows that under the alternative distribution Q_n, the limiting distribution of the test statistic X_n is also normal but has mean shifted by $\tau = \lim_{n\to\infty} \mathrm{cov}(X_n, \log(dQ_n/dP_n))$. In the testing situation, with asymptotically normal test statistics X_n, it follows that a change of measure from a null hypothesis to a contiguous alternative induces a change of asymptotic

mean in the test statistics equal to the asymptotic covariance between X_n and $\log \frac{dQ_n}{dP_n}$ and no change of variance. *It follows that good test statistics have a large (asymptotic) covariance with the log likelihood ratios.*

7.3 Wilcoxon signed rank statistic

Now we will use Le Cam's Third Lemma to analyze the local asymptotic power of the Wilcoxon signed rank statistic. For this we will use the same location example we used for the t-test and the sign test. Suppose P_θ is the distribution with density $f(x-\theta)$ on the real line. Suppose further that f is symmetric about 0, so that $f(x-\theta)$ is symmetric about θ. We observe $X_1, \ldots, X_n$ from f and wish to test the null $H_0 : \theta = 0$. The Wilcoxon signed rank statistic serves to test this null and takes the form

$$W_n = n^{-3/2} \sum_{i=1}^{n} R_{i,n}^{+} \operatorname{sign}(X_i) \ ,$$

where

$$\operatorname{sign}(X_i) = \begin{cases} 1 & \text{if } X_i \geq 0 \\ -1 & \text{otherwise} \end{cases} \ ,$$

and

$$R_{i,n}^{+} = \sum_{j=1}^{n} I\{|X_j| \leq |X_i|\}$$

is the rank of $|X_i|$ among $|X_1|, \ldots, |X_n|$. Under the null hypothesis, the behavior of W_n is fairly easy to obtain. If $\theta = 0$, the variables $\operatorname{sign}(X_i)$ are i.i.d. and equal to 1 with probability 1/2 and -1 with probability 1/2. Note that $E_{P_0}[\operatorname{sign}(X_i)] = 0$.

Now let $U_i = G(|X_i|)$ and G be the cdf of $|X_i|$. Note that

$$U_i - n^{-1} R_{i,n}^{+} = U_i - n^{-1} \sum_{j=1}^{n} I\{|X_j| \leq |X_i|\} = o_p(1) \ .$$

What's important is that the above convergence is valid uniformly over $i = 1, \ldots, n$ (Glivenko-Cantelli). Then

$$W_n = n^{-1/2} \sum_{i=1}^{n} n^{-1} R_{i,n}^{+} \operatorname{sign}(X_i) = n^{-1/2} \sum_{i=1}^{n} U_i \operatorname{sign}(X_i) + o_p(1) \ .$$

This shows that the asymptotic distribution of W_n equals the asymptotic distribution of $n^{-1/2} \sum_{i=1}^{n} U_i \operatorname{sign}(X_i)$. Let $Z_i = U_i \operatorname{sign}(X_i)$ and note that Z_i are i.i.d. with mean zero since $\operatorname{sign}(X)$ and $|X|$ are independent under the null. Then

$$E_{P_0}[W_n] = \sqrt{n} E_{P_0}[Z_i] = 0 \ .$$

To get the variance of Z_i under P_0, note that

$$\mathrm{var}_{P_0}[Z_i] = E_{P_0}[U_i^2] = \frac{1}{3} \ ,$$

since $U_i \sim U(0,1)$. Therefore, by the Central Limit Theorem

$$W_n \overset{P_0}{\rightsquigarrow} N\left(0, \frac{1}{3}\right) \ ,$$

as $n \to \infty$ and the test

$$\phi_{3,n} = I\{\sqrt{3}W_n > z_{1-\alpha}\}$$

is (pointwise) asymptotically of level α.

Now it is time to use Le Cam's third lemma to examine the power of this test under the sequence of alternatives P_{θ_n} where $\theta_n = h/\sqrt{n}$. Le Cam's lemma suggest that we look at,

$$(W_n, \log(dP_{\theta_n}/dP_0)) \ .$$

We first consider the case where $P_{\theta_n} = N(\theta_n, 1)$ and $P_0 = N(0,1)$. In this case, note that

$$p_{\theta_n}(X_1, \ldots, X_n) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp[-\frac{1}{2}(X_i - \theta_n)^2]$$

and then,

$$\begin{aligned} \log L_n = \log(dP_{\theta_n}/dP_0) &= \log \frac{e^{-\frac{1}{2}\sum_{i=1}^n (X_i^2 - 2X_i\theta_n + \theta_n^2)}}{e^{-\frac{1}{2}\sum_{i=1}^n X_i^2}} \\ &= \theta_n \sum_{i=1}^{n} X_i - \frac{n}{2}\theta_n^2 \\ &= hn^{-1/2} \sum_{i=1}^{n} X_i - \frac{1}{2}h^2 \ . \end{aligned}$$

We then have

$$(W_n, \log(dP_{\theta_n}/dP_0)) = \left(n^{-1/2} \sum_{i=1}^{n} U_i \,\mathrm{sign}(X_i) \ , \ hn^{-1/2} \sum_{i=1}^{n} X_i - h^2/2 \right) + o_p(1) \ ,$$

which is asymptotically bivariate normal under P_0 with covariance of the cross term

$$\tau = \ \mathrm{cov}_{P_0}\left[G(|X|)\,\mathrm{sign}(X) \ , \ hX\right] = hE_{P_0}\left[G(|X|)|X|\right] = \frac{h}{\sqrt{\pi}} \ ,$$

where the last equality is an exercise of integration. The local asymptotic power under the alternatives P_{θ_n} follows from the conclusion of Le Cam's third lemma, i.e,

$$W_n \overset{P_{\theta_n}}{\rightsquigarrow} N\left(\frac{h}{\sqrt{\pi}}, \frac{1}{3}\right) ,$$

since

$$\begin{aligned}\lim_{n\to\infty} P_{\theta_n}(W_n > z_{1-\alpha}/\sqrt{3}) &= \lim_{n\to\infty} P_{\theta_n}(W_n - h/\sqrt{\pi} > z_{1-\alpha}/\sqrt{3} - h/\sqrt{\pi}) \\ &= 1 - \Phi\left(z_{1-\alpha} - h\sqrt{\frac{3}{\pi}}\right) .\end{aligned}$$

In order to drop the simplifying normal assumption, we can obtain the asymptotic power under the assumption that the model $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean. This is the topic of next class.

Bibliography

Lehmann, E. and J. P. Romano (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Lecture 8

Local Asymptotic Normality

Suppose we observe a sample $X_1, \ldots, X_n$ from a distribution P_θ on some measurable space $(\mathcal{X}, \mathcal{A})$ indexed by a parameter θ in Θ open. Then the full observation is a single observation from the product P_θ^n of n copies of P_θ and the statistical model (also called *statistical experiment*) is completely described as the collection of probability measures $\mathbf{P} = \{P_\theta^n : \theta \in \Theta\}$. Today we will study conditions under which a statistical experiment can be approximated by a Gaussian experiment after a suitable reparametrization. Let's start with the trivial case where the approximation is exact.

Example 8.1 (Normal Location Model) *Suppose $P_\theta = N(\theta, \sigma^2)$, where σ^2 is known. In this case,*

$$\begin{aligned}
\log[dP^n_{\theta_0+h/\sqrt{n}}/dP^n_{\theta_0}] &= -\frac{1}{2\sigma^2}\sum_{i=1}^n \left(X_i - \theta_0 - \frac{h}{\sqrt{n}}\right)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\theta_0)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\theta_0)\frac{h}{\sqrt{n}} - \frac{h^2}{2\sigma^2} \\
&= \frac{h}{\sigma^2} n^{1/2}(\bar{X}_n - \theta_0) - \frac{h^2}{2\sigma^2} \\
&= h\Delta_n - \frac{1}{2} I_{\theta_0} h^2 \ , \qquad (8.1)
\end{aligned}$$

where

$$\Delta_n = n^{1/2}(\bar{X}_n - \theta_0)/\sigma^2 \sim N(0, I_{\theta_0}) \text{ and } I_{\theta_0} = 1/\sigma^2 \ ,$$

under P_{θ_0}. It follows that

$$\log[dP^n_{\theta_0+h/\sqrt{n}}/dP^n_{\theta_0}] \sim N\left(-\frac{1}{2}\frac{h^2}{\sigma^2}, \frac{h^2}{\sigma^2}\right) \text{ under } P_{\theta_0} \ . \qquad (8.2)$$

The expansion in (8.1) *has no reminder, is a linear function of Δ_n (which is exactly sufficient), and a simple quadratic function of h, with the coefficient on h^2 nonrandom (i.e., I_{θ_0}).*

8.1 Local Asymptotic Normality

The traditional regularity conditions for maximum likelihood theory involve existence of two or three derivatives of the density function, together with domination assumptions to justify differentiation under integrals. Le Cam (1970) noted that such conditions are unnecessarily stringent. He showed that the traditional conditions can be replaced by a simple assumption of differentiability in quadratic mean (QMD). In particular, Le Cam showed that it implies a quadratic approximation property for the log-likelihoods known as Local Asymptotic Normality (LAN).

Definition 8.1 *The statistical experiment is called LAN at $\theta_0 \in \Theta$ if there exist a sequence of stochastic vectors Δ_{n,θ_0} and a nonsingular $(k \times k)$ matrix I_{θ_0} such that $\Delta_{n,\theta_0} \xrightarrow{d} N(0, I_{\theta_0})$ under $P^n_{\theta_0}$ and such that,*

$$\log \left[\frac{dP^n_{\theta_0 + h/\sqrt{n}}}{dP^n_{\theta_0}} \right] = h\Delta_{n,\theta_0} - \frac{1}{2} h' I_{\theta_0} h + o_{P_{\theta_0}}(1) \ . \tag{8.3}$$

Traditional arguments to show LAN go as follows. Consider an i.i.d. sequence $X_1, \ldots, X_n$ from a density $p_\theta = dP_\theta/d\mu$ (for some dominating measure μ) such that the map $\theta \mapsto p_\theta$ is twice differentiable. Let $\ell_\theta(x) = \log p_\theta(x)$, with derivatives $\dot{\ell}_\theta(x)$ and $\ddot{\ell}_\theta(x)$ with respect to θ. Now do a Taylor expansion of $\ell_\theta(x)$ for fixed x,

$$\log p_{\theta + h/\sqrt{n}}(x) = \log p_\theta(x) + \frac{h}{\sqrt{n}} \dot{\ell}_\theta(x) + \frac{h^2}{2n} \ddot{\ell}_\theta(x) + o_x(h^2/n) \ , \tag{8.4}$$

where the subscript x in the reminder term denotes its dependence on x. It then follows,

$$\sum_{i=1}^n \log \left(\frac{p_{\theta + h/\sqrt{n}}}{p_\theta}(X_i) \right) = \frac{h}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\theta(X_i) + \frac{h^2}{2n} \sum_{i=1}^n \ddot{\ell}_\theta(X_i) + o_{p_\theta}(1) \ , \tag{8.5}$$

provided the sum of n reminders $o_x(h^2/n)$ is asymptotically negligible (i.e., $o_{p_\theta}(1)$). Here, the expected score is zero, $E_\theta \dot{\ell}_\theta = 0$, and $-E_\theta \ddot{\ell}_\theta = E_\theta \dot{\ell}^2_\theta$ equals the Fisher information for θ. Hence, the first term can be rewritten as $h\Delta_{n,\theta}$, where $\Delta_{n,\theta} \xrightarrow{d} N(0, I_\theta)$ under P_θ by the CLT. Furthermore, the second term in the expansion is asymptotically equivalent to $-1/2h^2 I_\theta$ by the LLN. This expansion can be made rigorous by assuming continuity conditions on the second derivative of $\ell_\theta(x)$. Le Cam's contribution is to show that one can get the benefit of the quadratic expansion without paying the twice-differentiability price usually demanded by such a Taylor expansion.

8.2 Differentiability in Quadratic Mean

How can we get the aforementioned benefit? Le Cam showed all that is required is a single condition that only involves a first derivative: differentiability of the root density $\theta \mapsto \sqrt{p_\theta}$ in quadratic mean as defined below.

Definition 8.2 (QMD) *A model $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ is called differentiable in quadratic mean (or Hellinger differentiable or QMD) at θ if there exists a vector of measurable functions $\eta_\theta = (\eta_{\theta,1}, \ldots, \eta_{\theta,k})'$ such that, as $h \to 0$,*

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h' \eta_\theta \sqrt{p_\theta} \right]^2 d\mu = o(||h||^2) \ , \tag{8.6}$$

where p_θ is the density of P_θ w.r.t. some measure μ.

If the model is QMD for every θ then we say the family if QMD. Usually, $\frac{1}{2} h \eta_\theta(x) \sqrt{p_\theta(x)}$ is the derivative of the map $h \mapsto \sqrt{p_{\theta+h}(x)}$ at $h = 0$ for (almost) every x. In this case,

$$\frac{\partial}{\partial \theta} \sqrt{p_\theta} = \frac{1}{2\sqrt{p_\theta}} \frac{\partial}{\partial \theta} p_\theta = \frac{1}{2} \left(\frac{\partial}{\partial \theta} \log p_\theta \right) \sqrt{p_\theta} \ ,$$

so the function $\eta_\theta(x)$ is $\dot{\ell}_\theta(x) = (\frac{\partial}{\partial \theta} \log p_\theta)$, the score function of the model. Condition (8.6) does not require differentiability of the map $\theta \mapsto p_\theta(x)$ for any single x, but rather differentiability in (quadratic) mean. Many commonly encountered families of distributions are differentiable in quadratic mean, including exponential families and location models with smooth underlying densities. See Chapter 12 of Lehmann and Romano (2005). However, some are not as the following example ilustrates.

Example 8.2 (Uniform Distribution) *The family of uniform distributions on $[0, \theta]$ is nowhere differentiable in quadratic mean. The reason is that the support depends too much on the parameter. Restricting* (8.6) *to the set $\{p_\theta = 0\}$ yields,*

$$P_{\theta+h}(p_\theta = 0) = \int_{p_\theta = 0} p_{\theta+h} d\mu = o(h^2) \ . \tag{8.7}$$

This is not true for the uniform distribution, because, for $h \geq 0$,

$$P_{\theta+h}(p_\theta = 0) = \int_{(\theta, \theta+h]} \frac{1}{\theta + h} dx = \frac{h}{\theta + h} \ . \tag{8.8}$$

Differentiability in quadratic mean (8.6) does not require that all densities p_θ have the same support. But it imposes restrictions on how much it may depend on θ.

The vital ingredient that makes QMD work is the fact that the square root of a density satisfies $\sqrt{p_\theta(x)} \in L^2(\mu)$, where $L^2(\mu)$ denotes the space of functions g such that $\int g^2(x)d\mu(x) < \infty$. But, in fact, $\sqrt{p_\theta}$ is not just an element of $L^2(\mu)$: it is an element with constant norm 1, i.e.,

$$\int (\sqrt{p_\theta})^2 d\mu = \int p_\theta d\mu = 1 \text{ for all } \theta \in \Theta \ .$$

The details of why having a constant norm matters can be found in Pollard's notes "Another Look at Differentiability in Quadratic Mean". It turns out, as shown by Le Cam, that condition (8.6) is exactly what we need to get LAN.

Theorem 8.1 *Suppose that Θ is an open subset of $\mathbf{R}^k$ and that the model $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at θ. Then $E_\theta \dot{\ell}_\theta = 0$ and the Fisher information matrix $I_\theta = E_\theta \dot{\ell}_\theta \dot{\ell}_\theta{}'$ exists. Furthermore,*

$$\log \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h' \dot{\ell}_\theta(X_i) - \frac{1}{2} h' I_\theta h + o_{p_\theta}(1) \ . \tag{8.9}$$

PROOF. See (van der Vaart, 1998, Theorem 7.2) ■

Next, we would like to determine simple sufficient conditions for QMD to hold. Usually one proceeds by showing differentiability of the map $\theta \mapsto \sqrt{p_\theta(x)}$ for almost every x plus μ-equi-integrability (which in turn will imply a convergence theorem for integrals). These conditions are stated in the following lemma.

Lemma 8.1 *For every θ in an open subset of $\mathbf{R}^k$ let p_θ be a μ-probability density. Assume that the map $\theta \mapsto s_\theta \equiv \sqrt{p_\theta(x)}$ is continuously differentiable for every x. If the elements of the matrix*

$$I_\theta = \int (\dot{p}_\theta / p_\theta)(\dot{p}_\theta / p_\theta)' p_\theta d\mu$$

are well defined and continuous in θ, then the map $\theta \mapsto \sqrt{p_\theta}$ is differentiable in quadratic mean with $\eta_\theta = \dot{p}_\theta / p_\theta$.

PROOF. For simplicity we consider the one dimensional case and divide the proof in steps. We wish to prove that

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h \eta_\theta \sqrt{p_\theta} \right]^2 d\mu = o(h^2) \ ,$$

for η_θ as defined above.

Step 1. We first show that $\dot{p}_\theta$ exists and find an expression for $\dot{s}_\theta$. By the chain rule, the map $\theta \mapsto p_\theta(x) = s_\theta^2(x)$ is differentiable for every x with

gradient $\dot{p}_\theta = 2s_\theta \dot{s}_\theta$. Because $s_\theta \geq 0$, its gradient $\dot{s}_\theta$ at a point at which $s_\theta = 0$ must be zero. It follows that

$$\dot{s}_\theta(x) \equiv \frac{\partial}{\partial\theta} s_\theta = \frac{1}{2\sqrt{p_\theta}} \dot{p}_\theta = \frac{1}{2} \eta_\theta(x) \sqrt{p_\theta} \ ,$$

where $\eta_\theta = \dot{p}_\theta / p_\theta$ may be arbitrarily defined if $p_\theta = 0$. By our assumptions, the map

$$\theta \mapsto I_\theta = \int \eta_\theta^2 p_\theta d\mu = 4 \int \dot{s}_\theta^2 d\mu$$

is well defined and continuous.

Step 2. We invoke Vitali's theorem (see Proposition 2.29 van der Vaart, 1998), which states that if $f_n(x) \to f(x)$ for μ-almost every x (both real-valued measurable functions) and

$$\limsup_{n\to\infty} \int f_n^2(x) d\mu(x) \leq \int f^2(x) d\mu(x) < \infty \ , \tag{8.10}$$

it follows that

$$\lim_{n\to\infty} \int |f_n(x) - f(x)|^2 d\mu(x) = 0 \ .$$

We therefore need to check the two conditions. First, since the map $\theta \mapsto s_\theta = \sqrt{p_\theta}$ is continuously differentiable,

$$\frac{1}{h}(s_{\theta+h}(x) - s_\theta(x)) \to \dot{s}_\theta(x) \ \text{ as } \ h \to 0 \ . \tag{8.11}$$

Second, the difference $s_{\theta+h}(x) - s_\theta(x)$ can be written as

$$s_{\theta+h}(x) - s_\theta(x) = \int_\theta^{\theta+h} \dot{s}_v dv = h \int_0^1 \dot{s}_{\theta+uh} du \ .$$

Using this last result we can prove (8.10) by arguing as follows. Note that

$$\left[\frac{1}{h}(s_{\theta+h}(x) - s_\theta(x))\right]^2 = \left[\int_0^1 \dot{s}_{\theta+uh}(x) du\right]^2 \leq \int_0^1 \dot{s}_{\theta+uh}^2(x) du \ ,$$

where the inequality follows from Jensen's inequality. Further note that

$$\begin{aligned} \int \left[\frac{1}{h}(s_{\theta+h}(x) - s_\theta(x))\right]^2 d\mu &\leq \int \int_0^1 \dot{s}_{\theta+uh}^2(x) du d\mu \\ &= \frac{1}{4} \int_0^1 I_{\theta+uh} du \to \frac{1}{4} I_\theta < \infty \ , \end{aligned}$$

where the equality follows from Fubini's Theorem and the last limit holds for $h \to 0$ by continuity of I_θ. Conclude that

$$\lim_{h\to 0} \int \left[\frac{1}{h}(s_{\theta+h}(x) - s_\theta(x))\right]^2 d\mu \leq \frac{1}{4} I_\theta = \int \dot{s}_\theta^2(x) d\mu < \infty \ ,$$

and so condition (8.10) holds. It then follows from Vitali's theorem that

$$\lim_{h\to 0}\int\left[\frac{1}{h}(s_{\theta+h}(x)-s_\theta(x))-\dot{s}_\theta(x)\right]^2 d\mu = 0\ . \tag{8.12}$$

Replacing $s_\theta = \sqrt{p_\theta}$ and $\dot{s}_\theta(x) = \frac{1}{2}\eta_\theta(x)\sqrt{p_\theta}$ completes the proof. ■

Example 8.3 (Location Model) *Let $\{p_\theta(x) = f(x-\theta) : \theta \in \Theta\}$ be a location model, where $f(\cdot)$ is continuously differentiable. Let*

$$\dot{\ell}_\theta(x) = \frac{\dot{p}_\theta}{p_\theta} = \frac{-f'(x-\theta)}{f(x-\theta)} \tag{8.13}$$

if $f(x-\theta) > 0$ and $f'(x-\theta)$ exists and zero otherwise. Assume

$$I_0 = \int \dot{\ell}_0^2(x) f(x)dx < \infty\ , \tag{8.14}$$

Since in this model the Fisher information is equal to I_0 for all θ (just set $y = x-\theta$ in the integral for I_θ), and thus continuous in θ, it follows that the family is QMD.

8.3 Limit Distributions under Contiguous Alternatives

Local asymptotic normality is a convenient tool in the study of the behavior of statistics under "contiguous alternatives". Under the LAN assumption,

$$\log\frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta} \xrightarrow{d} N\left(-\frac{1}{2}h'I_\theta h, h'I_\theta h\right), \tag{8.15}$$

under P_θ, so the sequences of distributions $dP^n_{\theta+h/\sqrt{n}}$ and dP^n_θ are mutually contiguous. With the help of Le Cam's third lemma it allows to obtain limit distributions of statistics under the parameters $\theta + h/\sqrt{n}$, once the limit behavior under θ is known.

The general scheme is as follows. Many sequences of statistics T_n allow an approximation of the type,

$$\sqrt{n}(T_n - \mu_\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_\theta(X_i) + o_{p_\theta}(1)\ . \tag{8.16}$$

By Theorem (8.1), the sequence of log likelihood ratios can be approximated by

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n h'\dot{\ell}_\theta(X_i) - \frac{1}{2}h'I_\theta h + o_{p_\theta}(1)\ . \tag{8.17}$$

The sequence of *joint* averages $n^{-1/2}\sum(\psi_\theta(X_i), h\dot{\ell}_\theta(X_i))$ is asymptotically multivariate normal under P_θ by the CLT and so,

$$\left(\sqrt{n}(T_n-\mu_\theta), \log\frac{dP^n_{\theta+h/\sqrt{n}}}{dP^n_\theta}\right) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ -1/2h'I_\theta h \end{pmatrix}, \begin{pmatrix} E_\theta\psi_\theta\psi_\theta' & E_\theta\psi_\theta h'\dot{\ell}_\theta \\ E_\theta\psi_\theta{}'h\dot{\ell}_\theta & h'I_\theta h \end{pmatrix}\right) .$$

Finally, we can apply Le Cam's third lemma to obtain the limit distribution of $\sqrt{n}(T_n-\mu_\theta)$ under $\theta+h/\sqrt{n}$.

8.3.1 Symmetric Location Model

We conclude today's lecture with an example. Consider the location model from the previous lectures. P_θ is the distribution with density $f(x-\theta)$ on the real line. Suppose further that f is symmetric about 0, so that $f(x-\theta)$ is symmetric about θ. We observe $X_1,\ldots,X_n$ from f and wish to test the null $H_0:\theta=0$. We will use the additional assumption that $\mathbf{P}=\{P_\theta:\theta\in\Theta\}$ is differentiable in quadratic mean. By Example 8.3, $p_\theta(x)=f(x-\theta)$ is QMD at $\theta=0$ if $f(\cdot)$ is absolutely continuous with finite Fisher information,

$$I_0=\int \ell_0^2(x)f(x)dx \ , \tag{8.18}$$

where

$$\dot{\ell}_\theta(x)=\frac{-f'(x-\theta)}{f(x-\theta)} \ . \tag{8.19}$$

It follows by Theorem (8.1) that,

$$\log L_n=\log(dP_{\theta_n}/dP_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n -h\frac{f'(X_i)}{f(X_i)}-\frac{1}{2}h^2I_0+o_p(1) \ . \tag{8.20}$$

We can now easily derive the local asymptotic power of the t-test, the sign-test and the Wilcoxon signed rank test by using Le Cam's third lemma.

T-test

First note that we can define the t-test as an asymptotically linear statistic,

$$t_n=\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n}=\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{X_i}{\sigma}+o_{P_0}(1) \ ,$$

so (8.16) holds for $\psi_\theta(X_i)=X_i/\sigma$. We are interested in the behavior of t_n under the alternative $\theta_n=h/\sqrt{n}$. Although we know this can be obtained by direct means, let us obtain the result using Le Cam's third lemma. The covariance term of interest is,

$$\begin{aligned}\tau &= E_0\left[\frac{X}{\sigma}\times -h\frac{f'(X)}{f(X))}\right]=-\frac{h}{\sigma}\int x\frac{f'}{f}fdx=-\frac{h}{\sigma}\int xf'dx \\ &= \frac{h}{\sigma} \ ,\end{aligned}$$

where the last equality follows from integration by parts. Thus, under P_{θ_n},

$$t_n \xrightarrow{d} N(h/\sigma, 1) \ .$$

Sign Test

Assume the median is unique. The sign test is,

$$S_n = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (I\{X_i > 0\} - 1/2) \ ,$$

so (8.16) holds for $\psi_\theta(X_i) = I\{X_i > 0\} - 1/2$ and then $S_n \xrightarrow{d} N(0, 1/4)$ under P_0. Under P_{θ_n}, $S_n \xrightarrow{d} N(\tau, 1/4)$ by Le Cam's third lemma where the covariance term of interest is

$$\tau = E_0 \left[I\{X > 0\} \times -h \frac{f'(X)}{f(X)} \right] = -h \int_0^\infty f' dx = hf(0) \ .$$

Thus, under the alternative P_{θ_n},

$$S_n \xrightarrow{d} N(hf(0), 1/4) \ .$$

Wilcoxon signed rank test

We previously showed that the Wilcoxon signed rank statistic can be written as,

$$W_n = n^{-1/2} \sum_{i=1}^n U_i \operatorname{sign}(X_i) + o_p(1) \ .$$

where

$$\operatorname{sign}(X_i) = \begin{cases} 1 & \text{if } X_i \ge 0 \\ -1 & \text{otherwise} \end{cases} \ ,$$

and $U_i = G(|X_i|)$ with G the cdf of $|X_i|$. Here (8.16) holds for $\psi_\theta(X_i) = U_i \operatorname{sign}(X_i)$ and then, under P_0,

$$W_n \xrightarrow{d} N(0, 1/3) \ .$$

For the behavior under P_{θ_n}, we compute the covariance again,

$$\begin{aligned} \tau &= E_0 \left[G(|X|) \operatorname{sign}(X) \times -h \frac{f'(X)}{f(X)} \right] \\ &= -h \left(\int_0^\infty (2F(x) - 1) f'(x) dx - \int_{-\infty}^0 (2F(x) - 1) f'(x) dx \right) \\ &= 2h \int_{-\infty}^\infty f^2(x) dx \ , \end{aligned}$$

where the last equality follows from integration by parts and $G(|X|) = 2F(X) - 1$ (by symmetry). Thus, under the alternative P_{θ_n},

$$W_n \xrightarrow{d} N\left(2h \int f^2, 1/3\right) \ .$$

Bibliography

LEHMANN, E. AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Lecture 9

Convolution Theorems[1]

Consider the following generic version of an estimation problem. One observes data $X_i, i = 1, \ldots, n$ i.i.d. with distribution $P \in \mathbf{P} = \{P_\theta : \theta \in \Theta\}$. Suppose we wish to estimate $\psi(\theta)$ using the data and that we have an estimator $T_n = T_n(X_1, \ldots, X_n)$ such that for each $\theta \in \Theta$,

$$\sqrt{n}\,(T_n - \psi(\theta)) \xrightarrow{d} L_\theta$$

under P_θ - for short we may write "under θ" today. What is the "best" possible limit distribution for such an estimator?

It is natural to measure "best" in terms of concentration, and we can measure concentration with a loss function. A loss function $\ell(x)$ is simply any function that takes values in $[0, \infty)$. A loss function is said to be "bowl-shaped" if the sublevel sets $\{x : \ell(x) \leq c\}$ are convex and symmetric about the origin. A common bowl-shaped loss function on $\mathbf{R}$ is mean-squared error loss, that is, $\ell(x) = x^2$. For a given loss function $\ell(x)$, a limit distribution will be considered "good" if

$$\int \ell(x) dL_\theta$$

is small.

If the estimator T_n is asymptotically normal in the sense that

$$L_\theta = N(\mu(\theta), \sigma^2(\theta)) \ ,$$

then in order to minimize the mean-squared error loss it is optimal to have $\mu(\theta) = 0$ and $\sigma^2(\theta)$ as small as possible. Of course, for estimators that are not asymptotically normal, this may not be true, and we do not wish to restrict attention a priori to asymptotically normal estimators.

[1]Today's notes are based on Azeem Shaikh's notes. I want to thank him for kindly sharing them.

9.1 Hodges' Estimator and Superefficiency

Suppose $\mathbf{P} = \{P_\theta = N(\theta, 1) : \theta \in \mathbf{R}\}$ and $\psi(\theta) = \theta$. A natural estimator of θ is the sample mean, that is, $T_n = \bar{X}_n$. As you already know, this estimator has many finite-sample optimality properties (it's minimax for every bowlshaped loss function, it's minimum variance unbiased, etc.), so we might reasonably expect it to be optimal asymptotically as well.

A second estimator of θ, S_n, can be defined as follows:

$$S_n = \begin{cases} T_n & \text{if } |T_n| \geq n^{-1/4} \\ 0 & \text{if } |T_n| < n^{-1/4} \end{cases} .$$

In words, $S_n = T_n$ when T_n is "far" from zero and $S_n = 0$ when T_n is "close" to zero. It is easy to see that

$$\sqrt{n}\,(T_n - \theta) \sim N(0, 1) .$$

But how does S_n behave asymptotically? To answer this question, first consider $\theta \neq 0$. For any such θ,

$$P_\theta \left\{ |T_n| \geq n^{-1/4} \right\} \to 1 .$$

To see this, let $Z_n = \sqrt{n}(T_n - \theta)$ and note that

$$\begin{aligned} P_\theta \left\{ |T_n| < n^{-1/4} \right\} &= P_\theta \left\{ -n^{-1/4} < T_n < n^{-1/4} \right\} \\ &= P_\theta \left\{ \sqrt{n}(-n^{-1/4} - \theta) < Z_n < \sqrt{n}(n^{-1/4} - \theta) \right\} . \end{aligned}$$

For $\theta > 0$, $n^{-1/4} - \theta < 0$ for n sufficiently large, so the probability tends to 0. For $\theta < 0$, $-n^{-1/4} - \theta > 0$ for n sufficiently large, so the probability tends to 0. The desired result thus follows. From the definition of S_n, we have that $S_n = T_n$ with probability approaching 1 for $\theta \neq 0$.

Now consider $\theta = 0$. In this case,

$$P_\theta \left\{ |T_n| \geq n^{-1/4} \right\} \to 0 .$$

To see this note that

$$\begin{aligned} P_\theta \left\{ |T_n| \geq n^{-1/4} \right\} &= P_\theta \left\{ T_n \geq n^{-1/4} \cup T_n \leq -n^{-1/4} \right\} \\ &= P_\theta \left\{ Z_n \geq n^{1/4} \cup Z_n \leq -n^{1/4} \right\} \\ &\leq P_\theta \left\{ Z_n \geq n^{1/4} \right\} + P_\theta \left\{ Z_n \leq -n^{1/4} \right\} . \end{aligned}$$

Both of the probabilities in the last expression tend to 0, so the result follows. From the definition of S_n, we have that $S_n = 0$ with probability approaching 1 for $\theta = 0$.

Thus, for $\theta \neq 0$

$$\sqrt{n}(S_n - \theta) \xrightarrow{d} N(0,1)$$

under P_θ and for $\theta = 0$

$$a_n(S_n - \theta) \xrightarrow{d} 0$$

under any sequence a_n, including $\sqrt{n}$. The estimator is said to be superefficient at $\theta = 0$.

Let L_θ denote the limit distribution of T_n and L'_θ denote the limit distribution of S_n. It follows from the above discussion that for $\theta \neq 0$

$$\int x^2 dL_\theta = \int x^2 dL'_\theta$$

and for $\theta = 0$,

$$\int x^2 dL'_\theta = 0 < 1 = \int x^2 dL_\theta \ .$$

Thus, S_n appears, at least in terms of its limiting distribution, to be a better estimator of θ than T_n. But appearances can be deceiving. This reasoning again reflects the poor use of asymptotics. Our hope is that

$$\int x^2 dL'_\theta$$

is a reasonable approximation to the finite-sample expected loss

$$E_\theta \left[\left(\sqrt{n} \left(S_n - \theta \right) \right)^2 \right] \ .$$

In finite-samples, for θ "far" from zero, we might expect $S_n = T_n$, and so we might expect L'_θ to be a reasonable approximation to the distribution of $\sqrt{n}\left(S_n - \theta\right)$; for "close" to zero, on the other hand, S_n will frequently differ from T_n, so the distribution of $\sqrt{n}\left(S_n - \theta\right)$ may be quite different from L'_θ. As before, the definition of "close" and "far" will differ with the sample size n. We must therefore consider the behavior of S_n under sequences $\theta_n \to 0$.

To illustrate this point, consider $\theta_n = \frac{h}{n^{1/4}}$ where $0 < h < 1$. (Implicitly, we are redefining $T_n = \bar{X}_{n,n}$, where $X_{i,n}, i = 1, \ldots, n$ are i.i.d. with distribution $P_{\theta_n} = N(\theta_n, 1)$). As before,

$$\sqrt{n}\left(T_n - \theta_n\right) \sim N(0,1) \ ,$$

but how does S_n behave under θ_n? To answer this, note that

$$\begin{aligned} P_{\theta_n}\left\{ |T_n| < n^{-1/4} \right\} &= P_{\theta_n}\left\{ -n^{-1/4} < T_n < n^{-1/4} \right\} \\ &= P_{\theta_n}\left\{ \sqrt{n}(-n^{-1/4} - \theta_n) < Z_n < \sqrt{n}(n^{-1/4} - \theta_n) \right\} \\ &= P_{\theta_n}\left\{ -n^{1/4}(1+h) < Z_n < n^{1/4}(1-h) \right\} \ . \end{aligned}$$

We saw earlier that this probability tended to 0 under $\theta \neq 0$, but under $\theta_n = \frac{h}{n^{1/4}}$, this probability tends to 1. Thus, under θ_n, we have that $S_n = 0$ with probability approaching 1. Hence, under θ_n

$$\sqrt{n}(S_n - \theta_n) = -n^{1/4}h$$

with probability approaching 1, and $-n^{1/4}h \to -\infty$. Denote by L the limiting distribution of T_n under θ_n and by L' the limiting distribution of S_n under θ_n (in this case L' is degenerate at $-\infty$). It follows that

$$\int x^2 dL' = \infty > 1 = \int x^2 dL \ .$$

Thus, S_n "buys" its better asymptotic performance at 0 at the expense of worse behavior for points "close" to zero. The definition of "close" changes with n, so this feature is not borne out by a pointwise asymptotic comparison for every $\theta \in \Theta$, but we can see it if we consider a sequence θ_n. We can also see it graphically by plotting the finite-sample expected losses, $E_\theta[\ell(\sqrt{n}(S_n - \theta))]$ versus $E_\theta[\ell(\sqrt{n}(T_n - \theta))] = 1$, for different samples sizes n.

This example is quite famous and is due to Hodges. The estimator S_n is often referred to as Hodges' estimator.

9.2 Efficiency of Maximum likelihood

Theorems that in some way show that a normal distribution with mean zero and covariance matrix equal to the inverse of the Fisher information is a "best possible" limit distribution have a long history, starting with Fisher in the 1920s and with important contributions by Cramér, Rao, Stein, Rubin, Chernoff and others. Of course, "the" theorem referred to is not true, at least not without a number of qualifications. The above example illustrates this and shows that it is impossible to give a nontrivial definition of "best" to the limit distributions L_θ. In fact, it is not even enough to consider L_θ under every $\theta \in \Theta$. For some fixed $\theta' \in \Theta$, we could always construct an estimator whose limit distribution was equal to L_θ for $\theta \neq \theta'$, but "better" at $\theta = \theta'$ by using the trick due to Hodges.

Le Cam contributed in various ways to an understanding of this issue, and eventually gave a complete explanation. Hájek formulated and proved two theorems, using different types of qualifications, which are now considered as most appropriate. It turns out that under certain conditions, the "best" limit distributions are in fact the limit distributions of maximum likelihood estimators, but to make this idea precise is a bit tricky. We need a few definitions first.

Definition 9.1 *T_n is called a sequence of locally regular estimators of $\psi(\theta)$ at the point θ_0 if, for every h*

$$a_n(T_n - \psi(\theta_0 + h/a_n)) \xrightarrow{d} L_{\theta_0} \text{ under } P_{\theta_0 + h/a_n}$$

as $a_n \to \infty$ (typically, $a_n = \sqrt{n}$), where the limit distribution might depend on θ_0 but not on h.

Note that a regular estimator sequence attains its limit distribution in a "locally uniform" manner. Intuitively, a small change in the parameter should not change the distribution of the estimator too much; a disappearing small change should not change the (limit) distribution at all.

Recall also that a model $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ is called differentiable in quadratic mean at θ if there exists a measurable function $\dot{\ell}_\theta$ such that, as $h \to 0$,

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h' \dot{\ell}_\theta \sqrt{p_\theta} \right]^2 d\mu = o(||h||^2),$$

where p_θ is the density of P_θ w.r.t. some measure μ.

9.2.1 Convolution Theorems

Hájek's convolution theorem shows that the limiting distribution of any *regular* estimator T_n can be written as a convolution of $N(0, \cdot)$ and "noise".

Theorem 9.1 (Hájek Convolution Theorem) *Suppose that $\mathbf{P}$ is differentiable in quadratic mean at each θ with non-singular Fisher information matrix $I_\theta = E_\theta[\dot{\ell}_\theta \dot{\ell}_\theta']$, and that ψ is differentiable at every θ. Let T_n be an at θ regular estimator sequence with limit distribution L_θ. Then, there exist distributions M_θ such that*

$$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta') * M_\theta \ .$$

In particular, if L_θ has covariance matrix Σ_θ, then the matrix $\Sigma_\theta - \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta'$ is nonnegative-definite.

The notation $*$ denotes the "convolution" operation between two distributions and should be interpreted as follows: If $X \sim F$ and $Y \sim G$ and $X \perp Y$, then $X+Y \sim F*G$. Theorem 9.1 is referred to as the Hájek convolution theorem. This theorem does not contradict the results of the previous section since it is easy to show that Hodges' estimator is not regular.

So, Hájek's convolution theorem puts a regularity restriction on the estimator sequence. Le Cam realized that the corresponding regularity restriction on estimators T_n in the limit experiment is location equivariance and that estimators T_n that are "equivariant-in-law" are rare. Another way to save the Cramér-Rao bound is to note that asymptotic superefficiency can occur only on very small sets of parameters, for instance null sets for the Lebesgue measure. Le Cam proved this for the first time in 1953, in his thesis. The following is a much nicer result, discovered by Le Cam later on.

Theorem 9.2 (Almost Everywhere Convolution Theorem) *Suppose that* $\mathbf{P}$ *is differentiable in quadratic mean at each* θ *with norming rate* a_n *and non-singular Fisher information matrix* $I_\theta = E_\theta[\dot{\ell}_\theta \dot{\ell}'_\theta]$*, and that* ψ *is differentiable at every* θ*. Let* T_n *be any estimator such that for every* θ

$$a_n(T_n - \psi(\theta)) \xrightarrow{d} L_\theta$$

under θ*. Then, there exist distributions* M_θ *such that for almost every* θ *w.r.t. Lebesgue measure*

$$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}'_\theta) * M_\theta \ .$$

This remarkable theorem yields the assertion of Hájek's convolution theorem at almost every parameter value θ, without having to impose the regularity requirement on the estimator sequence. This is however not really surprising, since Le Cam showed that it is roughly true that any estimator sequence T_n is "almost Hájek regular" at almost every parameter θ, at least along a subsequence of $\{n\}$ (see van der Vaart, 1998, Lemma 8.10). The convolution property implies that the covariance matrix of L_θ, if it exists, must be bounded below by the inverse Fisher information. This theorem does not contradict the results of the previous section. In that case, $\mathbf{P} = \{N(\theta, 1) : \theta \in \mathbf{R}\}$, $\psi(\theta) = \theta$, and $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}'_\theta) = N(0, 1)$. For every $\theta \neq 0$, $\sqrt{n}(S_n - \theta) \xrightarrow{d} N(0, 1)$ under θ, so the theorem is satisfied for M_θ the distribution with unit mass at 0.

Note that $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}'_\theta)$ is the limit distribution of the maximum likelihood estimator of $\psi(\theta)$. In order to assert that this is in fact the "best" limit distribution for more general loss functions, we need the following lemma.

Lemma 9.1 (Anderson's Lemma) *For any bowl-shaped loss function* ℓ *on* $\mathbf{R}^k$*, every probability distribution* M *on* $\mathbf{R}^k$*, and every covariance matrix* Σ*,*

$$\int \ell(x) dN(0, \Sigma) \leq \int \ell(x) d(N(0, \Sigma) * M) \ .$$

Thus, if "best" is measured by any bowl-shaped loss function (including mean-squared error loss), then, under the assumptions of Theorem 9.2, maximum likelihood estimators are "best" for almost every θ w.r.t. Lebesgue measure. For a proof of these two results, see van der Vaart (1998).

The almost-everywhere convolution theorem imposes no serious restrictions but yields no information about some parameters, albeit a null set of parameters. However, the lesson is that the possibility of improvement over the $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}'_\theta)$-limit is restricted on a null set of parameters. Improvement is also possible by considering special loss function (the James-Stein's estimator is an example), but improvement for one loss function necessarily implies worse performance for other loss functions (see van der Vaart, 1998).

An important part of these convolution theorems is the assumption that the model is QMD. The next example illustrates a situation in which maximum likelihood is not necessarily best. In addition, the differentiability of ψ is also key, see Hirano and Porter (2012) and Santos and Fang (2014) for recent contributions on non-differentiable functionals.

Example 9.1 *Suppose* $\mathbf{P} = \{P_\theta = U(0,\theta) : \theta > 0\}$ *and* $\psi(\theta) = \theta$ *(Recall that* $\mathbf{P}$ *is nowhere QMD so the model does not satisfy the conditions of the previous Theorems). We know that the MLE of* θ *is* $X_{(n)} = \max\{X_1, \ldots, X_n\}$ *and that*

$$n(\theta - X_{(n)}) \xrightarrow{d} L_\theta, \quad \text{where } L_\theta \text{ has density } \frac{1}{\theta}\exp\{-w/\theta\} \ . \tag{9.1}$$

Clearly, the estimator is not asymptotically normal. Although it converges at rate n*, much faster than the usual* $\sqrt{n}$ *rate, the fact that the limiting distribution lies completely to one side of the true parameter suggests that even better estimators may exists.*

To see that this is the case, note that for $W \sim L_\theta$*,* $E(W) = \theta$ *and* $Med(W) = \log(2)\theta$*. It is easy to see that for* $\ell(x) = x^2$*, MLE is sub-optimal and dominated by* $\tilde{\theta} = X_{(n)} + X_{(n)}/n$*. Note that*

$$\begin{aligned} n(\theta - \tilde{\theta}) &= n(\theta - X_{(n)} - X_{(n)}/n) \\ &= n(\theta - X_{(n)}) - X_{(n)} \\ &\xrightarrow{d} L'_\theta = L_\theta - \theta \ . \end{aligned}$$

It then follows that

$$\begin{aligned} \int x^2 dL'_\theta &= E(W-\theta)^2 = E(W^2 + \theta^2 - 2\theta W) \\ &= E(W^2) + \theta^2 - 2\theta^2 \\ &= E(W^2) - \theta^2 < E(W^2) = \int x^2 dL_\theta \end{aligned}$$

so that $E_{L'_\theta}(x)^2 < E_{L_\theta}(x)^2$*. On the other hand, for* $\ell(x) = |x|$*, MLE is sub-optimal and dominated by* $\tilde{\theta} = X_{(n)} + \log(2)X_{(n)}/n$*. These* $\tilde{\theta}$ *estimators turn out to be Bayes estimators.*

Proof of (9.1). Recall that $W \sim \exp(1)$ if

$$P\{W \le w\} = \begin{cases} 1 - \exp(-w) & \text{if } w \ge 0 \\ 0 & \text{otherwise} \end{cases} ,$$

which implies that

$$P\{-W\theta \le w\} = P\{W \ge -w/\theta\} = \begin{cases} \exp(w/\theta) & \text{if } w \ge 0 \\ 1 & \text{otherwise} \end{cases} .$$

Next note that,

$$\begin{aligned} P\{n(X_{(n)} - \theta) \leq x\} =& P\{X_{(n)} \leq \theta + \frac{x}{n}\} \\ =& P\{X_i \leq \theta + \frac{x}{n}\}^n. \end{aligned}$$

And then,

$$P\{X_i \leq \theta + \frac{x}{n}\} = \begin{cases} 0 & \text{if } x \leq -n\theta \\ \frac{1}{\theta}(\theta + \frac{x}{n}) & \text{if } -n\theta < x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}.$$

Therefore,

$$P\{X_{(n)} \leq \theta + \frac{x}{n}\} = \begin{cases} 0 & \text{if } x \leq -n\theta \\ (\frac{1}{\theta}(\theta + \frac{x}{n}))^n & \text{if } -n\theta < x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}.$$

It follows that,

$$(\frac{1}{\theta}(\theta + \frac{x}{n}))^n = (1 + \frac{x}{n\theta})^n \to \exp(x/\theta)$$

because of the identity

$$\exp(x) = \lim_{n\to\infty} (1 + \frac{x}{n})^n.$$

Hence,

$$P\{n(X_{(n)} - \theta) \leq x\} \to \begin{cases} \exp(x/\theta) & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}.$$

■

Bibliography

Hirano, K. and J. R. Porter (2012): "Impossibility results for nondifferentiable functionals," *Econometrica*, 80, 1769–1790.

Santos, A. and Z. Fang (2014): "Inference on Directionally Differentiable Functions," ArXiv:1404.3763.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Part III

Uniformly Valid Inference

Lecture 10

Uniformity

Let $X_i, i = 1, \ldots, n$ be an i.i.d. sample from some distribution $P \in \mathbf{P}$. Suppose one wishes to test the null hypothesis $H_0 : P \in \mathbf{P}_0 \subseteq \mathbf{P}$. To this end, one may consider a test function $\phi_n = \phi_n(X_1, \ldots, X_n)$ (that maps data into a binary decision) such that it controls the probability of a Type I error in some sense. It is clear that we must distinguish the exact size of a test from its approximate or asymptotic size. Ideally, we would like the test to satisfy

$$E_P[\phi_n] \leq \alpha \text{ for all } P \in \mathbf{P}_0 \text{ and } n \geq 1 \ , \tag{10.1}$$

but many times this is too demanding of a requirement. As a result, we may settle instead for tests such that

$$\limsup_{n \to \infty} E_P[\phi_n] \leq \alpha \text{ for all } P \in \mathbf{P}_0 \ . \tag{10.2}$$

Test satisfying (10.1) are said to be of level α for $P \in \mathbf{P}_0$, whereas tests satisfying (10.2) are said to be *pointwise asymptotically of level* α for $P \in \mathbf{P}_0$. The hope is that if (10.2) holds, then (10.1) holds approximately, at least for large enough n. However, asymptotic constructions that merely assert that the rejection probability of a test tends to the nominal level α under any *fixed* distribution P in the null hypothesis guarantee nothing about the exact finite sample size of a test. All that (10.2) ensures is that for each $P \in \mathbf{P}_0$ and $\epsilon > 0$ there is an $N(P)$ such that for all $n > N(P)$

$$E_P[\phi_n] \leq \alpha + \epsilon \ .$$

Importantly, the sample size required for the approximation to work, $N(P)$, may depend on P. As a result, it could be the case that for every sample size n (even, e.g., for $n = 10^{10}$) there could be $P = P_n \in \mathbf{P}_0$ such that

$$E_P[\phi_n] \gg \alpha \ .$$

Consider the following concrete example of this phenomenon.

Example 10.1 *Suppose* $\mathbf{P} = \{P \text{ on } \mathbf{R} : 0 < \sigma^2(P) < \infty\}$ *and* $\mathbf{P}_0 = \{P \in \mathbf{P} : \mu(P) = 0\}$. *Let* ϕ_n *be the t-test; that is,* $\phi_n = I\{\sqrt{n}\bar{X}_n > \hat{\sigma}_n z_{1-\alpha}\}$, *where* $z_{1-\alpha}$ *is the* $1-\alpha$ *quantile of the standard normal distribution. We know that*

$$E_P[\phi_n] \to \alpha \text{ for all } P \in \mathbf{P}_0 \ ,$$

but it turns out that the t-test suffers from the problem described above. In fact, we can show that for every $0 < c < 1$ *and every sample size* n *there exists a* $P_{n,c} \in \mathbf{P}_0$ *such that*

$$E_{P_{n,c}}[\phi_n] \geq c \ .$$

To see this, let n *and* c *be given. Let* $P_{n,c}$ *be the distribution that puts mass* $1-p_n$ *at* $p_n > 0$ *and mass* p_n *at* $-(1-p_n)$. *We will specify* p_n *in a minute, but first note that for such a distribution* $P_{n,c}$ *all of the* X_i *are in fact equal to* $p_n > 0$ *with probability* $(1-p_n)^n$. *For such a sequence of observations,* $\hat{\sigma}_n = 0$ *and* $\sqrt{n}\bar{X}_n > 0$, *so* $\phi_n = 1$. *The probability of rejection,* $E_{P_{n,c}}[\phi_n]$, *is therefore at least* $(1-p_n)^n$. *Now all that remains is to choose* p_n *so that* $(1-p_n)^n = c$; *that is,* $p_n = 1 - c^{1/n}$.

To rule this very disturbing possibility out, we need to ensure that the convergence in (10.2) is uniform for $P \in \mathbf{P}_0$.

Definition 10.1 *The sequence* $\{\phi_n\}$ *is uniformly asymptotically level* α *if*

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_P[\phi_n] \leq \alpha \ . \tag{10.3}$$

If instead of (10.3), the sequence $\{\phi_n\}$ *satisfies*

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_P[\phi_n] = \alpha \ , \tag{10.4}$$

then this value of α *is called the asymptotic size of* $\{\phi_n\}$.

The requirement in (10.3) implies that for each $\epsilon > 0$ there is an N (which does not depend on P) such that for all $n > N$

$$E_P[\phi_n] \leq \alpha + \epsilon \ .$$

In the case of the t-test, the above example shows us that this is not true for $\mathbf{P} = \{P \text{ on } \mathbf{R} : 0 < \sigma^2(P) < \infty\}$ and $\mathbf{P}_0 = \{P \in \mathbf{P} : \mu(P) = 0\}$. We will also study the behavior of tests under the alternative hypothesis and for that we will use the following definition.

Definition 10.2 *The sequence* $\{\phi_n\}$ *is pointwise consistent in power if, for an* $P \in \mathbf{P}_1$,

$$E_P[\phi_n] \to 1 \tag{10.5}$$

as $n \to \infty$.

10.1 A result of Bahadur and Savage (1956)

We know then that the t-test, $\phi_n = I\{\sqrt{n}\bar{X}_n > \hat{\sigma}_n z_{1-\alpha}\}$, has size one if when $\mathbf{P} = \{P \text{ on } \mathbf{R} : 0 < \sigma^2(P) < \infty\}$ and $\mathbf{P}_0 = \{P \in \mathbf{P} : \mu(P) = 0\}$:

$$\sup_{P \in \mathbf{P}_0} E_P[\phi_n] = 1 \ .$$

This result was perhaps a bit shocking, but it is possible that it is unique to the t-test - perhaps there are other tests of the same null hypothesis that would behave more reasonably. Unfortunately, we can show that this is not the case, provided that $\mathbf{P}$ is "sufficiently rich". Formally, we have the following result.

Theorem 10.1 *Let $\mathbf{P}$ be a class of distributions on $\mathbf{R}$ such that*

(i) For every $P \in \mathbf{P}$, $\mu(P)$ exists and is finite;

(ii) For every $m \in \mathbf{R}$, there is $P \in \mathbf{P}$ such that $\mu(P) = m$;

(iii) $\mathbf{P}$ is convex in the sense that if $P_1, P_2 \in \mathbf{P}$, then $\gamma P_1 + (1-\gamma)P_2 \in \mathbf{P}$.

Let $X_i, i = 1, \ldots, n$ be i.i.d. with distribution $P \in \mathbf{P}$. Let ϕ_n be any test sequence of the null hypothesis $H_0 : \mu(P) = 0$. Then,

(a) Any test of H_0 which has size α for $\mathbf{P}$ has power $\leq \alpha$ for any alternative $P \in \mathbf{P}$.

(b) Any test of H_0 which has power β against some alternative $P \in \mathbf{P}$ has size $\geq \beta$.

The proof of this result will follow from the following lemma.

Lemma 10.1 *Let $X_i, i = 1, \ldots, n$ be i.i.d. with distribution $P \in \mathbf{P}$, where $\mathbf{P}$ is the class of distributions on $\mathbf{R}$ satisfying (i)-(iii) in Theorem 10.1. Let ϕ_n be any test function. Define*

$$\mathbf{P}_m = \{P \in \mathbf{P} : \mu(P) = m\} \ .$$

Then,

$$\inf_{P \in \mathbf{P}_m} E_P[\phi_n] \text{ and } \sup_{P \in \mathbf{P}_m} E_P[\phi_n]$$

are independent of m.

PROOF. We show first that $\sup_{P \in \mathbf{P}_m} E_P[\phi_n]$ does not depend on m. Let m be given and choose $m' \neq m$. We wish to show that

$$\sup_{P \in \mathbf{P}_{m'}} E_P[\phi_n] = \sup_{P \in \mathbf{P}_m} E_P[\phi_n]$$

To this end, choose $P_j \in \mathbf{P}_m, j \geq 1$ so that

$$\lim_{j\to\infty} E_{P_j}[\phi_n] = \sup_{P\in\mathbf{P}_m} E_P[\phi_n] \ .$$

Let $h_j = j(m' - m) + m$ so that

$$m' = (1 - \frac{1}{j})m + \frac{1}{j}h_j \ .$$

Choose H_j so that $\mu(H_j) = h_j$ and define,

$$G_j = (1 - \frac{1}{j})P_j + \frac{1}{j}H_j \ .$$

Thus, $G_j \in \mathbf{P}_{m'}$. An observation from G_j can be obtained through a two-stage procedure. First, a coin is flipped with probability of heads $1/j$. If the outcome is a head, then the observation has the distribution H_j; otherwise, the observation is from P_j. Thus, with probability $(1 - \frac{1}{j})^n$, a sample of size n from G_j is simply a sample of size n from P_j. Therefore,

$$\sup_{P\in\mathbf{P}_{m'}} E_P[\phi_n] \geq E_{G_j}[\phi_n] \geq (1 - \frac{1}{j})^n E_{P_j}[\phi_n] \ .$$

But $(1 - \frac{1}{j})^n \to 1$ and $E_{P_j}[\phi_n] \to \sup_{P\in\mathbf{P}_m} E_P[\phi_n]$ as $j \to \infty$. Therefore,

$$\sup_{P\in\mathbf{P}_{m'}} E_P[\phi_n] \geq \sup_{P\in\mathbf{P}_m} E_P[\phi_n] \ .$$

Interchanging the roles of m and m', we can establish the reverse inequatlity

$$\sup_{P\in\mathbf{P}_{m'}} E_P[\phi_n] \leq \sup_{P\in\mathbf{P}_m} E_P[\phi_n] \ .$$

We could replace ϕ_n with $1 - \phi_n$ to establish that $\inf_{P\in\mathbf{P}_m} E_P[\phi_n]$ does not depend on m. ■

Proof of Theorem 10.1. To prove (a) let ϕ_n be a test of size α for $\mathbf{P}$. Let P' be any alternative. Define $m' = \mu(P')$. Then,

$$E_{P'}[\phi_n] \leq \sup_{P\in\mathbf{P}_{m'}} E_P[\phi_n] = \sup_{P\in\mathbf{P}_0} E_P[\phi_n] = \alpha \ .$$

The proof of (b) is similar. ■

The Bahadur-Savage result holds in the multivariate case as well. The theorem reads exactly, except that $\mathbf{P}$ refers to a family of distributions on $\mathbf{R}^k$ satisfying (i)-(iii) above with m a vector.

The class of distributions with finite second moment satisfies the requirements of the theorem, as does the class of distributions with infinitely many

moments. Thus, the failure of the t-test is not special to the t-test; in this setting, there simply exist no "reasonable" tests. Evidently, the problem is due to the fact that the mean $\mu(P)$ is quite sensitive for the tails of P, and one sample yields little information about the tails. But this does not mean that all hope is lost. Fortunately, the t-test does satisfy (10.3) for certain large classes of distributions that are somewhat smaller than $\mathbf{P}$ in Theorem 10.1. We will discuss this next class.

10.2 Extension of the Result by Bahadur-Savage

In this section we generalize the result of Bahadur and Savage following Romano (2004), by providing a constructive sufficient condition that applies to other testing problems as well. Although the idea is similar to theirs, it allows one to answer a conjecture of Bahadur and Savage concerning testing the existence of a mean. In addition, the idea of the Theorem was key in proving results about the testability of completeness conditions in non-parametric models with endogeneity, see Canay et al. (2013).

Suppose data X are observed on a sample space S with probability law P. A model is assumed only in the sense that P is known to belong to $\mathbf{P}$, some family of distributions on S. Consider the problem of testing the null hypothesis $H_0 : P \in \mathbf{P}_0$ versus the alternative hypothesis $H_1 : P \in \mathbf{P}_1 = \mathbf{P} \setminus \mathbf{P}_0$.

A convenient way to discuss the non-existence of tests with good power properties is in terms of the total variation metric, defined by

$$\tau(P,Q) \equiv \sup_{\{g:|g|\leq 1\}} \left| \int g dQ - \int g dP \right|. \tag{10.6}$$

Consider the following condition.

Condition 10.1 *For every $Q \in \mathbf{P}_1$ there exists a sequence $P_k \in \mathbf{P}_0$ such that $\tau(Q, P_k) \to 0$ as $k \to \infty$.*

Evidently, condition 10.1 asserts that $\mathbf{P}_0$ is dense in $\mathbf{P}$ with respect to the metric τ. We will also assume $\mathbf{P}_0$ and $\mathbf{P}_1$ satisfy the following (stronger) condition.

Condition 10.2 *For every $Q \in \mathbf{P}_1$ and any $\epsilon > 0$, there exists a subset $A = A_\epsilon$ of S satisfying $Q(A_\epsilon) \geq 1 - \epsilon$ and such that, if X has distribution Q, the conditional distribution of X given $X \in A_\epsilon$ is a distribution in $\mathbf{P}_0$.*

We can now prove, under conditions 10.1 or 10.2, that no test has power against Q greater than the size of the test.

Theorem 10.2 *Let $\phi_n(X)$ be any test of $\mathbf{P}_0$ versus $\mathbf{P}_1$.*

(i) If Condition 10.1 holds, then

$$\sup_{Q\in\mathbf{P}_1} E_Q[\phi_n(X)] \le \sup_{P\in\mathbf{P}_0} E_P[\phi_n(X)]. \tag{10.7}$$

Hence, if ϕ_n has size α, then

$$\sup_{Q\in\mathbf{P}_1} E_Q[\phi_n(X)] \le \alpha; \tag{10.8}$$

that is, the power function if bounded by α.

(ii) Assume Condition 10.2 holds. Then Condition 10.1 holds and therefore (10.7) *and* (10.8) *hold as well.*

PROOF. To prove (i), fix $Q \in \mathbf{P}_1$ and let g be a function with $|g| \le 1$. Take any $\epsilon_k \to 0$ and let $P_k \in \mathbf{P}_0$ satisfy $\tau(Q, P_k) \le \epsilon_k$. Then,

$$\begin{aligned} E_Q[g] &= \int g dQ - \int g dP_k + \int g dP_k \\ &\le E_{P_k}[g] + \tau(Q, P_k) \le \sup_{P\in\mathbf{P}_0} E_P[g] + \epsilon_k \ . \end{aligned}$$

Let $\epsilon_k \to 0$ and the result follows.

To prove (ii), let $\epsilon_k \to 0$, let A_{ϵ_k} be the subset in Condition 10.2, and let P_k denote the distribution of X given $X \in A_{\epsilon_k}$ when X has distribution Q. Then, for any g with $|g| \le 1$,

$$\begin{aligned} E_Q[g(X)] &= E_Q[g(X)|A_{\epsilon_k}]Q(A_{\epsilon_k}) + E_Q[g(X)|A^c_{\epsilon_k}]Q(A^c_{\epsilon_k}) \\ &\le E_{P_k}[g(X)]Q(A_{\epsilon_k}) + Q(A^c_{\epsilon_k}) \\ &\le E_{P_k}[g(X)] + \epsilon_k \ . \end{aligned}$$

The above assumed $E_{P_k}[g(X)] > 0$, but if $E_{P_k}[g(X)] < 0$ a similar argument gives $\le E_{P_k}[g(X)] + 2\epsilon_k$. In addition,

$$E_Q[g(X)] \ge E_{P_k}[g(X)](1 - \epsilon_k) - \epsilon_k \ge E_{P_k}[g(X)] - 2\epsilon_k. \tag{10.9}$$

Hence, $\tau(Q, P_k) \le 2\epsilon_k$ and the result follows by letting $\epsilon_k \to 0$. ■

Note that the hypothesis testing framework does not have to be cast in terms of testing a particular parameter as $\mathbf{P}_0$ and $\mathbf{P}_1$ are quite general. However, the main point is that condition 10.2, although stronger than condition 10.1, is easily verified is some novel examples.

When $X_1, \ldots, X_n$ is a vector of n i.i.d. random variables, then it suffices to verify condition 10.2 for $n = 1$. To produce the set A_ϵ, for X, simply take n-fold product set A_δ obtained from the case $n = 1$, where δ is taken small enough to guarantee with probability $1 - \epsilon$ that all n observations fall in A_δ. But, the chance that all observations fall in A_δ is at least $(1 - \delta)^n$. Thus, choose δ no bigger than $1 - (1 - \epsilon)^{1/n}$.

Example 10.2 (Finite versus not finite mean) *Let X be $X_1, \ldots, X_n$, n i.i.d. observations on the real line. As remarked by Bahadur and Savage (1956), "it would be interesting to know whether, in comparable nonparametric situations, tests of the existence of μ, are equally unsuccessful"; here, μ refers to the mean of an observation. So, let $\mathbf{P}_0$ be the family of distributions on the real line with a finite mean, and let $\mathbf{P}_1$ be the distributions without a finite mean. Condition 10.2 readily holds.*

To see why, suppose Q is a distribution without a mean. Given ϵ, let A be any bounded subset of the real line with probability at least $(1-\epsilon)$ under Q. Moreover, the conditional distribution of an observation given that it falls in A is some distribution on a bounded set, i.e. a distribution in the null hypothesis parameter space. Hence, the conclusion of Theorem 10.2 holds, and so it is impossible to construct a test with power greater than the size of the test.

For a real-valued parameter θ the impossibility of testing a hypothesis like $H_0 : \theta \neq \theta_0$ versus $H_1 : \theta = \theta_0$ is well known. More generally, it is impossible to test $H_0 : P \in \mathbf{P}_0$ versus $H_1 : P \in \mathbf{P}_1$ when $\mathbf{P}_0$ is dense in $\mathbf{P} = \mathbf{P}_1 \cup \mathbf{P}_0$. For testing goodness-of-fit, it is then impossible to conclude that the underlying distribution is normal, or any other family that falls in a lower dimensional subspace of the a priori model space. To make this precise, consider the following condition.

Condition 10.3 *For any $Q \in \mathbf{P}_1$ and any $\epsilon > 0$, there exists some distribution R such that $(1-\epsilon)Q + \epsilon R \in \mathbf{P}_0$.*

It is easy to see that Condition 10.3 implies Condition 10.1. To see this, pick $Q \in \mathbf{P}_1$ and let $\epsilon_k > 0$. Flip a coin with probability $1-\epsilon_k$ of heads, and let A_{ϵ_k} be the event "the toss is a head". Let $Y(\omega)$ be a random variable (on some probability space) that has distribution Q conditional on $\omega \in A_{\epsilon_k}$, and has distribution R condition on $\omega \in A^c_{\epsilon_k}$, for some distribution R.

It follows from Condition 10.3 that for any $Q \in \mathbf{P}_1$ and any $\epsilon_k > 0$, there exists a random variable Y (on some probability space) with distribution $P_k = (1-\epsilon_k)Q + \epsilon_k R \in \mathbf{P}_0$ and a subset $A = A_{\epsilon_k}$, with $P_k(A) \geq 1 - \epsilon_k$ such that the conditional distributionof Y given A, is Q. Then,

$$\begin{aligned} E_{P_k}[g(Y)] &= E_{P_k}[g(Y)|A_{\epsilon_k}]P_k(A_{\epsilon_k}) + E_{P_k}[g(Y)|A^c_{\epsilon_k}]P_k(A^c_{\epsilon_k}), \\ &\leq E_Q[g(Y)] + P_k(A^c_{\epsilon_k}), \\ &\leq E_Q[g(Y)] + \epsilon_k \ . \end{aligned}$$

Similarly,

$$E_{P_k}[g(Y)] \geq E_Q[g(Y)](1-\epsilon_k) - \epsilon_k \geq E_Q[g(Y)] - 2\epsilon_k \ . \tag{10.10}$$

Hence, $\tau(Q, P_k) \leq 2\epsilon_k$ and the result follows by letting $\epsilon_k \to 0$.

Example 10.3 (Goodness-of-fit testing) *The usual approach to testing goodness-of-fit runs as follows. Assume $X_1, \ldots, X_n$ are i.i.d. S-valued random variables with distribution P. The null hypothesis asserts P belongs to some class $\{P_\theta : \theta \in \Theta\}$ and the alternative hypothesis asserts $P \in \bar{\mathbf{P}}$, the family of all other distributions on S. Reversing the roles of the null and alternative is not possible. For example, consider the problem of testing uniformity on $S = (0, 1)$. We verify condition 10.3 to show the (fairly obvious) results that it is impossible to test the null hypothesis that P is not uniform on $(0, 1)$ versus the alternative that P is uniform on $(0, 1)$, at least not with any degree of power. In this example, $\mathbf{P}_1$ consists of U, the uniform distribution on $(0, 1)$. Then, for any other distribution R and any $\epsilon > 0$, $(1 - \epsilon)U + \epsilon R$ is not U, and so condition 10.3 holds.*

Similar considerations apply when $\mathbf{P}_1$ is a larger parametric model, such as the family of normal distributions. In summary, one can apply a goodness-of-fit test (such as Kolmogorov-Smirnov) to show that the data are consistent with the model, but one cannot definitely conclude that the model holds.

Bibliography

Bahadur, R. and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 25, 1115–1122.

Canay, I. A., A. Santos, and A. M. Shaikh (2013): "On the Testability of Identification in Some Nonparametric Models with Endogeneity," *Econometrica*, 81, 2535 – 2559.

Lehmann, E. and W.-Y. Loh (1990): "Pointwise versus uniform robustness in some large-sample tests and confidence intervals," *Scandinavian Journal of Statistics*, 17, 177–187.

Romano, J. P. (2004): "On Non-parametric Testing, the Uniform Behaviour of the t-test, and Related Problems," *Scandinavian Journal of Statistics*, 31, 567–584.

Lecture 11

Uniformity of the t-test

Last class we covered the important result by Bahadur and Savage, and some generalizations. We learned that the class of distributions that satisfy the conditions of their theorem is large. For example, the class of distributions with finite second moment satisfies the requirements of the theorem, as does the class of distributions with infinitely many moments. We concluded that the failure of the t-test is not special to the t-test; in this setting, there simply exist no "reasonable" tests. Evidently, the problem is due to the fact that the mean $\mu(P)$ is quite sensitive for the tails of P, and one sample yields little information about the tails. But this does not mean that all hope is lost. Fortunately, the t-test does satisfy (10.3) for certain large classes of distributions that are somewhat smaller than $\mathbf{P}$ in Theorem 10.1.

11.1 Distributions with Compact Support

The Bahadur-Savage result does not apply to the family of distributions supported on a compact set (condition (ii) is violated). However, if we now restrict our attention to distributions supported on a compact set, the size of the t-test is still 1, as the following calculation demonstrates.

Let $X_1, \ldots, X_n$ be i.i.d. P. Consider the one-sided α t-test with test function ϕ_n, for testing $\mu(P) = 0$ versus $\mu(P) > 0$. Let $\mathbf{P}$ be the set of distributions supported on $[-1, 1]$, and $\mathbf{P}_0$ those distributions on $[-1, 1]$ with mean 0. We will show,

$$\sup_{P \in \mathbf{P}_0} E_P[\phi_n] = 1, \quad \forall n \geq 2 \ .$$

It suffices to show that there exists a $P \in \mathbf{P}_0$ such that the probability of rejection under P is arbitrarily close to 1. To this end, we can use the same construction we used in Example 10.1. Fix $n > 1$ and any $c < 1$. Then, choose $p_n > 0$ so that $(1 - p_n)^n = c$. Let $P = P_{n,c}$ be the distribution that places mass $1 - p_n$ at p_n and mass p_n at $-(1 - p_n)$, so that $\mu(P) = 0$.

The idea is that, with probability at least c, a random sample of size n from P will be the sample having all observations equal to $p_n > 0$. For such a sequence of observations, $\hat{\sigma}_n = 0$ and $\sqrt{n}\bar{X}_n > 0$, so $\phi_n = 1$. The probability of rejection, $E_P[\phi_n]$, is therefore at least c, which can be arbitrarily close to 1. The problem here is that we have no control over the skewness in the class $P_{n,c}$. In fact, the skewness of the two-point distribution is

$$\frac{E_{P_{n,c}}[X^3]}{\sigma^3(P_{n,c})} = \frac{p_n^2 + (1-p_n)^2}{\sqrt{p_n(1-p_n)}} \to \infty \ .$$

In fact, it follows that for $\delta > 0$,

$$\frac{E_{P_{n,c}}[|X|^{2+\delta}]}{\sigma^{2+\delta}(P_{n,c})} = \frac{p_n^{1+\delta} + (1-p_n)^{1+\delta}}{(\sqrt{p_n(1-p_n)})^{\delta}} \to \infty \ ,$$

which is a condition that will be meaningful in the next section.

One can make the example more convincing to ensure that the underlying distribution is continuous and the observations are distinct. Let $X^*_{n,i} = X_{n,i} + U_{n,i}$, where $X_{n,i}$ has the distribution $P_{n,c}$ above and, independently, $U_{n,i}$ is uniform on $[-\tau_n, \tau_n]$. Assume τ_n is small enough such that,

$$\tau_n < \frac{\sqrt{n}p_n}{\sqrt{n} + z_{1-\alpha}} \ . \tag{11.1}$$

Also, notice that

$$\bar{X}^*_{n,n} = \bar{X}_{n,n} + \bar{U}_{n,n} \ ,$$

which with probability at least c satisfies,

$$\bar{X}^*_{n,n} = p_n + \bar{U}_{n,n} \geq p_n - \tau_n > 0 \ ,$$

since $\bar{U}_{n,n} \geq -\tau_n$ and, from (11.1),

$$p_n - \tau_n > \frac{\tau_n}{\sqrt{n}} z_{1-\alpha} > 0 \ .$$

Also, with probability at least c,

$$\hat{\sigma}^{2,*}_n = \frac{1}{n}\sum (X^*_{n,i} - \bar{X}^*_{n,n})^2 = \frac{1}{n}\sum (U_{n,i} - \bar{U}_{n,n})^2 \leq \tau_n^2 \ ,$$

and then it follows that

$$\begin{aligned} T_n &= \frac{\sqrt{n}\bar{X}^*_{n,n}}{\hat{\sigma}^*_n} \\ &\geq \frac{\sqrt{n}p_n - \sqrt{n}\tau_n}{\hat{\sigma}^*_n} \\ &> \frac{\tau_n z_{1-\alpha}}{\hat{\sigma}^*_n} \\ &\geq \frac{\tau_n z_{1-\alpha}}{\tau_n} = z_{1-\alpha} \ , \end{aligned}$$

by our choice of τ_n. Again, the probability of such event is c and this can be made arbitrarily close to 1. It is important to recognize however that by restricting attention to distributions on a compact set, it is possible to construct tests of size α (other than the t-test) that have reasonable power properties. See Lehmann and Romano (2005, ch. 11) for such a result.

As an alternative to the family of distributions supported on a compact set, it is interesting to study the behavior of the t-test under the assumption of symmetry (so again the Bahadur-Savage result does not hold since condition (iii) is violated). You will show in the problem set that the t-test is not uniformly asymptotically level α for the family of symmetric distributions. In fact, the size of the t-test under symmetry is one for moderate values of α (although it can be shown that the size is bounded away from 1 for small values of α). In general, the t-test does not behave uniformly well across distributions with large skewness, as the limiting normal theory fails.

11.2 Distributions with $2+\delta$ Moments

Fortunately, not everything is lost for the t-test. We will now show that the the t-test is uniformly consistent over certain large subfamilies of distributions with two finite moments. For this purpose, consider the family of distributions $\mathbf{P}$ on the real line satisfying,

$$\lim_{\lambda\to\infty}\sup_{P\in\mathbf{P}} E_P\left[\frac{|X-\mu(P)|^2}{\sigma^2(P)} I\left\{\frac{|X-\mu(P)|}{\sigma(P)}>\lambda\right\}\right]=0\ . \tag{11.2}$$

In particular, if we let $\mathbf{P}^{2+\delta}$ be the set of distributions satisfying

$$\mathbf{P}^{2+\delta}=\left\{P: E_P\left[\frac{|X-\mu(P)|^{2+\delta}}{\sigma^{2+\delta}(P)}\right]\le M\right\}\ , \tag{11.3}$$

for some $\delta>0$ and $M<\infty$, it follows that $\mathbf{P}^{2+\delta}\subseteq\mathbf{P}$. To see why, let

$$Y=\frac{X-\mu(P)}{\sigma(P)}$$

and note that from the inequality,

$$\lambda^\delta Y^2 I\{|Y|>\lambda\}\le |Y|^{2+\delta}\ ,$$

it follows that

$$\lim_{\lambda\to\infty}\sup_{P\in\mathbf{P}^{2+\delta}} E_P\left[|Y|^2 I\left\{|Y|>\lambda\right\}\right]\le\lim_{\lambda\to\infty}\frac{1}{\lambda^\delta}\sup_{P\in\mathbf{P}^{2+\delta}} E_P\left[|Y|^{2+\delta}\right]=0\ .$$

In addition, let $\mathbf{P}_0$ be the set of distributions in $\mathbf{P}$ with $\mu(P)=0$. For testing $\mu(P)=0$ versus $\mu(P)>0$, the t-test is defined as $\phi_n=I\{T_n>z_{1-\alpha}\}$, where

$$T_n=\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n}\ ,$$

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. The next Theorem shows that the t-test is uniformly asymptotically level α over $\mathbf{P}_0$.

Theorem 11.1 *Suppose $X_{n,1}, \ldots, X_{n,n}$ are i.i.d. with distribution $P_n \in \mathbf{P}$, where $\mathbf{P}$ satisfies* (11.2)*. Then, under P_n*

$$\frac{\sqrt{n}(\bar{X}_{n,n} - \mu(P_n))}{\hat{\sigma}_{n,n}} \xrightarrow{d} N(0,1) \ . \tag{11.4}$$

In addition, for testing $\mu(P) = 0$ versus $\mu(P) > 0$, the t-test is uniformly asymptotically level α over $\mathbf{P}_0$; that is,

$$\lim_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_P[\phi_n] = \alpha \ . \tag{11.5}$$

Proof of Theorem 11.1. First note that we can write,

$$\frac{n^{1/2}(\bar{X}_{n,n} - \mu(P_n))}{\hat{\sigma}_{n,n}} = \frac{n^{1/2}(\bar{X}_{n,n} - \mu(P_n))}{\sigma(P_n)} \times \frac{\sigma(P_n)}{\hat{\sigma}_{n,n}} \ .$$

We now want to apply the CLT for triangular arrays (Lindeberg-Feller) to the first term. To this end, let $Y_{n,i} = (X_{n,i} - \mu(P_n))/\sigma(P_n)$. We need to check the condition

$$\limsup_{n\to\infty} E_{P_n}[Y_{n,i}^2 I\{|Y_{n,i}| > \epsilon\sqrt{n}\}] = 0 \ ,$$

for every $\epsilon > 0$ (see Theorem 6.1). But, for every $\lambda > 0$,

$$\limsup_{n\to\infty} E_{P_n}[Y_{n,i}^2 I\{|Y_{n,i}| > \epsilon\sqrt{n}\}] \le \limsup_{n\to\infty} E_{P_n}[Y_{n,i}^2 I\{|Y_{n,i}| > \lambda\}] \ .$$

Let $\lambda \to \infty$ and the right side tends to zero. It follows that,

$$\frac{\sqrt{n}(\bar{X}_{n,n} - \mu(P_n))}{\sigma(P_n)} \xrightarrow{d} N(0,1) \ . \tag{11.6}$$

It remains to show that $\sigma(P_n)/\hat{\sigma}_{n,n} \to 1$ in probability. In order to do this, assume wlog that $\mu(P_n) = 0$ and note that

$$\frac{\hat{\sigma}_{n,n}^2}{\sigma^2(P_n)} = \frac{1}{n}\sum_{i=1}^n \frac{X_{n,i}^2}{\sigma^2(P_n)} - \frac{1}{n}\left(\frac{\sqrt{n}\bar{X}_{n,n}}{\sigma(P_n)}\right)^2 = \frac{1}{n}\sum_{i=1}^n \frac{X_{n,i}^2}{\sigma^2(P_n)} + o_{P_n}(1) \ ,$$

where the second line follows from (11.6). The proof is then completed by

$$\frac{1}{n}\sum_{i=1}^n \frac{X_{n,i}^2}{\sigma^2(P_n)} \to 1 \ ,$$

in probability under P_n, which in turn follows from Lemma 11.2 below.

To prove (11.5) note that, if the result failed, one could extract a subsequence $\{P_n\}$ with $P_n \in \mathbf{P}_0$ such that,

$$E_{P_n}[\phi_n] \to \alpha' \neq \alpha \ .$$

This contradicts (11.4) as T_n is asymptotically standard normal under P_n. ■

As we can see, to prove the theorem all we need is a CLT for triangular arrays (see Theorem 6.1) and a law of large numbers for triangular arrays. The latter is handled by the following two lemmas.

Lemma 11.1 *Let $Y_{n,1}, \ldots, Y_{n,n}$ be i.i.d. with cdf G_n and finite mean $\mu(G_n)$ satisfying*

$$\lim_{\beta\to\infty} \limsup_{n\to\infty} E_{G_n}[|Y_{n,i} - \mu(G_n)| I\{|Y_{n,i} - \mu(G_n)| \geq \beta\}] = 0 \ . \tag{11.7}$$

Let $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_{n,i}$. Then, under G_n, $\bar{Y}_n - \mu(G_n) \to 0$ in probability.

PROOF. Assume $\mu(G_n) = 0$ wlog and note that

$$0 = E[Y_{n,i}] = E[Y_{n,i} I\{|Y_{n,i}| \leq n\}] + E[Y_{n,i} I\{|Y_{n,i}| > n\}] \ . \tag{11.8}$$

We will study each of the expectations on the right hand side separately. In order to do this, define

$$Z_{n,i} = Y_{n,i} I\{|Y_{n,i}| \leq n\} \ .$$

Let $m_n = E[Z_{n,i}]$ and $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_{n,i}$. For any $\epsilon > 0$, we have

$$P\{|\bar{Y}_n - m_n| > \epsilon\} \leq P\{|\bar{Z}_n - m_n| > \epsilon\} + P\{\bar{Y}_n \neq \bar{Z}_n\} \ , \tag{11.9}$$

since the event $\{|\bar{Y}_n - m_n| > \epsilon\}$ implies either $\{|\bar{Z}_n - m_n| > \epsilon\}$ or $\{\bar{Y}_n \neq \bar{Z}_n\}$. In turn,

$$\begin{aligned} P\{\bar{Y}_n \neq \bar{Z}_n\} &\leq P\left\{ \bigcup_{1\leq i\leq n} \{Y_{n,i} \neq Z_{n,i}\} \right\} \\ &\leq \sum_{1\leq i\leq n} P\{Y_{n,i} \neq Z_{n,i}\} \\ &= \sum_{1\leq i\leq n} P\{|Y_{n,i}| > n\} \\ &= nP\{|Y_{n,i}| > n\} \ . \end{aligned}$$

By Chebychev's inequality,

$$P\{|\bar{Z}_n - m_n| > \epsilon\} \leq \frac{\mathrm{var}[\bar{Z}_n]}{\epsilon^2} = \frac{\mathrm{var}[Z_{n,i}]}{n\epsilon^2} \leq \frac{E[Z_{n,i}^2]}{n\epsilon^2} \ .$$

Hence,

$$P\{|\bar{Y}_n - m_n| > \epsilon\} \leq (n\epsilon^2)^{-1} E[Z_{n,i}^2] + nP\{|Y_{n,i}| > n\} \ .$$

For $t > 0$, let

$$\tau_n(t) = tP\{|Y_{n,i}| > t\} = t(1 - G_n(t) + G_n(-t))$$
$$\kappa_n(t) = \frac{1}{t} E[Y_{n,i}^2 I\{|Y_{n,i}| \leq t\}] = \frac{1}{t} \int_{-t}^{t} y^2 dG_n(y) \ .$$

In this notation,

$$P\{|\bar{Y}_n - m_n| > \epsilon\} \leq \epsilon^{-2} \kappa_n(n) + \tau_n(n) \ .$$

Since

$$tP\{|Y_{n,i}| > t\} \leq E[|Y_{n,i}| I\{|Y_{n,i}| > t\}] \ ,$$

it follows that $\tau_n(n) \to 0$ by (11.7). Now, using integration by parts it is possible to show that

$$\kappa_n(t) = -\tau_n(t) + \frac{2}{t} \int_0^t \tau_n(x) dx \ . \tag{11.10}$$

Therefore, in order to show that $P\{|\bar{Y}_n - m_n| > \epsilon\} \to 0$, it suffices to argue that

$$\frac{2}{n} \int_0^n \tau_n(x) dx \to 0 \ . \tag{11.11}$$

To this end, note that

$$\frac{2}{n} \int_0^n \tau_n(x) dx \leq \frac{2}{n} \int_0^n E[|Y_{n,i}| I\{|Y_{n,i}| > x\}] dx \ . \tag{11.12}$$

Let $\delta > 0$ be given and choose n_0 and β_0 so that

$$E[|Y_{n,i}| I\{|Y_{n,i}| > x\}] < \frac{\delta}{2}$$

whenever $n > n_0$ and $x > \beta_0$. For $x \leq \beta_0$ and $n > n_0$ we have that

$$\begin{aligned} E[|Y_{n,i}| I\{|Y_{n,i}| > x\}] &\leq E[|Y_{n,i}|] \\ &= E[|Y_{n,i}| I\{|Y_{n,i}| \leq \beta_0\}] + E[|Y_{n,i}| I\{|Y_{n,i}| > \beta_0\}] \\ &= \beta_0 + \frac{\delta}{2} \ . \end{aligned}$$

It follows that

$$\frac{2}{n} \int_0^n E[|Y_{n,i}| I\{|Y_{n,i}| > x\}] dx \leq \frac{\beta_0(2\beta_0 + \delta)}{n} + \delta \ , \tag{11.13}$$

which is less than δ for n sufficiently large. Since the choice of $\delta > 0$ was arbitrary, it follows that $\kappa_n(n) \to 0$. Therefore, $\bar{Y}_n - m_n \to 0$ in probability. To complete the proof note that

$$0 = E[Y_{n,i}] = m_n + E[Y_{n,i} I\{|Y_{n,i}| > n\}] \tag{11.14}$$

so that

$$|m_n| \leq E[|Y_{n,i}| I\{|Y_{n,i}| > n\}] \to 0 \ .$$

■

Lemma 11.2 *Let* $\mathbf{P}$ *be a family of distributions satisfying* (11.2)*. Suppose* $X_{n,1}, \ldots, X_{n,n}$ *are i.i.d.* $P_n \in \mathbf{P}$ *and* $\mu(P_n) = 0$*. Then, under* P_n*,*

$$\frac{1}{n} \sum_{i=1}^{n} \frac{X_{n,i}^2}{\sigma^2(P_n)} \to 1 \text{ in probability} \ .$$

PROOF. Apply Lemma 11.1 to

$$Y_{n,i} = [X_{n,i}^2 / \sigma^2(P_n)] - 1 \ .$$

To see that Lemma 11.1 applies, consider $\beta > 1$. In this case, the event $\{|Y_{n,i}| > \beta\}$ implies $X_{n,i}^2/\sigma^2(P_n) > \beta + 1$ since $|Y_{n,i}| > 1$ cannot happen if $X_{n,i}^2/\sigma^2(P_n) - 1 < 0$. In addition, note that

$$|Y_{n,i}| < X_{n,i}^2/\sigma^2(P_n) \ .$$

Hence, for $\beta > 1$,

$$E[|Y_{n,i}| I\{|Y_{n,i}| > \beta\}] \leq E\left[\frac{X_{n,i}^2}{\sigma^2(P_n)} I\left\{\frac{|X_{n,i}|}{\sigma(P_n)} > \sqrt{\beta+1}\right\}\right] \ .$$

Condition (11.2) therefore immediately implies (11.7). ■

11.2.1 Power of the *t*-test

So far we know that the t-test behaves uniformly well across a fairly large class of distributions. We will now study some power properties of the t-test. In particular, we will show that the t-test is uniformly consistent in level, and derive a limiting power calculation. The result is summarized in the following Theorem.

Theorem 11.2 *Let* $\mathbf{P}$ *be a family of distributions satisfying* (11.2) *and let* $\mathbf{P}_0$ *be the set of distributions in* $\mathbf{P}$ *with* $\mu(P) = 0$ *(assumed non-empty).*

Then, for testing $\mu(P) = 0$ versus $\mu(P) > 0$, the limiting power of the t-test against $P_n \in \mathbf{P}$ with $n^{1/2}\mu(P_n)/\sigma(P_n) \to \delta$ is given by

$$\lim_{n\to\infty} E_{P_n}[\phi_n] = 1 - \Phi(z_{1-\alpha} - \delta) \ . \tag{11.15}$$

Furthermore,

$$\lim_{n\to\infty} \inf_{\{P\in\mathbf{P}: n^{1/2}\mu(P)/\sigma(P)\geq\delta\}} E_P[\phi_n] = 1 - \Phi(z_{1-\alpha} - \delta) \ . \tag{11.16}$$

PROOF. Let $X_{n,1}, \ldots, X_{n,n}$ be i.i.d. with distribution P_n and consider the t-statistic $T_n = \bar{X}_{n,n}/\hat{\sigma}_{n,n}$. Write

$$T_n = \frac{n^{1/2}(\bar{X}_{n,n} - \mu(P_n))}{\hat{\sigma}_{n,n}} + \frac{n^{1/2}\mu(P_n)/\sigma(P_n)}{\hat{\sigma}_{n,n}/\sigma(P_n)} \ .$$

By Theorem 11.1 the first term converges weakly to $N(0,1)$ under P_n, and by the proof of the same Theorem, the denominator of the second term converges to 1 in probability under P_n. It follows that $T_n \stackrel{d}{\to} N(\delta, 1)$ under P_n and so (11.15) follows.

To prove (11.16), argue by contradiction and assume there exists a subsequence $\{P_n\}$ with $n^{1/2}\mu(P_n)/\sigma(P_n) \geq \delta$ such that

$$E_{P_n}[\phi_n] \to \gamma < 1 - \Phi(z_{1-\alpha} - \delta) \ .$$

This, however, would violate (11.15) if $n^{1/2}\mu(P_n)/\sigma(P_n)$ has a limit. If it does not have a limit, pass to any convergent subsequence and apply the same argument. ■

We will conclude today's lecture by noting that condition (11.16) fails if $\mathbf{P}$ is replaced by all distributions with finite second moments or finite fourth moments, or even the more restricted family of distributions supported on a compact set (as in the previous section).

Bibliography

BAHADUR, R. AND L. J. SAVAGE (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 25, 1115–1122.

LEHMANN, E. AND W.-Y. LOH (1990): "Pointwise versus uniform robustness in some large-sample tests and confidence intervals," *Scandinavian Journal of Statistics*, 17, 177–187.

LEHMANN, E. AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

ROMANO, J. P. (2004): "On Non-parametric Testing, the Uniform Behaviour of the t-test, and Related Problems," *Scandinavian Journal of Statistics*, 31, 567–584.

Lecture 12

Uniformity of Subsampling

12.1 Intuition Behind Subsampling

Suppose $X_i, i = 1, \ldots, n$ is an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$. Let $\theta(P)$ be some real-valued parameter of interest, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be some estimate of $\theta(P)$. Consider the root

$$T_n = \sqrt{n}(\hat{\theta}_n - \theta(P)) \ ,$$

where root stands for a functional depending on both, the data and $\theta(P)$. Let $J_n(P)$ denote the sampling distribution of T_n and define the corresponding cumulative distribution function as,

$$J_n(x, P) = P\{T_n \leq x\} \ . \tag{12.1}$$

We wish to estimate $J_n(x, P)$ so we can make inferences about $\theta(P)$. For example, we would like to estimate quantiles of $J_n(x, P)$, so we can construct confidence sets for $\theta(P)$. Unfortunately, we do not know P, and, as a result, we do not know $J_n(x, P)$.

The bootstrap solved this problem simply by replacing the unknown P with an estimate $\hat{P}_n$. In the case of i.i.d. data, a typical choice of $\hat{P}_n$ is the empirical distribution of the $X_i, i = 1, \ldots, n$. For this approach to work, we essentially required that $J_n(x, P)$ when viewed as a function of P was continuous in a certain neighborhood of P. An alternative to the bootstrap known as subsampling, originally due to Politis and Romano (1994), does not impose this requirement but rather the following much weaker condition.

Assumption 12.1 *There exists a limiting law $J(P)$ such that $J_n(P)$ converges weakly to $J(P)$ as $n \to \infty$.*

In order to motivate the idea behind subsampling, consider the following thought experiment. Suppose for the time being that $\theta(P)$ is known. Suppose that, instead of n i.i.d. observations from P, we had a very, very

large number of i.i.d. observations from P. For concreteness, suppose $X_i, i = 1, \ldots, m$ is an i.i.d. sequence of random variables with distribution P with $m = nk$ for some very big k. We could then estimate $J_n(x, P)$ by looking at the empirical distribution of

$$\sqrt{n}(\hat{\theta}_n(X_{n(j-1)+1}, \ldots, X_{nj}) - \theta(P)),\ j = 1, \ldots, k\ .$$

This is an i.i.d. sequence of random variables with distribution $J_n(x, P)$. Therefore, by the Glivenko-Cantelli theorem, we know that this empirical distribution is a good estimate of $J_n(x, P)$, at least for large k. In fact, with a simple trick, we could show that it is even possible to improve upon this estimate by using all possible sets of data of size n from the m observations, not just those that are disjoint; that is, estimate $J_n(x, P)$ with the empirical distribution of the

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta(P)),\ j = 1, \ldots, \binom{m}{n}\ .$$

where $\hat{\theta}_{n,j}$ is the estimate of $\theta(P)$ computed using the jth set of data of size n from the original m observations.

In practice $m = n$, so, even if we knew $\theta(P)$, this idea won't work. The key idea behind subsampling is the following simple observation: replace n with some smaller number b that is much smaller than n. We would then expect

$$\sqrt{b}(\hat{\theta}_{b,j} - \theta(P)),\ j = 1, \ldots, \binom{n}{b}\ ,$$

where $\hat{\theta}_{b,j}$ is the estimate of $\theta(P)$ computed using the jth set of data of size b from the original n observations, to be a good estimate of $J_b(x, P)$, at least if $\binom{n}{b}$ is large. Of course, we are interested in $J_n(x, P)$, not $J_b(x, P)$. We therefore need some way to force $J_n(x, P)$ and $J_b(x, P)$ to be close to one another. To ensure this, it suffices to assume that $J_n(x, P) \to J(x, P)$. Therefore, $J_b(x, P)$ and $J_n(x, P)$ are both close to $J(x, P)$, and thus close to one another as well, at least for large b and n. In order to ensure that both b and $\binom{n}{b}$ are large, at least asymptotically, it suffices to assume that $b \to \infty$, but $b/n \to 0$.

This procedure is still not feasible because in practice we typically do not know $\theta(P)$. But we can replace $\theta(P)$ with $\hat{\theta}_n$. This would cause no problems if

$$\sqrt{b}(\hat{\theta}_n - \theta(P)) = \frac{\sqrt{b}}{\sqrt{n}}\sqrt{n}(\hat{\theta}_n - \theta(P))$$

is small, which follows from $b/n \to 0$ in this case.

Essentially, all we required was that $J_n(x, P)$ converged in distribution to a limit distribution $J(x, P)$, whereas for the bootstrap we required this and additionally that $J_n(x, P)$ was continuous in a certain sense. Showing

continuity of $J_n(x, P)$ is problem specific, as the example in the next section illustrates.

Theorem 12.1 *Assume Assumption 12.1. Also, let $J_n(P)$ denote the sampling distribution of $\tau_n(\hat{\theta}_n - \theta(P))$ for some normalizing sequence $\tau_n \to \infty$, $N_n = \binom{n}{b}$, and assume that $\tau_b/\tau_n \to 0$, $b \to \infty$, and $b/n \to 0$ as $n \to \infty$.*

i) If x is a continuity point of $J(\cdot, P)$, then $L_{n,b}(x) \to J(x, P)$ in probability, where

$$L_{n,b}(x) = \frac{1}{N_n} \sum_{j=1}^{N_n} I\{\tau_b(\hat{\theta}_{n,b,j} - \hat{\theta}_n) \leq x\} \ . \tag{12.2}$$

ii) If $J(\cdot, P)$ is continuous, then

$$\sup_x |L_{n,b}(x) - J_n(x, P)| \to 0 \text{ in probability} \ . \tag{12.3}$$

iii) Let

$$c_{n,b}(1-\alpha) = \inf\{x : L_{n,b}(x) \geq 1-\alpha\} \ ,$$
$$c(1-\alpha, P) = \inf\{x : J(x, P) \geq 1-\alpha\} \ .$$

If $J(\cdot, P)$ is continuous at $c(1-\alpha, P)$, then

$$P\{\tau_n(\hat{\theta}_n - \theta(P)) \leq c_{n,b}(1-\alpha)\} \to 1-\alpha \text{ as } n \to \infty \ . \tag{12.4}$$

PROOF. See Politis et al. (1999, Page 43) ■

It is worth noticing that the assumption $b/n \to 0$ and $b \to \infty$ need not imply $\tau_b/\tau_n \to 0$. In regular cases, $\tau_n = n^{1/2}$ and in such case the result follows from $b/n \to 0$.

The result in Theorem 12.1 only requires a very mild condition (i.e., Assumption 12.1). For this reason, subsampling has been traditionally advocated in cases in which the bootstrap is known to be inconsistent (i.e., cases where $J_n(x, P)$ is discontinuous in a certain neighborhood of P). However, the result in Theorem 12.1 is just a pointwise result, and so it is often the case (but not always) that in cases where the bootstrap fails, subsampling fails to be uniformly valid. The next example illustrates this clearly.

12.2 Parameter at the Boundary

Andrews and Guggenberger (2010b) study the properties of subsampling in a broad class of non-regular models. They consider cases in which a test statistic has a discontinuity in its asymptotic distribution as a function of the

true distribution that generates the observations. In such cases bootstrap procedures typically do not provide pointwise asymptotically valid inference, and subsampling has often been advocated.

Consider the following example. Suppose $X_i, i = 1, \ldots, n$ are i.i.d. with distribution $P \in \mathbf{P} = \{N(\theta(P), 1) : \theta(P) \geq 0\}$. The maximum likelihood estimator is $\hat{\theta}_n = \max\{\bar{X}_n, 0\}$. Consider the root

$$\begin{aligned} T_n = \sqrt{n}(\hat{\theta}_n - \theta(P)) &= \sqrt{n}(\max\{\bar{X}_n, 0\} - \theta(P)) \\ &= \max\{\sqrt{n}(\bar{X}_n - \theta(P)), -\sqrt{n}\theta(P)\} \ . \end{aligned} \tag{12.5}$$

It follows that

$$T_n \xrightarrow{d} \begin{cases} \max\{Z, 0\} & \text{if } \theta(P) = 0 \\ Z & \text{if } \theta(P) > 0 \end{cases}$$

where $Z \sim N(0, 1)$. Below it will be convenient to label these distributions as $J_0 \equiv \max\{Z, 0\}$ and $J_\infty \equiv Z$. Before moving to subsampling, we will show that $J_n(x, \hat{P}_n)$ (the bootstrap approximation) does not converge to $J(x, P)$ a.s. in this particular case.

12.2.1 Failure of the Bootstrap

For each n, let $X_{n,i}, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution P_n (not necessarily in $\mathbf{P}$), where P_n converges in distribution to P, $\theta(P_n) \to \theta(P)$, and $\sigma^2(P_n) \to \sigma^2(P)$. The distribution of $J_n(x, P_n)$ is simply the distribution of

$$\begin{aligned} T_n = \sqrt{n}(\hat{\theta}_{n,n} - \theta(P_n)) &= \sqrt{n}(\max\{\bar{X}_{n,n}, 0\} - \theta(P_n)) \\ &= \max\{\sqrt{n}(\bar{X}_{n,n} - \theta(P_n)), -\sqrt{n}\theta(P_n)\} \ . \end{aligned}$$

under P_n. Suppose $\theta(P) = 0$. Let $c > 0$ and suppose $\sqrt{n}\theta(P_n) > c$ for all n. For such a sequence P_n,

$$T_n \leq \max\{\sqrt{n}(\bar{X}_{n,n} - \theta(P_n)), -c\} \xrightarrow{d} \max\{Z, -c\} \ ,$$

under P_n, which is dominated by the distribution of $\max\{Z, 0\}$.

To complete the argument, it suffices to show that $\hat{P}_n$ satisfies a.s. the requirements on P_n in the above discussion. By the SLLN $\hat{P}_n$ converges in distribution to P a.s., $\theta(\hat{P}_n) \to \theta(P)$ a.s., and $\sigma^2(\hat{P}_n) \to \sigma^2(P)$ a.s. It remains to determine whether $\sqrt{n}\theta(\hat{P}_n) > c$ for all n a.s. Equivalently, we need to determine whether

$$\bar{X}_n > \frac{c}{\sqrt{n}} \text{ for all } n \text{ a.s.}$$

Unfortunately, the SLLN will not suffice for this purpose. Instead, we will need the following refinement of the SLLN known as the Law of the Iterated Logarithm (LIL):

Theorem 12.2 *Let $Y_i, i = 1, ..., n$ be an i.i.d. sequence of random variables with distribution P on $\mathbf{R}$. Suppose $\mu(P) = 0$ and $\sigma^2(P) = 1$. Then,*

$$\limsup_{n\to\infty} \frac{\bar{Y}_n}{\sqrt{\frac{2\log\log n}{n}}} = 1 \text{ a.s. and } \liminf_{n\to\infty} \frac{\bar{Y}_n}{\sqrt{\frac{2\log\log n}{n}}} = -1 \text{ a.s.}$$

Recall that for a sequence of real numbers a_n, $n \geq 1$

$$\limsup_{n\to\infty} a_n = a$$

if and only if for any $\epsilon > 0$

$$a_n > a - \epsilon \text{ i.o.}$$

and

$$a_n < a + \epsilon$$

for all n sufficiently large. An implication of the LIL therefore is that for any $\epsilon > 0$,

$$\bar{Y}_n > (1-\epsilon)\sqrt{\frac{2\log\log n}{n}} \text{ i.o. a.s.}$$

In other words, for $W_n = \bar{Y}_n/\sqrt{\frac{2\log\log n}{n}}$, the LIL says that infinitely many of these sequences will come arbitrarily close to 1, but no more than a finite number will exceed it, a.s.

Since $(1-\epsilon)\sqrt{2\log\log n} > c$ for all n sufficiently large, it follows that

$$\bar{Y}_n > \frac{c}{\sqrt{n}} \text{ i.o. a.s.}$$

In other words, there exists a set Ω with $P\{\Omega\} = 1$ such that for all $\omega \in \Omega$,

$$\bar{X}_n(\omega) > \frac{c}{\sqrt{n}} \text{ i.o.}$$

Thus, for all $\omega \in \Omega$ there exists a subsequence $n_k = n_k(\omega), k \geq 1$ of $n \geq 1$ such that for all $k \geq 1$

$$\bar{X}_n(\omega) > \frac{c}{\sqrt{n_k}} \ .$$

It follows that $\hat{P}_n$ satisfies the requirement on P_n, at least along a subsequence, a.s. Thus, at least along the subsequence, $J_n(x, \hat{P}_n)$ does not converge to $J(x, P)$ a.s.

Remark 12.1 The LIL gives an interesting illustration of the difference between almost sure and distributional statements. Under the assumptions of the LIL Theorem, we know that

$$\sqrt{n}\bar{Y}_n \xrightarrow{d} N(0,1) \text{ and } \frac{\sqrt{n}\bar{Y}_n}{\sqrt{n}} \xrightarrow{a.s.} 0$$

so the LIL is giving the rate representing the "knife-egde" between the degenerate and non-degenerate asymptotic distribution (i.e., $\frac{\sqrt{n}}{2\log\log n}$). Note that since $\sqrt{n}\bar{Y}_n \xrightarrow{d} N(0,1)$, the sequence $W_n = \bar{Y}_n/\sqrt{\frac{2\log\log n}{n}}$ in is $(-\epsilon,\epsilon)$ eventually for any $\epsilon > 0$. This does not contradict the LIL that says that W_n reaches the interval $(1-\epsilon, 1+\epsilon)$ infinitely often with probability 1. The explanation is that the set of ω such that $W_n(\omega)$ is in $(-\epsilon,\epsilon)$ or $(1-\epsilon,1+\epsilon)$ fluctuates with n. For a given large value of n, the CLT asserts that a very large fraction of ω have $W_n(\omega) \in (-\epsilon,\epsilon)$. For a given ω, the LIL says that the sequence $W_n(\omega)$ drops in and out of the interval $(1-\epsilon, 1+\epsilon)$ infinitely often.

12.2.2 Subsampling: pointwise behavior

Let's see what happens if we use subsampling. The jth subsample estimator based on a subsample of size $b_n = o(n)$ is $\hat{\theta}_{b_n,j} = \max\{\bar{X}_{b_n,j}, 0\}$, where $\bar{X}_{b_n,j}$ is the sample average of the b_n observations in the jth subsample. In this case,

$$\begin{aligned} T_{b_n,j} = \sqrt{b_n}(\hat{\theta}_{b_n,j} - \theta(P)) &= \sqrt{b_n}(\max\{\bar{X}_{b_n,j}, 0\} - \theta(P)) \\ &= \max\{\sqrt{b_n}(\bar{X}_{b_n,j} - \theta(P)), -\sqrt{b_n}\theta(P)\} \ . \end{aligned} \tag{12.6}$$

It is immediate that for a fixed $\theta(P) = 0$, $T_{b_n,j}(\theta(P)) \xrightarrow{d} J_0$. Also, if $\theta(P) > 0$, we have $T_{b_n,j} \xrightarrow{d} J_\infty$ since $b_n \to \infty$. Thus, as opposed to the bootstrap, subsampling provides the right limiting behavior under standard asymptotics based on a fixed true probability distribution.

Andrews and Guggenberger show that if a sequence of test statistics has an asymptotic null distribution that is discontinuous in a nuisance parameter (as in the previous example), then a subsample test does not necessarily yield the desired asymptotic level. Specifically, the limit of the finite-sample size of the test can exceed its nominal level. The potential problem is not just a small sample problem - it arises with all sample sizes. In particular, subsample tests can have an asymptotic null rejection rate that equals its nominal level under any fixed true distribution, but still the limit of its finite-sample size can be greater than its nominal level. This is due to a *lack of uniformity in the pointwise asymptotics.*

12.2.3 Subsampling: uniform behavior

We will now show, using the previous example, that there are two different rates of drift such that over-rejection and under-rejection can occur. To this end, let γ_n be a "localization sequence" that measures how "far" or "close" we are from $\theta(P) = 0$. This is, we consider a sequence of null distributions P_n

such that $\theta_n = \theta(P_n) = \gamma_n$ and look at the behavior of T_n and $T_{b_n,j}$ along the sequence. The complication arises from the fact that the asymptotic distribution of T_n is discontinuous at $\gamma = 0$.

Remark 12.2 In this example θ_n and γ_n are the same parameter. The reason we introduce the γ_n notation is to allow (in the generalization that follows) for the possibility that the parameter introducing a discontinuity in the asymptotic distribution of T_n is a nuisance parameter different from θ.

Let's start by setting $\gamma_n = h/\sqrt{n}$. In this case,

$$\begin{aligned} T_n = \sqrt{n}(\hat{\theta}_n - \theta_n) &= \max\{\sqrt{n}(\bar{X}_n - \theta_n), -\sqrt{n}\theta_n\} \\ &= \max\{\sqrt{n}(\bar{X}_n - \theta_n), -h\} \\ &\xrightarrow{d} J_h \equiv \max\{Z, -h\} \ , \end{aligned}$$

under P_n and

$$\begin{aligned} T_{b_n,j} &= \sqrt{b_n}(\max\{\bar{X}_{b_n,j}, 0\} - \theta_n) \\ &= \max\{\sqrt{b_n}(\bar{X}_{b_n,j} - \theta_n), -\sqrt{b_n}\theta_n\} \\ &= \max\{\sqrt{b_n}(\bar{X}_{b_n,j} - \theta_n), -(b_n/n)^{1/2}\sqrt{n}\theta_n\} \\ &= \max\{\sqrt{b_n}(\bar{X}_{b_n,j} - \theta_n), -(b_n/n)^{1/2}h\} \\ &\xrightarrow{d} J_0 = \max\{Z, 0\} \ . \end{aligned}$$

Thus, the full-sample test statistic has an asymptotic distribution that depends on a "localization parameter", h, and the subsample critical values behave like the critical value from the asymptotic distribution of the statistic under $h = 0$. Note that for $J_h(x) = P\{\max\{Z, -h\} \leq x\}$,

$$J_h(x) = J_0(x) \text{ for all } x \geq 0 \text{ while } J_h(x) > J_0(x) \text{ for all } x \in [-h, 0) \ , \quad (12.7)$$

so that $J_h(x) \geq J_0(x)$ for all x. That is, the subsample distribution gives a very good approximation of the full-sample distribution in the right tail, but a poor one in the left-tail. Hence, an upper one-sided subsample CI for $\theta(P)$, which relies on a subsample critical value from the right tail of the subsample distribution, will perform well. To see this, let $c_h(1-\alpha)$ denote the $1-\alpha$-quantile of J_h (i.e., $1 - J_h(c_h(1-\alpha)) = \alpha$), we have

$$P_n\{T_n > c_0(1-\alpha)\} \to 1 - J_h(c_0(1-\alpha)) \ ,$$

and

$$1 - J_h(c_0(1-\alpha)) \leq 1 - J_0(c_0(1-\alpha)) = \alpha \ .$$

Indeed, for $\alpha \in (0, 1/2)$ it follows that $c_h(1-\alpha) = c_0(1-\alpha)$, while for $\alpha < 1/2$, $c_h(1-\alpha) < c_0(1-\alpha)$. Thus, a subsample lower one-sided CI

will perform poorly. Furthermore, equal-tailed and symmetric two-sided subsample CIs will perform poorly. To see this, let $\bar{c}_h(1-\alpha)$ be such that

$$1 - J_h(\bar{c}_h(1-\alpha)) + J_h(-\bar{c}_h(1-\alpha)) = \alpha \ .$$

For this critical value, we have

$$P_n\{|T_n| > \bar{c}_0(1-\alpha)\} \to 1 - J_h(\bar{c}_0(1-\alpha)) + J_h(\bar{c}_0(1-\alpha)) \ ,$$

and hence

$$1 - J_h(\bar{c}_0(1-\alpha)) + J_h(-\bar{c}_0(1-\alpha)) \geq 1 - J_0(\bar{c}_0(1-\alpha)) + J_0(-\bar{c}_0(1-\alpha)) = \alpha \ .$$

Thus, subsampling may lead to over-rejection. For example, for $h = 3$ the 95% quantile of the distribution of $|\max\{Z, 0\}|$ is 1.63, while the 95% quantile of the distribution of $|\max\{Z, -h\}|$ is 1.96.

Now let's consider the sequence $\gamma_n = g/\sqrt{b_n}$. In this case,

$$\begin{aligned} T_n = \sqrt{n}(\hat{\theta}_n - \theta_n) &= \max\{\sqrt{n}(\bar{X}_n - \theta_n), -\sqrt{n}\theta_n\} \\ &= \max\{\sqrt{n}(\bar{X}_n - \theta_n), -(n/b_n)^{1/2}\sqrt{b_n}\theta_n\} \\ &= \max\{\sqrt{n}(\bar{X}_n - \theta_n), -(n/b_n)^{1/2} g\} \\ &\xrightarrow{d} J_\infty \equiv Z \end{aligned}$$

under P_n since $(n/b_n)^{1/2} \to \infty$ and

$$\begin{aligned} T_{b_n,j} = \sqrt{b_n}(\max\{\bar{X}_{b_n,j}, 0\} - \theta_n) &= \max\{\sqrt{b_n}(\bar{X}_{b_n,j} - \theta_n), -\sqrt{b_n}\theta_n\} \\ &= \max\{\sqrt{b_n}(\bar{X}_{b_n,j} - \theta_n), -g\} \\ &\xrightarrow{d} J_g = \max\{Z, -g\} \ . \end{aligned}$$

Thus, in this case the full-sample test statistic has an asymptotic distribution that is the same as for fixed $\gamma \neq 0$ and the subsample critical values behave like the critical values from the asymptotic distribution of the full-sample statistic under the drifting sequence with localization parameter g. Thus, here again we might have over-rejection. It turns out that these two sequences determine the limit of the finite-sample size of the test as discussed next.

12.3 Asymptotic Size of Subsampling

Let's consider a family of distributions $\mathbf{P} = \{P_{\theta,\gamma} : \theta \in \Theta, \gamma \in \Gamma\}$. Here γ might be infinite dimensional. The exact size of a test that rejects $H_0 : \theta(P) = \theta_0$ when $T_n(\theta_0) > c_{1-\alpha}$, $ExSz_n$, is the supremum over $\gamma \in \Gamma$ (i.e., the supremum over $\mathbf{P}_0$) of the null rejection probability under γ:

$$ExSz_n = \sup_{\gamma \in \Gamma} P_{\theta_0,\gamma}\{T_n(\theta_0) > c_{1-\alpha}\} \ ,$$

and $P_{\theta_0,\gamma}$ denotes the probability when the true parameters are (θ_0, γ). The asymptotic size of the test is defined by

$$AsySz = \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} P_{\theta_0,\gamma} \{T_n(\theta_0) > c_{1-\alpha}\} \ .$$

Recall that our interest is in the exact finite-sample size of the test; we just use asymptotics to approximate this. Uniformity over $\gamma \in \Gamma$, which is built into the definition of $AsySz_n$, is necessary for the asymptotic size to give a good approximation to the finite-sample size. If $AsySz > \alpha$, the nominal level α test has asymptotic size greater than α and the test does not have correct asymptotic level.

What we learned from the example in the previous section is in fact a general result. Let J_h denote the asymptotic distribution of T_n under a sequence γ_n (i.e., a sequence of distributions P_n) such that

$$h = \lim_{n\to\infty} \sqrt{n}\gamma_n \text{ and } g = \lim_{n\to\infty} \sqrt{b_n}\gamma_n \ , \tag{12.8}$$

for some $h \in H = [0, \infty]$ and $g \in H = [0, \infty]$. Under the *same* sequence P_n, let J_g denote the asymptotic distribution of $T_{b_n,j}$ for the g defined above. The set of all possible pairs of localization parameters (g, h) is denoted by GH and is defined by

$$GH = \{(g, h) \in H \times H : g = 0 \text{ if } h < \infty \ \& \ g \in [0, \infty] \text{ if } h = \infty\} \ .$$

Note that $g \leq h$ for all $(g, h) \in GH$. In the previous example, we got $(g, h) = (0, 0)$ and $(g, h) = (\infty, \infty)$ by standard asymptotics; and $(g, h) = (0, h)$ and $(g, h) = (g, \infty)$ using different drifting sequences.

Lemma 12.1 *Suppose that for all $h \in H$ and all sequences $\{\gamma_n : n \geq 1\}$, $T_n \xrightarrow{d} J_h$ under $\{P_{\theta_0,\gamma_n} : n \geq 1\}$ for some distribution J_h. Then,*

$$AsySz = \sup_{(g,h)\in GH} [1 - J_h(c_g(1-\alpha))] \ ,$$

provided Assumption S in Andrews and Guggenberger (2010b) holds.

Therefore, $AsySz \leq \alpha$ iff $c_g(1-\alpha) \geq c_h(1-\alpha)$. For the details of Assumption S and additional discussion, see Andrews and Guggenberger (2010b, Corollary 1). The general results can be used to show for example that: (i) in an instrumental variables (IVs) regression model with potentially weak IVs, all nominal level $1-\alpha$ one-sided and two-sided subsampling tests concerning the coefficient on an exogenous variable and based on the two-stage least squares (2SLS) estimator have asymptotic size equal to one; (ii) in models where (partially-identified) parameters are restricted by moment inequalities, subsampling tests and CIs based on suitable test statistics have correct asymptotic size.

The approach given above is based on sequences of nuisance parameters and it requires verifying certain assumptions for all possible sequences of nuisance parameters. In particular, proving that a test controls asymptotic size typically requires to argue that it is not possible to find a non-stochastic (sub)sequence of parameters γ_n such that:

$$\lim_{n\to\infty} P_{\theta(P),\gamma_n}\{T_n > c_{1-\alpha}\} > \alpha \ . \tag{12.9}$$

Proving the latter typically involves deriving the asymptotic distribution of the test statistic along all possible non-stochastic sequences $\gamma_n \in \Gamma$.

A different approach includes the one in Romano and Shaikh (2012). In that paper the authors show that subsampling tests are valid whenever the family $\mathbf{P}$ satisfies,

$$\lim_{n\to\infty} \sup_{P\in\mathbf{P}} \sup_{x\in\mathbf{R}} |J_b(x,P) - J_n(x,P)| = 0 \ , \tag{12.10}$$

see Theorem 2.1 in Romano and Shaikh (2012). The authors also provide uniform results for Bootstrap tests.

Bibliography

Andrews, D. W. K. and P. Guggenberger (2009a): "Hybrid and size-corrected subsample methods," *Econometrica*, 77, 721–762.

——— (2009b): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669–709.

——— (2010a): "Applications of Subsampling, Hybrid, and Size-Correction Methods," *Journal of Econometrics*, 158, 285–305.

——— (2010b): "Asymptotic Size and a Problem with Subsampling and with the m Out of n Bootstrap," *Econometric Theory*, 26, 426–468.

Politis, D. N. and J. P. Romano (1994): "Large sample confidence regions based on subsamples under minimal assumptions," *Annals of Statistics*, 22, 2031–2050.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer, New York.

Romano, J. P. and A. M. Shaikh (2012): "On the Uniform Asymptotic Validity of Subsampling and the Bootstrap," *The Annals of Statistics*, 40, 2798–2822.

Lecture 13

Moment Inequality Models I

There are no lecture notes for this topic. You are supposed to read two papers and the slides we used in class. The two papers are Canay and Shaikh (2017) and Ho and Rosen (2017).

Bibliography

Canay, I. A. and A. M. Shaikh (2017): "Practical and Theoretical Advances for Inference in Partially Identified Models," in *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, vol. 2, 271–306.

Ho, K. and A. M. Rosen (2017): "Partial Identification in Applied Research: Benefits and Challenges," in *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, vol. 2, 307–360, manuscript, UCL.

Lecture 14

Moment Inequality Models I

There are no lecture notes for this topic. You are supposed to read two papers and the slides we used in class. The two papers are Canay and Shaikh (2017) and Ho and Rosen (2017).

Bibliography

Canay, I. A. and A. M. Shaikh (2017): "Practical and Theoretical Advances for Inference in Partially Identified Models," in *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, vol. 2, 271–306.

Ho, K. and A. M. Rosen (2017): "Partial Identification in Applied Research: Benefits and Challenges," in *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, vol. 2, 307–360, manuscript, UCL.

Bibliography

Andrews, D. W. K. and P. Guggenberger (2009a): "Hybrid and size-corrected subsample methods," *Econometrica*, 77, 721–762.

——— (2009b): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669–709.

——— (2010a): "Applications of Subsampling, Hybrid, and Size-Correction Methods," *Journal of Econometrics*, 158, 285–305.

——— (2010b): "Asymptotic Size and a Problem with Subsampling and with the m Out of n Bootstrap," *Econometric Theory*, 26, 426–468.

Bahadur, R. and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 25, 1115–1122.

Billingsley, P. (1995): *Probability and Measure*, Wiley-Interscience.

Canay, I. A., A. Santos, and A. M. Shaikh (2013): "On the Testability of Identification in Some Nonparametric Models with Endogeneity," *Econometrica*, 81, 2535 – 2559.

Canay, I. A. and A. M. Shaikh (2017): "Practical and Theoretical Advances for Inference in Partially Identified Models," in *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, vol. 2, 271–306.

Hirano, K. and J. R. Porter (2012): "Impossibility results for nondifferentiable functionals," *Econometrica*, 80, 1769–1790.

Ho, K. and A. M. Rosen (2017): "Partial Identification in Applied Research: Benefits and Challenges," in *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, vol. 2, 307–360, manuscript, UCL.

Lehmann, E. and W.-Y. Loh (1990): "Pointwise versus uniform robustness in some large-sample tests and confidence intervals," *Scandinavian Journal of Statistics*, 17, 177–187.

Lehmann, E. and J. P. Romano (2005): *Testing Statistical Hypotheses*, Springer, New York, 3rd ed.

Politis, D. N. and J. P. Romano (1994): "Large sample confidence regions based on subsamples under minimal assumptions," *Annals of Statistics*, 22, 2031–2050.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer, New York.

Pollard, D. (2002): *A User's Guide to Measure Theoretic Probability*, Cambrigde University Press, New York.

Romano, J. P. (2004): "On Non-parametric Testing, the Uniform Behaviour of the t-test, and Related Problems," *Scandinavian Journal of Statistics*, 31, 567–584.

Romano, J. P. and A. M. Shaikh (2012): "On the Uniform Asymptotic Validity of Subsampling and the Bootstrap," *The Annals of Statistics*, 40, 2798–2822.

Santos, A. and Z. Fang (2014): "Inference on Directionally Differentiable Functions," ArXiv:1404.3763.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press, Cambridge.