

Chapter 10

Confidence Intervals and Tests

Until now we've focused primarily on point estimation. Since data are not infinite, point estimation involves uncertainty. We should not be overconfident in our claim to know the data-generating process. Rather, we should be only as confident as the data allows us to be.

We discussed uncertainty previously when we considered sampling distributions. In this chapter we further quantify estimation uncertainty using the language of confidence sets. We also cover hypothesis tests, which are a second key method of inference in classical statistics.

To tie in with standard presentations we'll switch to parametric language. The notation is as on page 214 except that

- (i) the marginal distribution P of each observation $\mathbf{z}_n$ is written as P_θ ,
- (ii) the universe $\mathcal{P}$ of possible distributions is $\{P_\theta : \theta \in \Theta\}$, and
- (iii) the joint distribution of the data $\mathbf{z}_D = (\mathbf{z}_1, \dots, \mathbf{z}_N)$ is written as P_θ^D .

However, we won't actually restrict ourselves to the case of finitely many parameters. Instead, θ can be thought of more generally as an index over the set of distributions.

10.1 Confidence Sets

Confidence sets are random subsets of model space that contain the true data generating process with high probability. Let's start with the key definitions and then move on to examples.

10.1.1 Finite Sample Confidence Sets

Fix $\alpha \in (0, 1)$. A random set $C(\mathbf{z}_D) \subset \Theta$ is called a **$1 - \alpha$ confidence set** for θ if $C(\mathbf{z}_D)$ is observable given the data and, for any $\theta \in \Theta$,

$$\mathbb{P}\{\theta \in C(\mathbf{z}_D)\} \geq 1 - \alpha \quad \text{whenever} \quad \mathcal{L}(\mathbf{z}_D) = P_\theta^D \quad (10.1)$$

Note that it's the set that's random here, not θ .

Confidence sets that are intervals are usually called **confidence intervals**.

Suppose, for example, that we want to estimate the mean of an unknown distribution based on a set of observations $\mathbf{z}_D$. If we can come up with a 95% confidence set $C(\mathbf{z}_D)$ in the sense of (10.1), then we know that the set will contain the mean in about 95% of our experiments, regardless of the underlying distribution.

The last part of the sentence is essential: We don't know what θ is, and so the set *must be designed* such that the probability of containing θ exceeds $1 - \alpha$ regardless of which θ is generating the data. The difficulty is to actually construct a confidence set with this property for a reasonably large class of distributions. Here's a traditional approach that uses strong parametric assumptions.

Example 10.1.1 Let $\mathbf{z}_D = \mathbf{x} = (x_1, \dots, x_N)$ where each x_n is an independent draw from $N(\mu, \sigma^2)$. Suppose for now that only $\mu \in \mathbb{R}$ is unknown. We wish to form a confidence interval for μ . By (9.3), we have

$$\mathcal{L} \left[\sqrt{N} \frac{(\bar{x}_N - \mu)}{\sigma} \right] = N(0, 1) \quad (10.2)$$

Notice that this statement is true regardless of the values of μ and σ . Applying (4.36) on page 115 now gives

$$\mathbb{P} \left\{ \frac{\sqrt{N}}{\sigma} |\bar{x}_N - \mu| \leq z_{\alpha/2} \right\} = 1 - \alpha \quad \text{when} \quad z_{\alpha/2} := \Phi^{-1} \left(1 - \frac{\alpha}{2} \right)$$

Here Φ is the standard normal CDF. Some rearranging gives

$$\mathbb{P} \left\{ \bar{x}_N - \frac{\sigma}{\sqrt{N}} z_{\alpha/2} \leq \mu \leq \bar{x}_N + \frac{\sigma}{\sqrt{N}} z_{\alpha/2} \right\} = 1 - \alpha$$

Since this argument is true regardless of the value of μ , we conclude that the interval

$$C(\mathbf{x}) := (\bar{x}_N - e_n, \bar{x}_N + e_n) \quad \text{with} \quad e_n := \frac{\sigma}{\sqrt{N}} z_{\alpha/2}$$

is a $1 - \alpha$ confidence interval for μ . (Note that $C(\mathbf{x})$ is indeed observable under our

assumption that σ is known.)

In example 10.1.1, the assumption that σ is known is clearly unrealistic. If σ is unknown, a natural approach is to replace this term with the sample standard deviation s_N . In doing so, we will use the following fact, which exercise 10.4.1 asks you to prove.

Fact 10.1.1 If $x_1, \dots, x_N$ are IID draws from $N(\mu, \sigma^2)$, then

$$\sqrt{N-1} \frac{(\bar{x}_N - \mu)}{s_N} \quad (10.3)$$

has a Student's t -distribution with $N - 1$ degrees of freedom.

The random variable (10.3) is called a **pivot**, which means that its distribution does not depend on the unknown parameters.

Example 10.1.2 Consider then the setting of example 10.1.1 but with σ also unknown. Let F_{N-1} be the CDF of the t -distribution with $N - 1$ degrees of freedom. Using the same reasoning as example 10.1.1 with $z_{\alpha/2}$ replaced by $t_{\alpha/2} := F_{N-1}^{-1}(1 - \alpha/2)$, we obtain

$$\mathbb{P} \left\{ \bar{x}_N - \frac{s_N}{\sqrt{N-1}} t_{\alpha/2} \leq \mu \leq \bar{x}_N + \frac{s_N}{\sqrt{N-1}} t_{\alpha/2} \right\} = 1 - \alpha$$

Recall that the standard deviation of $\bar{x}_N$ is $\sigma/\sqrt{N}$. The term $s_N/\sqrt{N-1}$ is a sample estimate of $\sigma/\sqrt{N}$. Following the discussion in §9.1.3.2, we call $s_N/\sqrt{N-1}$ the standard error of $\bar{x}_N$ and write it as $\text{se}(\bar{x}_N)$. We can then write the confidence interval in example 10.1.2 as

$$C(\mathbf{x}) := (\bar{x}_N - \text{se}(\bar{x}_N) t_{\alpha/2}, \bar{x}_N + \text{se}(\bar{x}_N) t_{\alpha/2}) \quad (10.4)$$

10.1.2 Asymptotic Methods

While strong parametric assumptions of the form used in examples 10.1.1 and 10.1.2 are difficult to justify in most econometric scenarios, there is another approach to forming confidence intervals that requires much less in terms of structure imposed on the underlying distributions, provided that the sample size is relatively large. Let's start with a definition: given $\alpha \in (0, 1)$, a set $C_N(\mathbf{z}_D) \subset \Theta$ is called an **asymptotic $1 - \alpha$ confidence set** for θ if $C_N(\mathbf{z}_D)$ is observable given the data and

$$\lim_{N \rightarrow \infty} \mathbb{P}\{\theta \in C_N(\mathbf{z}_D)\} \geq 1 - \alpha \quad (10.5)$$

for all $\theta \in \Theta$. Note that $C_N(\mathbf{z}_D)$ is really a sequence of sets and the definition concerns this sequence.

How can we create asymptotic confidence sets? Let's consider first the case of scalar θ . Suppose that we have an estimator $\hat{\theta}_N$ of θ that is asymptotically normal for all $\theta \in \Theta$. Specializing the results in §9.2.4.1 to the scalar case, this means that, for all $\theta \in \Theta$, there exists a positive constant $v(\theta)$ such that

$$\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, v(\theta)) \quad \text{as } N \rightarrow \infty \quad (10.6)$$

The constant $v(\theta)$ is called the asymptotic variance of $\hat{\theta}_N$. Suppose we have a sequence of statistics $\text{se}(\hat{\theta}_N)$ such that

$$\sqrt{N} \text{se}(\hat{\theta}_N) \xrightarrow{p} \sqrt{v(\theta)} \quad \text{as } N \rightarrow \infty \quad (10.7)$$

As exercise 10.4.2 asks you to show, (10.6) and (10.7) imply that

$$\frac{\hat{\theta}_N - \theta}{\text{se}(\hat{\theta}_N)} \xrightarrow{d} N(0, 1) \quad \text{as } N \rightarrow \infty \quad (10.8)$$

We can now create our asymptotic confidence intervals as

$$C_N := (\hat{\theta}_N - \text{se}(\hat{\theta}_N)z_{\alpha/2}, \hat{\theta}_N + \text{se}(\hat{\theta}_N)z_{\alpha/2}) \quad (10.9)$$

To see this, take the limit and apply (10.8) to get

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathbb{P}\{\theta \in C_N(\mathbf{z}_D)\} &= \lim_{N \rightarrow \infty} \mathbb{P}\{\hat{\theta}_N - \text{se}(\hat{\theta}_N)z_{\alpha/2} \leq \theta \leq \hat{\theta}_N + \text{se}(\hat{\theta}_N)z_{\alpha/2}\} \\ &= \lim_{N \rightarrow \infty} \mathbb{P}\left\{-z_{\alpha/2} \leq \frac{\hat{\theta}_N - \theta}{\text{se}(\hat{\theta}_N)} \leq z_{\alpha/2}\right\} \\ &= 1 - \alpha \end{aligned}$$

Example 10.1.3 Let $\bar{x}_N$ be the sample mean of IID data $\{x_n\}$. Suppose that these data come from some distribution with finite second moment. Let μ denote the common mean and let σ^2 denote the variance. Let s_N be the sample standard deviation. Combining the central limit theorem and consistency of s_N for σ (see page 263), we have

$$\sqrt{N}(\bar{x}_N - \mu) \xrightarrow{d} N(0, \sigma^2) \quad \text{and} \quad \sqrt{N} \text{se}(\bar{x}_N) \xrightarrow{p} \sigma \quad \text{for} \quad \text{se}(\bar{x}_N) := \frac{s_N}{\sqrt{N}} \quad (10.10)$$

It follows that, with this definition of $\text{se}(\bar{x}_N)$, the set

$$(\bar{x}_N - \text{se}(\bar{x}_N)z_{\alpha/2}, \bar{x}_N + \text{se}(\bar{x}_N)z_{\alpha/2})$$

is an asymptotic $1 - \alpha$ confidence interval for $\bar{x}_N$.

10.1.3 A Nonparametric Example

Here's an example of a nonparametric confidence set. In §8.1.3 we learned that if $\mathbf{x} = (x_1, \dots, x_N)$ is a vector of IID draws from some CDF F and $\hat{F}_N$ is the corresponding ECDF, then $\|F - \hat{F}_N\|_\infty$ converges in probability to 0. In 1933, A. N. Kolmogorov used an extension of the central limit theorem to obtain an asymptotic distribution for this term. In particular, he showed that when F is continuous,

$$\sqrt{N} \sup_{s \in \mathbb{R}} |\hat{F}_N(s) - F(s)| \xrightarrow{d} K \quad (10.11)$$

where K is the **Kolmogorov** CDF

$$K(s) := \frac{\sqrt{2\pi}}{s} \sum_{i=1}^{\infty} \exp \left[-\frac{(2i-1)^2 \pi^2}{8s^2} \right] \quad (s \geq 0) \quad (10.12)$$

As in the CLT, the limiting distribution K is independent of the CDF F that generates the data.

We can use (10.11) to produce an asymptotic $1 - \alpha$ confidence set for F . To do so, let $\mathfrak{F}$ be the set of all CDFs on $\mathbb{R}$, let $x_1, \dots, x_N$ be IID draws from $F \in \mathfrak{F}$, let $k_{1-\alpha} := K^{-1}(1 - \alpha)$, and let

$$C_N(\mathbf{x}) := \left\{ G \in \mathfrak{F} : \hat{F}_N(s) - \frac{k_{1-\alpha}}{\sqrt{N}} \leq G(s) \leq \hat{F}_N(s) + \frac{k_{1-\alpha}}{\sqrt{N}} \text{ for all } s \in \mathbb{R} \right\}$$

The set $C_N(\mathbf{x}) \subset \mathfrak{F}$ is an asymptotic $1 - \alpha$ confidence set for F . Indeed, after rearranging the expression, we get

$$\begin{aligned} F \in C_N(\mathbf{x}) &\iff -k_{1-\alpha} \leq \sqrt{N}(\hat{F}_N(s) - F(s)) \leq k_{1-\alpha} \text{ for all } s \\ &\iff \sqrt{N}|\hat{F}_N(s) - F(s)| \leq k_{1-\alpha} \text{ for all } s \\ &\iff \sup_s \sqrt{N}|\hat{F}_N(s) - F(s)| \leq k_{1-\alpha} \end{aligned}$$

Applying (10.11) now confirms our claim:

$$\lim_{N \rightarrow \infty} \mathbb{P}\{F \in C_N(\mathbf{x})\} = \lim_{N \rightarrow \infty} \mathbb{P} \left\{ \sup_s \sqrt{N}|\hat{F}_N(s) - F(s)| \leq k_{1-\alpha} \right\} = 1 - \alpha$$

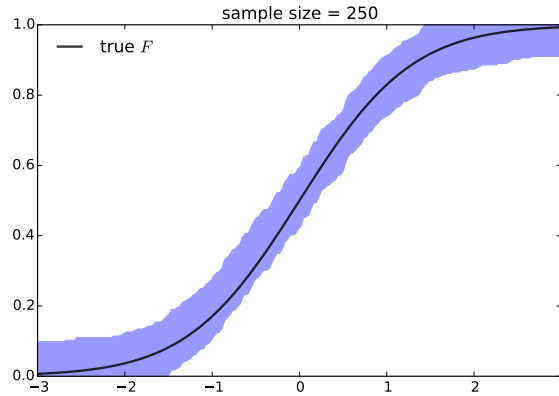


Figure 10.1 Confidence set for the ECDF with 250 observations

Given our data $x_1, \dots, x_N$ and the corresponding ECDF $\hat{F}_N$, we can present the confidence set $C_N(\mathbf{x})$ visually by highlighting the area between the lower bound $\hat{F}_N(s) - k_{1-\alpha}/\sqrt{N}$ and the upper bound $\hat{F}_N(s) + k_{1-\alpha}/\sqrt{N}$ over s . This is done in figures 10.1 and 10.2. The data are drawn from a t -distribution F with 10 degrees of freedom and α is set to 0.05. The values of N are 250 for figure 10.1 and 1,000 for figure 10.2.

10.2 Hypothesis Tests

This section considers problems of inference where we

- (i) hold a belief or theory concerning the probabilities generating the data and
- (ii) consider whether the data provides evidence for or against that theory.

Example 10.2.1 While asset price returns data tend to have heavier tails than the normal distribution, for certain assets normality might still be a reasonable and convenient approximation. Let's look at whether or not normality provides a reasonable fit in the case of daily returns on the Nikkei 225. Figure 10.3 shows a comparison of the empirical distribution of standardized returns from January 2014 until July 2015 with the CDF of the standard normal distribution.¹ While the ECDF of the data and the CDF are not identical, we wouldn't expect this even if the hypothesis of normality is completely true. After all, we are only observing a sample, and a sample is subject to

1. Returns are standardized by subtracting the sample mean and dividing by the sample standard deviation. Data are sourced from the Federal Reserve of St. Louis FRED data set.

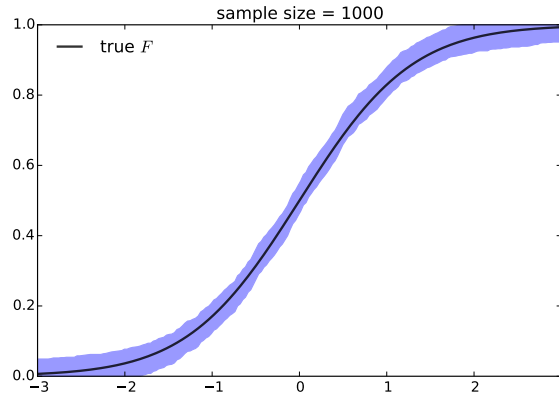


Figure 10.2 Confidence set for the ECDF with 1,000 observations

randomness. So the question becomes, is the sample “unlikely” given the hypothesis of normality?

To take the next steps, we must formalize the process of setting up a hypothesis and quantifying the extent to which the data points in this or some other direction.

Hypothesis testing begins with the specification of a **null hypothesis**, which is a statement that the observed data are being generated by a certain model, or one of a certain class of models. One might imagine that the standard procedure in statistics is to show the validity of the null hypothesis but it is not. Rather, a hypothesis test is an attempt to reject the null.

To see why, let’s recall a famous example due to Karl Popper (1902–1994). Consider the claim that all swans are white. It is futile to attempt to prove correctness of this hypothesis: No quantity of white swan sightings can confirm the claim that *all* swans on planet earth are white. Yet a single black swan can show the claim to be false. In this sense, attempting to falsify a theory on the basis of observation is more constructive than attempting to confirm it.

Although we attempt to reject the null hypothesis, we do so only if strong evidence against it is observed. The rationale is this: Suppose that we have a collection of theories about how the economy works. The procedure would then be to step through the theories, at each stage taking correctness of the theory as the null hypothesis and attempting to reject. If the theory is rejected, then we can discard it. This is a useful process of elimination. However, we don’t want to mistakenly discard a good theory. Hence we only reject when there is strong evidence against the null hypothesis.

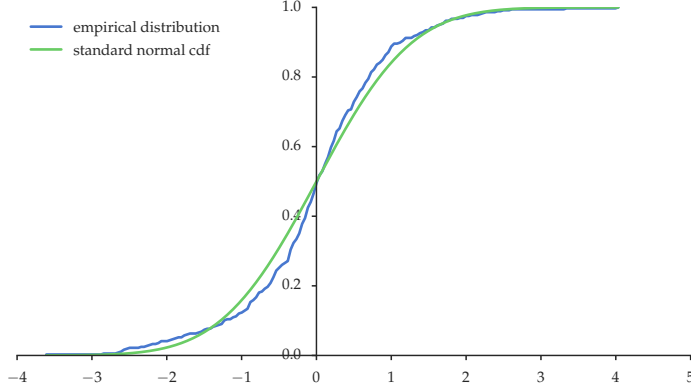


Figure 10.3 ECDF for standardized returns versus the standard normal CDF

10.2.1 Constructing Tests

We adopt the setting and notation used in §10.1.1. A null hypothesis is a specification of a set of models that we believe the data-generating process belongs to. This amounts to specifying a subset Θ_0 of Θ . The null hypothesis is often written as

$$H_0: \theta \in \Theta_0$$

If Θ_0 is a singleton, then the null hypothesis is called a **simple hypothesis**. If not, then the null hypothesis is called a **composite hypothesis**. A test of the null hypothesis is a test of whether or not the observed data were generated by P_θ for some $\theta \in \Theta_0$.

Example 10.2.2 The hypothesis of purchasing power parity is tested by considering models such as

$$p = a + \beta e p^* + \sigma u \quad (10.13)$$

where p is a price or index of domestic prices, e is an exchange rate, p^* is a corresponding foreign price, u is a disturbance term and a, β, σ are parameters. The null hypothesis for absolute purchasing power parity is that (10.13) holds with $a = 0$ and $\beta = 1$. Since σ is not pinned down by the null, this is a composite null hypothesis.

Formally, a **test** of H_0 is a binary function ϕ mapping the observed data $\mathbf{z}_D$ into $\{0, 1\}$. The decision rule is

$$\begin{aligned} &\text{if } \phi(\mathbf{z}_D) = 1, \text{ then reject } H_0 \\ &\text{if } \phi(\mathbf{z}_D) = 0, \text{ then do not reject } H_0 \end{aligned}$$

(Failing to reject H_0 should not be confused with accepting H_0 . More on this below.) Our aim will be to design ϕ such that it takes the value 1 when the data present strong evidence against H_0 .

Example 10.2.3 Let $\mathbf{z}_{\mathcal{D}}$ be a set of pairs $\mathbf{z}_n = (x_n, y_n)$ from some unknown bivariate distribution P_θ . If H_0 is the hypothesis that the correlation between x and y is negative, then a large positive sample correlation would constitute evidence against H_0 . Hence our test might take the form $\phi(\mathbf{z}_{\mathcal{D}}) = \mathbb{1}\{\hat{\rho} > c\}$, where $\hat{\rho}$ is the sample correlation. An appropriate value of c remains to be determined.

The outcome of our test depends on the random sample $\mathbf{z}_{\mathcal{D}}$, and being random, its realization can be misleading. There are two different ways in which the realization can mislead us. First, we can mistakenly reject the null hypothesis when it is in fact true. This is called **type I error**. Second, we can fail to reject the null hypothesis when it is false. This is called **type II error**.

The **power function** associated with the test ϕ is the function

$$\beta(\theta) := \mathbb{P}_\theta\{\phi(\mathbf{z}_{\mathcal{D}}) = 1\} \quad (\theta \in \Theta)$$

Here and below, the symbol $\mathbb{P}_\theta$ indicates that we are computing probabilities under the assumption that $\mathcal{L}(\mathbf{z}_{\mathcal{D}}) = P_\theta^{\mathcal{D}}$. Thus, $\beta(\theta)$ is the probability that the test rejects when the data are generated by the probabilistic model identified by θ .

Ideally, we would like $\beta(\theta) = 0$ when $\theta \in \Theta_0$, and $\beta(\theta) = 1$ when $\theta \notin \Theta_0$. In practice, this is difficult to achieve.

As discussed above, we tend to be conservative in rejecting the null because we don't want to discard good theories. For this reason it is traditional to keep the probability of type I error small. Then, if our test tells us to reject the null, it's unlikely the null is true. Because of this, the standard procedure is to choose a small number α such as 0.05 or 0.01, and then adjust the test such that

$$\beta(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0 \quad (10.14)$$

If (10.14) holds, then the test is said to be of **size α** . Similarly the **size** of any test with power function β is

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta) \quad (10.15)$$

This is the maximal rejection probability when the null hypothesis is true.

In constructing tests, a common (but not universal) setup is to define a real-valued **test statistic** T and a **critical value** c , and then set

$$\phi(\mathbf{z}_{\mathcal{D}}) := \mathbb{1}\{T(\mathbf{z}_{\mathcal{D}}) > c\} \quad (10.16)$$

The pair (T, c) then defines the test, and the rule becomes:

$$\text{reject } H_0 \text{ if and only if } T(\mathbf{z}_D) > c$$

Example 10.2.4 Suppose that $x_1, \dots, x_N$ are independent draws from $N(\mu, 1)$ where the value of μ is unknown. The null hypothesis is $\mu \leq 0$, or $\Theta_0 = (-\infty, 0]$. Since we want to make inference about the mean, a natural choice for our test statistic is the sample mean. Thus

$$T(\mathbf{z}_D) := T(x_1, \dots, x_N) := \bar{x}_N$$

Each $c \in \mathbb{R}$ now gives us a test via (10.16), with power function $\beta(\mu) = \mathbb{P}_\mu\{\bar{x}_N > c\}$. To evaluate β , recall that $\mathcal{L}(\bar{x}_N) = N(\mu, 1/N)$. As a result, if Φ is the CDF of the standard normal distribution and $\mathcal{L}(z) = \Phi$, then

$$\begin{aligned} \mathbb{P}_\mu\{\bar{x}_N \leq c\} &= \mathbb{P}\{\mu + N^{-1/2}z \leq c\} = \mathbb{P}\{z \leq N^{1/2}(c - \mu)\} = \Phi[N^{1/2}(c - \mu)] \\ \therefore \quad \beta(\mu) &= 1 - \Phi[N^{1/2}(c - \mu)] \end{aligned} \quad (10.17)$$

Given c , the power function is increasing in μ because higher μ pushes up the mean of $\bar{x}_N$, making the event $\{\bar{x}_N > c\}$ more likely. Given μ , the function is decreasing in c , because higher c makes the event $\{\bar{x}_N > c\}$ less likely.

Plots of the power function β in (10.17) are presented in figure 10.4 for two different values of c . Here N is fixed at 10. Since $\Theta_0 = (-\infty, 0]$, the size of the test in each case is $\sup_{\mu \leq 0} \beta(\mu)$. Since the power curves are increasing, this is just $\beta(0)$.

Figure 10.4 illustrates a typical trade-off between type I and type II error. If we increase c , we make rejection less likely for all values of μ . This pushes down type I error, but it also increases the probability that we fail to reject a false null.

10.2.2 Choosing Critical Values

Let's think a bit more about the test in (10.16). Typically, the choice of T is suggested by the problem. For example, if our hypothesis is a statement about the second moment of a random variable, then we might take T to be the sample second moment. Once T is fixed, we need to adjust c such that (T, c) attains the appropriate size. Thus the standard procedure is to

- (i) choose a desired size α according to our tolerance for type I error,
- (ii) identify a suitable test statistic T , and
- (iii) choose a critical value c so that (T, c) is of size α .

In performing the last step, we can balance our desire to minimize type II error while maintaining a size α test by choosing c such that (10.14) holds with equality. In

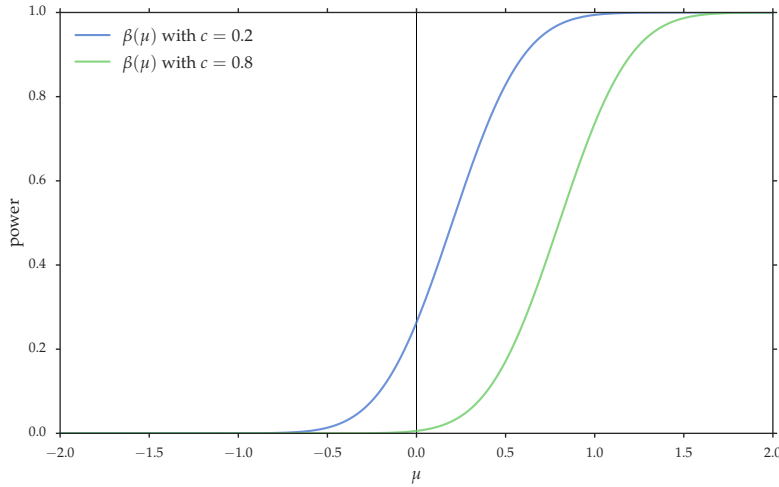


Figure 10.4 Power function β

this case, the problem is to choose c to solve

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta}\{T(\mathbf{z}_{\mathcal{D}}) > c\} \quad (10.18)$$

Figure 10.5 gives an illustration. In the figure, we've taken Θ_0 to be the two element set $\{\theta_a, \theta_b\}$. The blue line gives an imaginary distribution of $T(\mathbf{z}_{\mathcal{D}})$, represented as a density, when $\mathbf{z}_{\mathcal{D}}$ is generated by θ_a . The black line gives the same for θ_b . Assuming that a value of α is prescribed, our next step is to determine the c such that (10.18) holds. Here we choose c such that the largest of the two shaded areas is equal to α .

Example 10.2.5 Let's look again at example 10.2.4, where $x_1, \dots, x_N \stackrel{\text{iid}}{\sim} N(\mu, 1)$ for μ unknown, and our null hypothesis is $\mu \leq 0$. Given α , our task is to find the appropriate critical value c so that the test (T, c) is of size α . To solve for c given α , we use (10.18). Applying the expression for the power function in (10.17), this becomes

$$\alpha = \sup_{\mu \leq 0} \{1 - \Phi[N^{1/2}(c - \mu)]\}$$

The right-hand side is increasing in μ , so the supremum is obtained by setting $\mu = 0$.

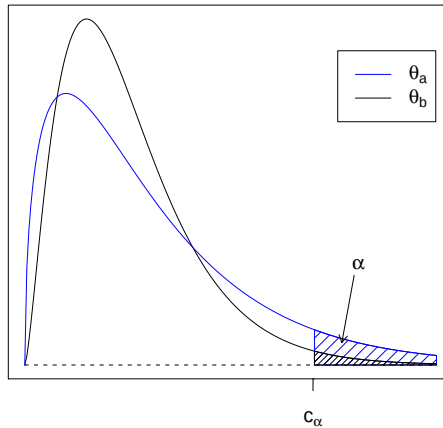


Figure 10.5 Determining the critical value

Setting $\mu = 0$ and solving for c , we obtain

$$c(\alpha) := N^{-1/2} \Phi^{-1}(1 - \alpha)$$

where Φ^{-1} is the quantile function of the standard normal distribution. In R, this function can be evaluated using `qnorm`. For example,

```
> alpha = 0.05
> qnorm(1 - alpha)
[1] 1.644854
```

Since Φ^{-1} is increasing, we see that smaller α corresponds to larger $c(\alpha)$ —we reduce the probability of type I error by increasing the critical value that the mean $\bar{x}_N$ must obtain for rejection to occur. Also higher N brings down $c(\alpha)$: more data allows us to reduce the critical value without increasing the probability of rejecting a true null.

10.2.3 Asymptotic Tests

In general, constructing a test of a given size requires some knowledge of the distribution of the test statistic under the null hypothesis. In many cases, however, we know relatively little about the distribution of the test statistic. Fortunately, we still have options. One approach is to bootstrap. Another is to try to determine the asymptotic distribution of the test statistic. As we've learned from the central limit theory,

asymptotics hold out the possibility that we can pin down distributions without actually assuming a parametric structure *a priori*.

To proceed, we need to switch to a notion of asymptotic size for tests, rather than finite sample size. Writing β_N instead of β to emphasize the fact that the power function typically depends on sample size, a test is called **asymptotically of size α** if

$$\lim_{N \rightarrow \infty} \beta_N(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0 \quad (10.19)$$

Example 10.2.6 Let $x_1, \dots, x_N$ be an IID sample with mean θ and variance σ^2 . Assume that both θ and σ are unknown. We wish to test the hypothesis $H_0: \theta = \theta_0$. Consider the statistic

$$t_N := \sqrt{N} \left\{ \frac{\bar{x}_N - \theta_0}{s_N} \right\} = \frac{\bar{x}_N - \theta_0}{\text{se}(\bar{x}_N)} \quad (10.20)$$

where $\text{se}(\bar{x}_N) := \frac{s_N}{\sqrt{N}}$ as in (10.10). As discussed in example 10.1.3, this expression converges in distribution to a standard normal. Hence

$$\phi_N(\mathbf{x}) := \mathbb{1}\{|t_N| > z_{\alpha/2}\} \quad (10.21)$$

is asymptotically of size α (see ex. 10.4.3).

10.2.3.1 A Test for the Empirical Distribution

As another example of how to construct tests which are asymptotically of size α , let's return to the standardized daily returns data presented in figure 10.3. Let Φ be the standard normal CDF as before, and let H_0 be that standardized returns are IID draws from Φ . Let α be given, and let $k_{1-\alpha} = K^{-1}(1 - \alpha)$ be the $1 - \alpha$ quantile of the Kolmogorov distribution K , as defined in (10.12). Finally, let $\hat{F}_N$ be the ECDF of the data. If the null hypothesis is true, then, by (10.11) on page 279, we have

$$\sqrt{N} \sup_{s \in \mathbb{R}} |\hat{F}_N(s) - \Phi(s)| \xrightarrow{d} K \quad (10.22)$$

For the test

$$\phi_N(\mathbf{x}) := \mathbb{1} \left\{ \sqrt{N} \sup_{s \in \mathbb{R}} |\hat{F}_N(s) - \Phi(s)| > k_{1-\alpha} \right\}$$

let $\beta_N(\Phi)$ be the value of the power function when the null hypothesis is true. By (10.22), we have

$$\lim_{N \rightarrow \infty} \beta_N(\Phi) = \lim_{N \rightarrow \infty} \mathbb{P} \left\{ \sqrt{N} \sup_{s \in \mathbb{R}} |\hat{F}_N(s) - \Phi(s)| > k_{1-\alpha} \right\} = \alpha$$

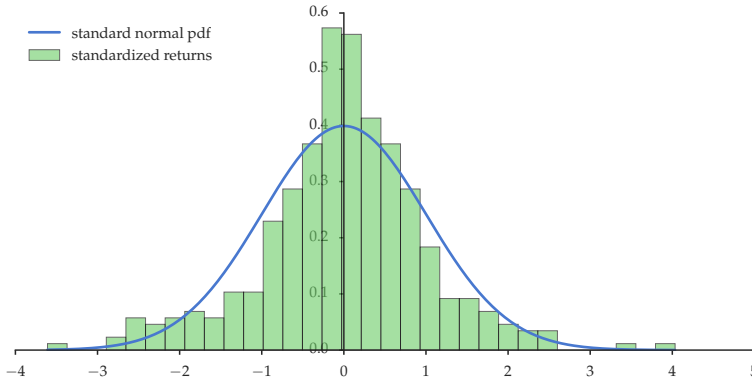


Figure 10.6 Standardized daily returns for the Nikkei 225

Hence (10.19) is verified, and the test is asymptotically of size α . Notice how the construction of the test inverts the confidence set in §10.1.3.

The Nikkei data set used here has 364 observations. The value of the test statistic $\sqrt{N} \sup_{s \in \mathbb{R}} |\hat{F}_N(s) - \Phi(s)|$ is 1.548. If $\alpha = 0.05$, then, as shown in §10.1.3, the critical value $k_{1-\alpha}$ is 1.36. Hence the test statistic exceeds the critical value, and the null hypothesis is rejected. See johnstachurski.net/emet.html for the code.

While the departure from normality is not easy to see in figure 10.3, it can be seen more clearly in a comparison of the histogram and density, as in figure 10.6. The data are noticeably more peaked and there is evidence of heavy tails.

This application also illustrates how tests often depend on a range of assumptions beyond those that are the focus of a particular hypothesis. Here, for example, a full description of H_0 is that standardized returns are IID draws from the standard normal density. The rejection suggests this null is false, but it may be the IID assumption rather than the normality assumption that makes our null a poor fit to the data.²

10.2.3.2 p -Values

Typically, a test that rejects at size 0.05 will also reject at size 0.1, but may not reject at size 0.01. Lower α means less tolerance for type I error, and forces the critical value to become larger. Hence, for a fixed value of the test statistic, the result of the test may switch from reject to don't reject. A natural question then is: What is the smallest

2. The test we've implemented can be modified to tackle the case of dependent data. See Negri and Nishiyama (2010) for one such test and more references.

value of α at which we can still reject a given test statistic? This value is called the p -value.

To state this more precisely, consider again the setting of §10.2. The null hypothesis is $H_0 : \theta \in \Theta_0$ and, for each $\alpha \in (0, 1)$, a test $(T, c(\alpha))$ of size α . We assume here that $c(\alpha)$ is determined via the relationship in (10.18). In this setting, the p -value of the test is defined as

$$p(\mathbf{z}_{\mathcal{D}}) := \inf\{\alpha \in (0, 1) : c(\alpha) < T(\mathbf{z}_{\mathcal{D}})\}$$

Roughly speaking, this is the α at which the test switches from reject to don't reject. Typically, $\alpha \mapsto c(\alpha)$ is continuous, and in this case the expression for $p(\mathbf{z}_{\mathcal{D}})$ reduces to

$$p(\mathbf{z}_{\mathcal{D}}) := \text{the } \alpha \text{ such that } c(\alpha) = T(\mathbf{z}_{\mathcal{D}}) \quad (10.23)$$

Example 10.2.7 Recall the test (10.21). Here $c(\alpha) := \Phi^{-1}(1 - \alpha/2)$, and $c(\alpha)$ is continuous in α , so we can apply the definition of $p(\mathbf{z}_{\mathcal{D}})$ in (10.23). To solve for $p(\mathbf{z}_{\mathcal{D}})$, then, we need to solve for α in the expression $\Phi^{-1}(1 - \alpha/2) = |t_N(\mathbf{z}_{\mathcal{D}})|$. With a bit of rearranging and an application of symmetry, we obtain

$$p(\mathbf{z}_{\mathcal{D}}) = 2\Phi(-|t_N(\mathbf{z}_{\mathcal{D}})|) \quad (10.24)$$

10.2.4 Accepting the Null?

Sometimes failure to reject a null hypothesis comes to be seen as presenting evidence in favor of that hypothesis. This is a slippery slope. Here's one illustration.

In the 1980s econometricians found a new set of toys called unit root tests. First it was discovered that a newly developed test due to Dickey and Fuller (1979) did not reject the null hypothesis of a unit root when applied to a variety of economic time series (Nelson and Plosser 1982). It was argued that this finding called into question a range of earlier time series methods that assumed trend stationarity, as well as a variety of business cycle studies that assumed stationary innovations around a trend. A raft of new unit root tests appeared. The idea that innovations to major economic time series contain a permanent component became something like a stylized fact.

Let's think about this carefully. For the sake of concreteness, suppose that we are considering unemployment rate data, where the unit root hypothesis was associated with the concept of hysteresis (see, e.g., Mitchell 1993). In this setting the unit root null can be expressed most simply as the hypothesis that $a = 1$ in the AR(1) process

$$u_{t+1} = au_t + b + \epsilon_{t+1} \quad (10.25)$$

Here $\{u_t\}$ is unemployment, a and b are parameters and $\{\epsilon_t\}$ is a zero-mean innovation (see eq. (5) of Mitchell 1993). Letting T be the length of the sample, the standard

one-sided Dickey–Fuller test of the unit root null is to

- (i) estimate the parameters as $\hat{a}$ and $\hat{b}$ by least squares (see page 224) and
- (ii) reject the null if $T(\hat{a} - 1) < c$, where c is a critical value.

Intuitively, when $a < 1$, the test statistic will evaluate to a negative number in large samples. While we omit the derivation, Dickey and Fuller (1979) showed that the test statistic converges in distribution under the null and tabulated critical values for different test sizes.

A number of studies have been unable to reject a unit root null for unemployment rate data (see Mitchell 1993 for an earlier study or Cheng et al. 2012 for a study using a newer set of tests). How should we interpret failure to reject the null in this setting? One possibility is that (10.25) is a good model for the data with $a = 1$. This is not the only possibility, however. Another possibility is that the data are being generated by some other process against which this test has little power.

The second possibility seems more plausible than the first. Indeed, if $a = 1$ in (10.25), then $\mathbb{P}\{u_t \in B\} \rightarrow 0$ as $t \rightarrow \infty$ for every bounded $\mathcal{B}$ -measurable $B \subset \mathbb{R}$. But employment rates don't diverge. A more likely story is that while unemployment rates are highly persistent for some range of values, there are equilibrium forces that become increasingly stronger outside this band. For example, the further the unemployment rate falls, the more tightness in the labor market boosts wages, encouraging employers to substitute toward other factors of production and putting upward pressure on the unemployment rate.

We can easily reconcile these equilibrium forces with the idea that the unemployment rate is highly persistent at commonly observed values by adopting a nonlinear model. For example, consider the following simple variation on a random walk:

$$u_{t+1} = h(u_t) + \epsilon_{t+1} \quad (10.26)$$

where h is the generalized logistic function depicted in figure 10.7, along with the 45 degree line. For unemployment rates in a band between about 5 and 15, the function has a slope close to one, and the process exhibits strong persistence, just like a random walk. For values above and below this number the shape of h causes drift back to the band.

Listing 7 gives code for a simulation when this model generates the data. The shocks $\{\epsilon_t\}$ are independent Gaussian. In each loop a time series of length 100 is drawn and the Dickey–Fuller test statistic is calculated. The size of the test is set to 0.05, in which case the critical value for the test is -13.7 (see, e.g., Hayashi 2000, p. 576). If you run the code you will see that for this random seed, the rejection frequency over 5,000 repetitions is around 0.05. This is the same rejection frequency as when the null is true. For this data size, the great majority of tests would fail to reject

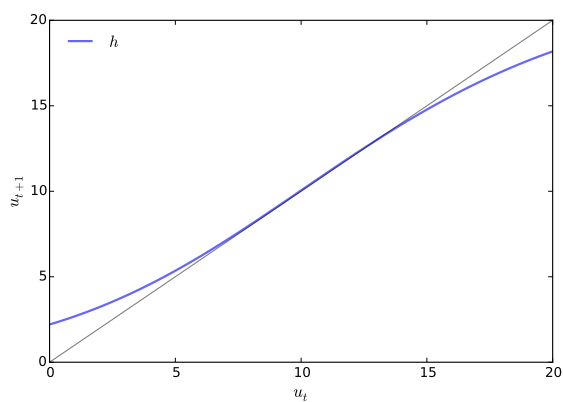


Figure 10.7 A nonlinear process for unemployment dynamics

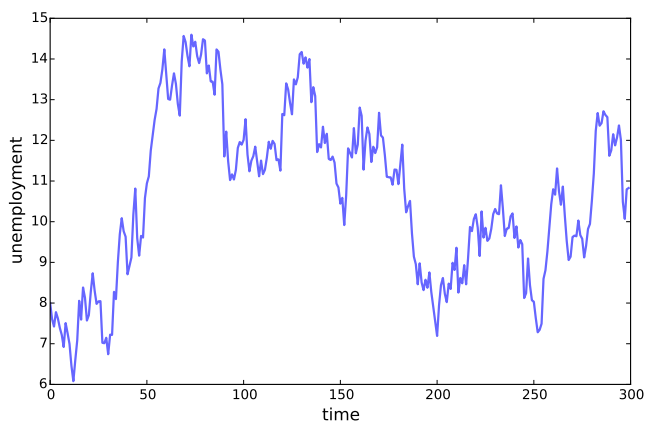


Figure 10.8 Simulated unemployment dynamics

the unit root model, even though the data-generating process is (10.26).

Listing 7 DF test applied to a nonlinear time series (Julia)

```

N = 100           # Length of time series
alpha = 0.05      # Size of test
critical = -13.7   # Critical value for DF test of size 0.05
srand(123)        # Set random seed

function h(x, a=0.0, k=22, b=.16, q=0.6, m=10)
    return a + (k - a) / ((1 + q * exp(-b * (x - m)))^(1/q))
end

function dickey_fuller_stat(y)
    y1 = y[1:end-1]
    y2 = y[2:end]
    z2 = y2 - mean(y2)
    z1 = y1 - mean(y1)
    rho = dot(z1, z2) / dot(z1, z1)
    return length(y1) * (rho - 1)
end

function sim_dgp(N)
    y = Array{Float64, N}
    y[1] = 8 # Initial value of unemployment rate
    for i in 1:N-1
        y[i+1] = h(y[i]) + 0.5 * randn()
    end
    return y
end

M = 5000
rejections = 0
for m in 1:M
    y = sim_dgp(N)
    rejections += (dickey_fuller_stat(y) < critical)
end
println("Fraction under critical value: ", rejections / M)

```

The interesting thing about this experiment is that the nonlinear data-generating process has very different properties from the null. Not only is it stationary, it is uni-

formly ergodic, as can be seen by checking the conditions of theorem 16.0.2 of Meyn and Tweedie (2009). This is the strongest possible form of stationarity after the IID property. It means that the effect of initial conditions dies out at a geometric rate that is bounded away from one uniformly over the state space. There is no “permanent component.”

10.2.5 Statistical Tests in Economics

While the mechanics of hypothesis testing are relatively straightforward, the scientific paradigm based around hypothesis testing is the subject of continual debate. Macroeconomics is an interesting case in point. In the 1970s the rational expectations revolution ushered in a new class of macroeconomic models. One of the first things the proponents of rational expectations did was to test these models against data. Thomas Sargent’s account runs as follows (Evans and Honkapohja 2005):

My recollection is that Bob Lucas and Ed Prescott were initially very enthusiastic about rational expectations econometrics. After all, it simply involved imposing on ourselves the same high standards we had criticized the Keynesians for failing to live up to. But after about five years of doing likelihood ratio tests on rational expectations models, I recall Bob Lucas and Ed Prescott telling me that those tests were rejecting too many good models.

In the following years, many macroeconomists moved away from formal hypothesis testing. A standard response to classical testing was: “I already know that my model is wrong. Rejection of my model using standard inference tells me nothing new.” For example, Kydland and Prescott (1991) define econometrics as a process without tests:

The main steps in econometric analysis are as follows: defining the question; setting up the model; calibrating the model; and reporting the findings.

How does this line of thinking compare with the standard statistical paradigm? The standard paradigm certainly recognizes that all models are wrong. All sensible models are wrong along some dimensions because they are, by definition, simplified representations of reality. Rejection of models is intended to facilitate survival of the fittest: Those models that explain observed phenomena effectively, and, at the same time, are most difficult to reject, will survive.³

3. Incidentally, given this description of model selection, it might appear there is an easy way to produce

At the same time, it's reasonable to feel sympathy for the frustration with classical hypothesis testing expressed by some economists. While the hypothesis that water boils at 100 degrees C will never be refuted (at least at sea level), in economics even the best models capture only some phenomena of interest. Inevitably they fail to match with reality when viewed in other ways. With enough data almost all sensible economic models can be rejected. But should we discard them? Perhaps some of these models are still useful in guiding our thinking.

Example 10.2.8 The consumption-based asset pricing model of Lucas (1978) fails to provide a good fit for returns on most classes of assets. But the key pricing equation still provides an important benchmark, as well as a way for econometricians to organize their thoughts on asset prices in a quantitative manner.

How to resolve these different points of view? Some researchers argue that calibration and matching of moments are sufficient as a means of validating their models. But this form of validation imposes little discipline on model builders. Clever economists can find multiple ways to replicate most interesting phenomena observed in the data, each with their own distinct policy implications. More important, the objective of matching moments is largely an in-sample criterion. The dangers of such an approach were illustrated in figure 8.4 on page 230.

One middle ground is testing of partially specified models, where tests are only concerned with a specific restriction implied by theory. A second path forward is more emphasis on out-of-sample performance in economic studies, which provides a much stricter test of true understanding. A third is offered by the Bayesian approach to model selection, where the emphasis is on comparing and improving models rather than rejection. A fourth is improvements in economic theory, to better match our models to the data.

10.3 Further Reading

For a technical discussion of the results surrounding the convergence to the Kolmogorov distribution in (10.12), see section 19.4 of Van der Vaart (2000).

For a discussions of statistical testing in econometrics see, for example, McCloskey and Ziliak (1996) and Freedman (2009). Discussions of calibration in relation to estimation and hypothesis testing can be found in Kydland and Prescott (1991), Kim and Pagan (1999), and many other places. Hansen (2014b) presents the main arguments

a model that survives forever in the pool of currently acceptable theories, without ever being culled via rejection: just make sure that the model has no testable implications. However, this strategy does not work because a model with no testable implications cannot be considered as a theory of anything. The *definition* of a scientific theory, as proposed by Karl Popper, is that the theory has one or more testable implications. In other words, the theory can be *falsified*.

in favor of tests of restrictions in partially specified models. Silver (2012) provides an interesting read on estimation and uncertainty with applications in forecasting.

An excellent summary of unit root tests is given in Hayashi (2000). For a critical overview of unit root testing, see Cochrane (1991).

10.4 Exercises

Ex. 10.4.1 Confirm fact 10.1.1 on page 277.

Ex. 10.4.2 Show that (10.8) is valid when (10.6) and (10.7) hold.

Ex. 10.4.3 Let $\alpha \in (0, 1)$ and set $z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2)$. Show that if $t_N(\mathbf{x}) \xrightarrow{d} N(0, 1)$ whenever H_0 is true, and $T_N(\mathbf{x}) = |t_N(\mathbf{x})|$, the sequence of tests $\phi_N(\mathbf{x}) := \mathbb{1}\{T_N(\mathbf{x}) > z_{\alpha/2}\}$ is asymptotically of size α .

Ex. 10.4.4 If a test is well-designed, then the rejection probability will converge to one as $N \rightarrow \infty$ whenever H_0 is false:

$$\beta_N(\theta) \rightarrow 1 \text{ as } N \rightarrow \infty \quad \text{whenever} \quad \theta \notin \Theta_0$$

Such a test is said to be **consistent**. Show that the test in example 10.2.6 is consistent whenever $\alpha \in (0, 1)$.⁴

Ex. 10.4.5 (Computational) The chi-squared goodness of fit test is used to test whether or not a given data set was generated from a particular discrete distribution, described by a PMF $p_1, \dots, p_J$ over values $1, \dots, J$. More precisely, let $x_1, \dots, x_N$ be N random variables, each x_n taking integer values between 1 and J . The null hypothesis of the test is that the sample is IID with $\mathbb{P}\{x_n = j\} = p_j$ for $n \in \{1, \dots, N\}$ and $j \in \{1, \dots, J\}$. The test statistic is given by

$$X := N \sum_{j=1}^J \frac{(q_j - p_j)^2}{p_j} \tag{10.27}$$

where q_j is the fraction of the sample $x_1, \dots, x_N$ taking the value j . Write a function called `chsqts` that takes arguments `observations` and `p`, where `observations` is a vector storing the sample $x_1, \dots, x_N$ and `p` is a vector storing the values $p_1, \dots, p_J$. The function call `chsqts(observations, p)` should return the value X in equation (10.27).⁵

4. Hint: In essence, you need to show that the absolute value of the test statistic $|t_N|$ in (10.20) diverges to infinity when $\theta \neq \theta_0$. Try the add and subtract strategy, replacing the expression $\bar{x}_N - \theta_0$ in (10.20) with $(\bar{x}_N - \theta) + (\theta - \theta_0)$.

5. R has a built-in function called `chisq.test` that you can use to test your results.

Ex. 10.4.6 (Computational) This exercise continues on from ex. 10.4.5. Under the null hypothesis, X in (10.27) is asymptotically chi-squared with $J - 1$ degrees of freedom. Let $J = 3$ with $p_1 = 0.2$, $p_2 = 0.2$ and $p_3 = 0.6$. Let $N = 20$. By repeatedly simulating 20 IID observations $x_1, \dots, x_{20}$ from this PMF,⁶ generate 5,000 independent observations of the statistic X , and store them in a vector called `obsX`. Plot the ECDF of `obsX`. In the same figure, plot the chi-squared CDF with 2 degrees of freedom. The functions should be close. (You can improve the fit further by taking N larger.)

10.4.1 Solutions to Selected Exercises

Solution to Ex. 10.4.1. Assume the conditions of fact 10.1.1. Let s_N^2 be the sample variance. Combining fact 5.1.16 on page 140, example 9.1.4 on page 249, fact 9.1.1 on page 249 and equation (10.2), the distribution of

$$\sqrt{N} \frac{(\bar{x}_N - \mu)}{\sigma} \sqrt{\frac{N-1}{Ns_N^2/\sigma^2}}$$

has a t distribution with $N - 1$ degrees of freedom. Some algebra gives (10.3). $\square$

Solution to Ex. 10.4.2. We aim to show that $(\hat{\theta}_N - \theta)/\text{se}(\hat{\theta}_N)$ converges in distribution to a standard normal when $\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, v(\theta))$ and $\sqrt{N} \text{se}(\hat{\theta}_N) \xrightarrow{p} \sqrt{v(\theta)}$. To see this, observe that

$$\frac{\hat{\theta}_N - \theta}{\text{se}(\hat{\theta}_N)} = \zeta_N \eta_N \quad \text{where} \quad \zeta_N := \frac{\sqrt{N}(\hat{\theta}_N - \theta)}{\sqrt{v(\theta)}} \quad \text{and} \quad \eta_N := \frac{\sqrt{v(\theta)}}{\sqrt{N} \text{se}(\hat{\theta}_N)}$$

Using the various rules for convergence in probability and distribution (check them carefully—see facts 6.1.1 and 6.1.6) we obtain $\zeta_N \xrightarrow{d} Z \sim N(0, 1)$ and $\eta_N \xrightarrow{p} 1$. Applying Slutsky's theorem (page 168) now gives $\zeta_N \eta_N \xrightarrow{d} Z$, as was to be shown. $\square$

Solution to Ex. 10.4.3. Fix $\alpha \in (0, 1)$, and let $z \sim N(0, 1)$. In view of (4.36) on page 115, we have $\mathbb{P}\{|z| > z_{\alpha/2}\} = \alpha$. If H_0 is true, then $t_N(\mathbf{x}) \xrightarrow{d} z$ by assumption. Since $g(s) := |s|$ is continuous, fact 6.1.6 on page 167 implies that $|t_N(\mathbf{x})| \xrightarrow{d} |z|$. As a result we have

$$\lim_{N \rightarrow \infty} \beta_N(\theta) = \lim_{N \rightarrow \infty} \mathbb{P}_\theta\{|t_N(\mathbf{x})| > z_{\alpha/2}\} = \mathbb{P}\{|z| > z_{\alpha/2}\} = \alpha$$

This confirms (10.19), and the exercise is done. $\square$

6. In particular, draw each x_n such that $\mathbb{P}\{x_n = j\} = p_j$ for $j = 1, 2, 3$.