

These notes come from Chapters 13 and 14 of the textbook and provide an introduction to inference and hypothesis testing. This section will rely heavily on our results concerning asymptotic normality from the previous set of notes.

## Hypothesis Testing Notes

Often in empirical applications in economics/business/social sciences, we have some theory (or “hypothesis”) that we would like to test. For example, a theory might be that some economic policy has no effect on an outcome of interest.

Because we only have access to a sample rather than the full population, even if the theory is correct (e.g., no effect), then we still will not generally estimate, say,  $\hat{\theta} = 0$  (i.e., that our estimate of  $\theta$  is exactly equal to 0).

The idea of hypothesis testing is to try answer the question: Are the observed estimates compatible with the theory in the sense that the difference of the estimate relative to the theory can be explained by stochastic variation (i.e., that we collected a sample); or, alternatively, are the estimates incompatible with the theory in the sense that the estimate would be highly unlikely if the hypothesis were true?

### Hypotheses

PSE: 13.2

The **null hypothesis**  $\mathbb{H}_0$  is the restriction that  $\theta = \theta_0$  where  $\theta_0$  is a hypothesized value (and, therefore, known) value. A leading case is  $\theta_0 = 0$ .

The **alternative hypothesis**  $\mathbb{H}_1$  is the set  $\{\theta \in \Theta | \theta \neq \theta_0\}$  where  $\Theta$  is the set of possible values of  $\theta$  (aka the parameter space).

In some cases, it might make sense to consider a one-sided alternative hypothesis, that is,  $\mathbb{H}_1 : \theta > \theta_0$ , but because it is the most common case, I’ll stick to the case with the two sided alternative above.

### Acceptance and Rejection

PSE: 13.3

A hypothesis test either rejects the null hypothesis or fails to reject the null hypothesis. This amounts to making a decision based on the available data. The most common version of this is to construct a **test statistic** that is a function of the data; that is,

$$T = T(X_1, \dots, X_n)$$

and then to compare  $T$  to a **critical value**  $c$  and to use the decision rule: reject  $\mathbb{H}_0$  if  $T > c$ ; otherwise, fail to reject.

## Type I Error

PSE: 13.4

**Type I Error** means to reject  $\mathbb{H}_0$  when  $\mathbb{H}_0$  is true.

The probability of a Type I Error is called the **size** of a test. That is,

$$P(\text{Reject } \mathbb{H}_0 | \mathbb{H}_0 \text{ true}) = P(T > c | \mathbb{H}_0 \text{ true})$$

## Type II Error and Power

PSE: 13.4

**Type II Error** means to fail to reject  $\mathbb{H}_0$  when  $\mathbb{H}_1$  is true. The rejection probability under the alternative hypothesis is called the **power** of the test. The power of the test is equal to 1 minus the probability of a Type II error. It is given by

$$\pi(\theta) = P(\text{reject } \mathbb{H}_0 | \mathbb{H}_1 \text{ true}) = P(T > c | \mathbb{H}_1)$$

where we write this as a function of  $\theta$  to indicate that it depends on the true value of the parameter  $\theta$ .

The most common approach to hypothesis testing is to pre-select a **significance level**  $\alpha$  (e.g.,  $\alpha = 0.05$  or  $\alpha = 0.01$ ) and then to select the critical value  $c$  so that the (asymptotic) size of the test is no larger than  $\alpha$ . Subject to this constraint, the goal is then to have high power. The power of a test depends on the true value  $\theta$  (generally, power is higher for values of  $\theta$  further away from  $\theta_0$ ) and on the sample size  $n$  (power increases with larger sample sizes) – see simulations at the end of these notes for an example of this.

Furthermore, notice that there is a tradeoff between committing Type I and Type II errors. For example, increasing the critical value will decrease the likelihood of a Type I error (thus, decreasing the size of the test) but it also increases the likelihood of a Type II error (thus, decreasing the power of the test).

Table 9.1: Hypothesis Testing Decisions

	Accept $\mathbb{H}_0$	Reject $\mathbb{H}_0$
$\mathbb{H}_0$ true	Correct Decision	Type I Error
$\mathbb{H}_1$ true	Type II Error	Correct Decision

## Asymptotic Standard Errors

PSE: 13.9

Next, let us return to our results on asymptotic normality of  $\hat{\theta}$  and see how this is useful for hypothesis testing. In particular, suppose that we have (somehow) established that  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d}$

$N(0, \sigma^2)$ , for example, using techniques such as the central limit theorem, continuous mapping theorem, or delta method. Recall that this is equivalent to

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\sigma^2}} \xrightarrow{d} N(0, 1)$$

which we use below. Asymptotic normality will be the basis for the approach to inference/hypothesis testing below, and is the whole reason that we invested so much effort into learning about it in the previous set of notes.

We define the **standard error** of the  $\hat{\theta}$  to be

$$\text{se}(\hat{\theta}) = \frac{\sqrt{\hat{\sigma}^2}}{\sqrt{n}}$$

where  $\hat{\sigma}^2$  is an estimator of  $\sigma^2$ . We are operating in a quite generic framework here, so the precise expression for  $\hat{\sigma}^2$  depends on what  $\sigma^2$  is, and we discussed how to estimate  $\sigma^2$  in a variety of scenarios at the end of the last set of notes. The only requirement for us is that  $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$ . Just to repeat one concrete example though, if  $\theta = E[g(x)]$ , then, it is natural to use  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (g(X_i) - \hat{\theta})^2$ .

This also immediately extends to the case where  $\theta$  is a  $k \times 1$  vector. In that case, supposing that we know that  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V})$  and where  $\mathbf{V}$  is a  $k \times k$  matrix, the standard error of  $\hat{\theta}_j$  (the  $j$ th element of  $\hat{\theta}$ ) as

$$\text{se}(\hat{\theta}_j) = \frac{\sqrt{\hat{\mathbf{V}}_{jj}}}{\sqrt{n}}$$

where  $\hat{\mathbf{V}}$  is a consistent estimator of  $\mathbf{V}$ , and  $\hat{\mathbf{V}}_{jj}$  is the  $(j, j)$  element of  $\hat{\mathbf{V}}$  (which is an element along the diagonal of  $\hat{\mathbf{V}}$ ). As above, we are operating quite generally here, and an expression for  $\hat{\mathbf{V}}$  depends on the particular expression for  $\mathbf{V}$  (that you would know from deriving that  $\hat{\theta}$  is asymptotically normal). Here is concrete example though: if  $\theta = E[g(X)]$ , then a natural way to estimate the variance is  $\hat{\mathbf{V}} = \frac{1}{n} \sum_{i=1}^n (g(X_i) - \hat{\theta})(g(X_i) - \hat{\theta})'$  which is a  $k \times k$  matrix.

The standard error of an estimator is a key ingredient for the inference procedures discussed below. It is also very common in applications in economics to just report  $\hat{\theta}$  (i.e., your estimator) and  $\text{se}(\hat{\theta})$  (i.e., its standard error).

### t-statistic

PSE: 13.9

When  $\theta$  is a scalar and under  $\mathbb{H}_0 : \theta = \theta_0$ , the **t-statistic** is given by

$$t = \frac{\hat{\theta} - \theta_0}{\text{s.e.}(\hat{\theta})}$$

The t-statistic is the most common type of test-statistic. Notice that  $t$  is a number that we can calculate given the data that we have. To understand its usefulness (and the arguments below), it is useful to provide an alternative expression for  $t$ :

$$\begin{aligned} t &= \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{\sigma}^2}} \\ &= \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\sigma^2}} + \underbrace{\left( \frac{1}{\sqrt{\hat{\sigma}^2}} - \frac{1}{\sqrt{\sigma^2}} \right)}_{=o_p(1)} \underbrace{\sqrt{n}(\hat{\theta} - \theta_0)}_{O_p(1)} \\ &= \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\sigma^2}} + o_p(1) \end{aligned}$$

where the first equality holds by the definition of s.e.  $(\hat{\theta})$ , the second equality holds by adding and subtracting  $\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\sigma^2}}$  (and then re-arranging), and the third equality holds from the previous one because  $\hat{\sigma}^2$  is consistent for  $\sigma^2$  and by the continuous mapping theorem.

The basis for our hypothesis testing procedure below will be that  $t$  behaves differently depending on whether or not  $\mathbb{H}_0$  is true. Consider the two following cases:

- **Case 1 ( $\mathbb{H}_0$  true):** If  $\mathbb{H}_0$  is true, then, in the expression for  $t$  above, we can replace  $\theta_0$  (from our theory) with  $\theta$  (the actual population parameter); that is,

$$t = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\sigma^2}} + o_p(1) = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\sigma^2}} + o_p(1) \xrightarrow{d} N(0, 1)$$

In other words (given a large sample), if the  $\mathbb{H}_0$  is true, then  $t$  should behave like a draw from a standard normal distribution.

- **Case 2 ( $\mathbb{H}_0$  false):** Now consider the case where  $\mathbb{H}_0$  is false. In this case  $\hat{\theta} - \theta_0$  does not converge to 0; rather,  $(\hat{\theta} - \theta) \xrightarrow{p} (\theta - \theta_0) \neq 0$ . Thus,

$$\begin{aligned} t &= \frac{\sqrt{n}((\hat{\theta} - \theta_0))}{\sqrt{\sigma^2}} + o_p(1) = \frac{\sqrt{n}(\hat{\theta} - \theta) + (\theta - \theta_0)}{\sqrt{\sigma^2}} + o_p(1) \\ &= \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\sigma^2}} + \frac{\sqrt{n}(\theta - \theta_0)}{\sigma^2} + o_p(1) \end{aligned}$$

which holds just by adding and subtracting  $\theta$  and re-arranging. The first term converges to  $N(0, 1)$  as in the previous case. But the second term is  $\sqrt{n}$  multiplied by a non-zero constant  $(\theta - \theta_0)/\sigma^2$ . This term will, therefore, diverge as  $n \rightarrow \infty$ . It is the dominant term here and will cause  $t$  itself to diverge. [Note that  $t$  can go to positive or negative infinity depending on the sign of  $\theta - \theta_0$ , which basically amounts to whether or not the “theory” from  $\mathbb{H}_0$  is above or below the the true value of the population parameter  $\theta$ .]

The above discussion suggests notably different behavior of  $t$  depending on whether or not

$\mathbb{H}_0$  holds — this will be the basis for our inference procedure. In particular, our idea will be the following: if, when we calculate  $t$  using the available data,  $t$  “looks like” a draw from a normal distribution, then we will not reject  $\mathbb{H}_0$ . On the other hand, if the magnitude of  $t$  is large, so that it appears “unlikely” that  $t$  is a draw from a normal distribution, we will take that as strong evidence against  $\mathbb{H}_0$  and reject it.

All that’s left is to formalize what “looks like” means. It is typical to use the decision rule: Reject  $\mathbb{H}_0$  if  $|t| > c$  where  $c$  is a critical value. As we discussed earlier, the critical value depends on the significance level (which we have chosen *ex ante*) and denoted by  $\alpha$ . The significance level is the probability with which we’d be willing to reject  $\mathbb{H}_0$  when it is true (given that we have a large sample). Now let’s discuss how to choose a critical value  $c$  so that, in large samples, we reject  $\mathbb{H}_0$  with probability  $\alpha$  when  $\mathbb{H}_0$  is true. In particular, we would like  $c$  to satisfy

$$\alpha = \lim_{n \rightarrow \infty} P(|t| > c | \mathbb{H}_0)$$

which is just what we said before — we are aiming to, asymptotically, commit a Type I error  $\alpha$  percent of the time, and we will next solve this equation for  $c$  as a function of  $\alpha$ .

$$\begin{aligned} \alpha &= \lim_{n \rightarrow \infty} P(|t| > c | \mathbb{H}_0) \\ &= P(|Z| > c) \\ &= P(Z < -c) + P(Z > c) \\ &= 2P(Z > c) \\ &= 2(1 - \Phi(c)) \\ \implies c &= \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \end{aligned}$$

where the second equality holds because, if  $\mathbb{H}_0$  is true, then  $t \xrightarrow{d} Z \sim N(0, 1)$  and by the extended continuous mapping theorem (to deal with the absolute value part), the third equality holds immediately, the fourth equality holds because a normal distribution is symmetric, the fifth holds by properties of cdfs (and where  $\Phi$  is the cdf of a standard normal random variable), and the last equality is just algebra.

This suggests a way to choose a critical value, given a value of  $\alpha$ . For example, when  $\alpha = 0.05$ , you would compute  $c = \Phi^{-1}(1 - .05/2) = \Phi^{-1}(0.975)$ . You can look up the quantiles of a standard normal random variable in a table or use R to do it using the `qnorm` function. For example: `c <- qnorm(.975) = 1.96`, which is probably a familiar critical value.

Finally, notice that if you make different choices of  $\alpha$ , then this will lead to different critical values. For example, if you set  $\alpha = 0.01$  (which indicates that you have less tolerance for committing Type I errors), then the critical value would be `qnorm(1 - .01/2) = 2.58`. It is also worth mentioning that it makes sense that you get a larger critical value when  $\alpha = 0.01$  than before. In this case, we are less willing to make Type I errors. The way to avoid making as many Type I errors is to reject

$\mathbb{H}_0$  less, and you will reject less if you use a larger critical value. The “cost” of the larger critical value is that the test will have lower power.

### p-value

PSE: 13.15

The approach to inference discussed so far has been to compute a t-statistic and then to make a binary decision to either reject or fail to reject  $\mathbb{H}_0$ . This approach has some inherent issues. The textbook gives the example of a t-statistic equal to 1.7 relative to one that is equal to 2.0. Given a 5% significance level, these t-statistics lead to different decisions. However, it is immediately clear that the strength of evidence against  $\mathbb{H}_0$  is not really much different between these two cases.

An alternative approach is to report an (asymptotic) p-value. The p-value is the probability of getting a test-statistic as large (in absolute value) as we did given that  $\mathbb{H}_0$  is true (an alternative interpretation is that  $p$  is the smallest value of  $\alpha$  for which the test would reject  $\mathbb{H}_0$ ). A p-value can be calculated as

$$p = 2\Phi(-|\tilde{t}|)$$

where  $\Phi$  is the cdf of a standard normal random variable and  $\tilde{t}$  is the t-statistic that we calculated using the data that we have (as a side-comment, below we need to distinguish between the estimator  $t$ , which has a sampling distribution, and  $\tilde{t}$  which is the particular value of the t-statistic that we calculated). To understand where this comes from, notice that

$$\begin{aligned} p &= \lim_{n \rightarrow \infty} \mathbb{P}(|t| > -|\tilde{t}| \mid \mathbb{H}_0) \\ &= \mathbb{P}(|Z| > -|\tilde{t}|) \\ &= \mathbb{P}(Z < -|\tilde{t}|) + \mathbb{P}(Z > -|\tilde{t}|) \\ &= 2\Phi(-|\tilde{t}|) \end{aligned}$$

where the first line comes from just writing in math how we described the p-value above – it is the (asymptotic) probability of getting a t-statistic as extreme as  $\tilde{t}$  if  $\mathbb{H}_0$  were true, the second equality holds because, under  $\mathbb{H}_0$ ,  $t \xrightarrow{d} Z \sim N(0, 1)$  and by the CMT, the third equality holds immediately, and the fourth equality holds by the symmetry of a standard normal distribution.

Unlike the binary decision rule that we have discussed previously, the p-value provides continuous information. For example, if we calculate that  $p = 0.06$ , we would not reject  $\mathbb{H}_0$  at the 5% significance level, but getting a t-statistic this large in absolute value is still relatively uncommon if  $\mathbb{H}_0$  is true. Similarly, if  $p = 0.00001$ , this would indicate very strong evidence against  $\mathbb{H}_0$ .

### Confidence intervals

PSE: 14.1-14.3, 14.5-14.6, 14.9

The previous discussion has been broadly about hypothesis testing, i.e., checking whether the data that you have is compatible with some theory about the population parameter being studied. An alternative approach, though closely related, is to provide an **interval estimator** for  $\theta$  (where “interval” is to distinguish from estimators like  $\hat{\theta}$  that can be referred to as a **point estimator** for  $\theta$ ; that is, it is a single value.). An interval estimator is typically of the form,  $\hat{C} = [\hat{L}, \hat{U}]$  (here:  $L$  stands for “lower” and  $U$  stands for “upper”), and where  $\hat{C}$  is called a **confidence interval**.

A confidence interval is useful for providing a measure of the accuracy of an estimator. And, for example, a main motivation for providing a confidence interval would be to provide a range of values that are “reasonably plausible” given the data that we have. We will construct a confidence interval below that exploits that we know that an estimator is asymptotically normal. The arguments will be broadly similar to the ones for hypothesis testing above.

One thing to point out is that, because  $\hat{C}$  will be a function of the data,  $\hat{C}$  itself is random. The **coverage probability** of  $\hat{C}$  is  $P(\theta \in \hat{C})$ . The randomness comes from  $\hat{C}$  because  $\theta$  is a fixed, though unknown, population parameter.  $\hat{C}$  is called a  $(1 - \alpha)\%$  confidence interval if, at least in large samples,  $P(\theta \in \hat{C}) = 1 - \alpha$ .

The most common version of a confidence interval is the one given by

$$\hat{C} = [\hat{\theta} - c_{1-\alpha/2}\text{s.e.}(\hat{\theta}), \hat{\theta} + c_{1-\alpha/2}\text{s.e.}(\hat{\theta})]$$

where, for example, if  $\alpha = 0.05$ , then  $c_{1-\alpha/2} = c_{.975} = 1.96$  (because 1.96 is the 97.5th percentile of a standard normal distribution). It is worth briefly considering where this confidence interval comes from. Notice that (below, I drop the  $(1 - \alpha/2)$  subscript on  $c$  to make the notation clearer, but  $c$  does depend on  $\alpha$ ):

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\theta \in \hat{C}) &= \lim_{n \rightarrow \infty} P(\hat{\theta} - c\text{s.e.}(\hat{\theta}) < \theta < \hat{\theta} + c\text{s.e.}(\hat{\theta})) \\ &= \lim_{n \rightarrow \infty} P\left(-c < \frac{\theta - \hat{\theta}}{\text{s.e.}(\hat{\theta})} < c\right) \\ &= \lim_{n \rightarrow \infty} P\left(c > \frac{\hat{\theta} - \theta}{\text{s.e.}(\hat{\theta})} > -c\right) \\ &= \lim_{n \rightarrow \infty} P\left(-c < \frac{\hat{\theta} - \theta}{\text{s.e.}(\hat{\theta})} < c\right) \\ &= P(-c < Z < c) \\ &= P(|Z| < c) \\ &= 1 - \alpha \end{aligned}$$

where the first line holds because (i) we are going to establish the asymptotic coverage probability and (ii) the definition of the confidence interval, the second line holds by subtracting  $\hat{\theta}$  from all terms and then dividing all terms by  $\text{s.e.}(\hat{\theta})$ , the third equality holds by multiplying all terms by  $-1$ , the fourth equality holds by just rearranging the inequalities inside the parentheses, the fifth

equality holds because

$$\frac{\hat{\theta} - \theta}{\text{s.e.}(\hat{\theta})} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{\sigma}^2}} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\sigma^2}} + o_p(1) \xrightarrow{d} Z \sim N(0, 1)$$

the sixth equality holds immediately, and the last equality holds by the definition of critical values (we constructed  $c_{1-\alpha/2}$  so that the probability that  $|Z| < c_{1-\alpha/2}$  is equal to  $1 - \alpha$ ).

## Statistical significance vs. economic significance

PSE: 13.14

My sense is that it is most common to report  $\hat{\theta}$  and  $\text{s.e.}(\hat{\theta})$  in applications in economics. I think it is very uncommon (and not good practice) to only report  $\hat{\theta}$  with a binary indicator of whether or not some  $\mathbb{H}_0$  is rejected (this used to be more common with statistical significance being indicated by “significance stars” — there is a lot of pushback lately against “significance stars”; in my view, they can be useful but should not be the only thing that is reported).

It is also common to report confidence intervals (though less common than the practice of just  $\hat{\theta}$  and its standard error). This is more useful in some particular cases. For example, it's fairly common that papers in the minimum wage literature that estimate “no effect” of the minimum wage on employment to report a confidence interval; the strength of this sort of argument is to say something along the lines of: “for any ‘reasonable’ estimate, we find economically small effects of the minimum wage on employment.” Another effective use of confidence intervals is, in cases where you want to report a lot of estimates (e.g., how treatment effects vary across different values of the covariates) to include a figure that includes point estimates and confidence intervals.

I think that it is less common (at least in economics) to report p-values or t-statistics. These contain essentially the same information as the combination of  $\hat{\theta}$  and its standard error anyway. That said, if you are sitting in a seminar and the researcher provides an estimate and its standard error, I think many people in the audience will (in their heads) be calculating  $\hat{\theta}/\text{s.e.}(\hat{\theta})$  and comparing it to 2 (i.e., close to the critical value 1.96).

More generally, particularly as datasets in economics tend to become larger over time, it is important to distinguish between **economic significance** and **statistical significance**. In particular, in most applications in economics, it is probably not reasonable to think that the effect of one variable on another is literally exactly equal to 0. Along these lines, it is certainly possible to estimate a statistically significant effect that is “economically small”. This is arguably self-evident — for example, suppose that you observed the entire population so that you could just directly calculate the population parameter of interest  $\theta$  rather than having to estimate  $\hat{\theta}$ ; it seems rather obvious that “how much”  $\theta$  is different from, say, 0 will matter a lot in many applications. This issue is still there even if you have to deal with additional complications due to having to estimate  $\theta$ .

The exact opposite issue is important too. Just because a statistical test does not reject  $\mathbb{H}_0$  does not imply that  $\mathbb{H}_0$  is true — rather it indicates that we cannot reject it given the data that we have.

The textbook gives an interesting related example of the “marriage premium” (i.e., the difference

between wages of married people and unmarried people) separately for men and women. The point estimate for men is 0.21 (p-value = 0.000, so strongly statistically significant), indicating that, on average, married men have 21% higher earnings than unmarried men, while for women the point estimate is 0.016 (p-value = 0.094 indicating marginally statistically significant). Supposing that you set the significance level to be  $\alpha = 0.1$ , then you would report that both marriage premiums are positive and statistically significant. However, this is probably not the best way to think about these result. Rather, a 95% confidence interval is given by  $[0.19, 0.23]$  for men and  $[0.00, 0.03]$ . This indicates a large difference in the marriage premium for men and women; and, in particular, the magnitudes of these differences are more informative than the results of the hypothesis testing.

## Wald Statistic

So far, we have mostly been focusing on the case where  $\theta$  is a scalar. Now suppose that  $\theta$  is a  $k \times 1$  vector. Further, suppose that we are interested in  $\mathbb{H}_0 : \theta = \theta_0$  (here:  $\theta_0$  is also a  $k \times 1$  vector). (Note the arguments below go through, essentially immediately, for the case where we are interested in conducting inference with respect to  $\beta = h(\theta)$ ). For this section, we suppose that we have already somehow established that  $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V})$  where  $\mathbf{V}$  is a  $k \times k$  variance matrix, and that we have a consistent estimator  $\hat{\mathbf{V}}$  of  $\mathbf{V}$ .

It is hard to operationalize the t-statistic that we talked about above – e.g., you could calculate a t-statistic for each element in  $\theta$ , but to conduct a test, you somehow need to boil that down into being a scalar test statistic. It is *possible* to make progress along those lines (e.g., by taking the maximum of the t-statistics), but these arguments turn out to be rather complicated. Instead, we will consider a Wald statistic:

$$W = n(\hat{\theta} - \theta_0)' \hat{\mathbf{V}}^{-1} (\hat{\theta} - \theta_0)$$

Notice that this is a number that we can compute (given a value of  $\theta_0$ ) and that it is a scalar. Using similar arguments as we did for the t-statistic above, it is helpful to notice that

$$\begin{aligned} W &= n(\hat{\theta} - \theta_0)' \mathbf{V}^{-1} (\hat{\theta} - \theta_0) + n(\hat{\theta} - \theta_0)' (\hat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}) (\hat{\theta} - \theta_0) \\ &= n(\hat{\theta} - \theta_0)' \mathbf{V}^{-1} (\hat{\theta} - \theta_0) + \underbrace{\sqrt{n}(\hat{\theta} - \theta_0)'}_{=O_p(1)} \underbrace{(\hat{\mathbf{V}}^{-1} - \mathbf{V}^{-1})}_{=o_p(1)} \underbrace{\sqrt{n}(\hat{\theta} - \theta_0)}_{=o_p(1)} \\ &= n(\hat{\theta} - \theta_0)' \mathbf{V}^{-1} (\hat{\theta} - \theta_0) + o_p(1) \end{aligned}$$

where the first equality holds by adding and subtracting  $n(\hat{\theta} - \theta_0)' \mathbf{V}^{-1} (\hat{\theta} - \theta_0)$ , and going from the second to the third line holds because  $\hat{\mathbf{V}}$  is consistent for  $\mathbf{V}$  and by the CMT.

Before proceeding, it is useful to recall the following result: If  $Z \sim N(0, \mathbf{I}_k)$ , then  $Z'Z \sim \chi_k^2$  (that is,  $Z'Z$  follows a chi-square distribution with  $k$  degrees of freedom).

As for  $t$  above, let's consider the behavior of  $W$  under  $\mathbb{H}_0$  and under  $\mathbb{H}_1 : \theta \neq \theta_0$ .

- **Case 1 ( $\mathbb{H}_0$  true):** If  $\mathbb{H}_0$  is true, then we can replace  $\theta_0$  in the expression for  $W$  with  $\theta$ . Then,

$$\begin{aligned} W &= \sqrt{n}(\hat{\theta} - \theta)' \mathbf{V}^{-1} \sqrt{n}(\hat{\theta} - \theta) + o_p(1) \\ &= \left( \mathbf{V}^{-1/2} \sqrt{n}(\hat{\theta} - \theta) \right)' \mathbf{V}^{-1/2} \sqrt{n}(\hat{\theta} - \theta) + o_p(1) \\ &\xrightarrow{d} Z'Z \sim \chi_k^2 \end{aligned}$$

where the first equality holds immediately, the second equality uses the square root matrix of  $\mathbf{V}^{-1}$  (which exists because  $\mathbf{V}$  is positive definite), and the last line holds because  $\mathbf{V}^{-1/2} \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} Z \sim N(0, \mathbf{I}_k)$ . All this to say, if  $\mathbb{H}_0$  is true, then  $W$  should behave like a draw from a  $\chi_k^2$  distribution.

- **Case 2 ( $\mathbb{H}_0$  false):** If  $\mathbb{H}_0$  is false, then  $\hat{\theta} - \theta_0 \xrightarrow{p} \theta - \theta_0 \neq 0$ . Thus, the terms  $\sqrt{n}(\hat{\theta} - \theta_0)$  both diverge which implies that  $W$  also diverges. s

As for the t-statistic, this different behavior under  $\mathbb{H}_0$  relative to  $\mathbb{H}_1$  provides an approach to inference. If  $W$  “looks like” a draw from a  $\chi_k^2$  distribution, then we should not reject  $\mathbb{H}_0$ , but, if not, then we will reject. The critical values here depend on the number of degrees of freedom; therefore, it is uncommon to memorize particular critical values. Rather, for testing multiple restrictions like this, I think that it is most common to report a p-value. Here, you can calculate a p-value by

$$p = 1 - G_k(W)$$

which is the probability of getting a test statistic as large as the  $W$  we actually calculated if  $\mathbb{H}_0$  were true. Here  $G_k$  is the cdf of a chi-square random variable with  $k$  degrees of freedom.

As mentioned above, because the distribution of  $W$  under  $\mathbb{H}_0$  depends on the degrees of freedom  $k$  (i.e., the number of restrictions being tested), it is harder to “just remember” critical values. That said, it is easy to compute the p-value above in R using the function `pchisq` (which computes the cdf of a chi-square random variable). For example, suppose that you calculate  $W = 7$  and that  $k = 2$ . Then, the p-value can be calculated as

```
p <- 1 - pchisq(7,df=2)
round(p,4)
```

```
## [1] 0.0302
```

so that the p-value is about 0.03 (indicating that you would reject  $\mathbb{H}_0$  at the 5% significance level).

## Monte Carlo Simulations

```

library(ggplot2)
library(dplyr)

set.seed(1234)

p <- 0.5 # prob of heads

# function to flip a coin with probability p
flip <- function(p) {
  sample(c(0,1), size=1, prob=c(1-p,p))
}

# function to generate a sample of size n
generate_sample <- function(n,p) {
  Y <- c()
  for (i in 1:n) {
    Y[i] <- flip(p)
  }
  Y
}

# function to carry out Monte Carlo simulations
# returns a vector of length nsims
# containing standardized versions of phat or bhat
# (depending on the argument `which_est`) from
# each simulation that come from, for example,
# sqrt(n)(phat-p)/sqrt(V)
mc2 <- function(n, H0=0.5, nsims=1000) {
  phat <- c() # vector to hold estimated p
  tstat <- c() # vector to hold estimated beta
  rej <- c() # vector to hold whether or not we reject H0
  for (i in 1:nsims) {
    Y <- generate_sample(n,p)
    phat[i] <- mean(Y)
    tstat[i] <- sqrt(n)*(phat[i]-H0)/sqrt(var(Y))
    rej[i] <- 1*(abs(tstat[i]) > qnorm(.975))
  }

  data.frame(phat=phat, tstat=tstat, rej=rej)
}

```

```

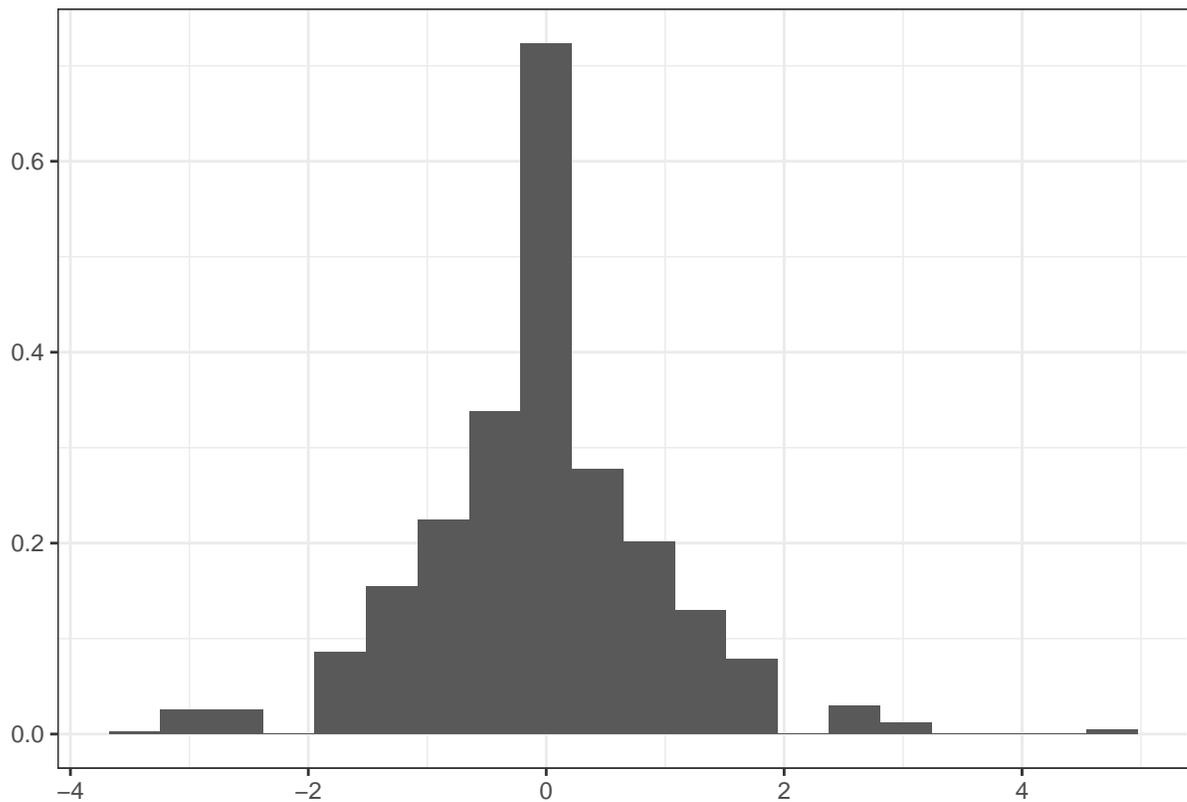
}

plot_t_results <- function(sim_results) {
  ggplot(sim_results, aes(x=tstat)) +
    geom_histogram(aes(y=..density..),
                  bins=20) +
    theme_bw() +
    ylab("") + xlab("")
}

# show results for phat for different values of n
# and p

# n=25
mc_res1 <- mc2(25)
plot_t_results(mc_res1)

```



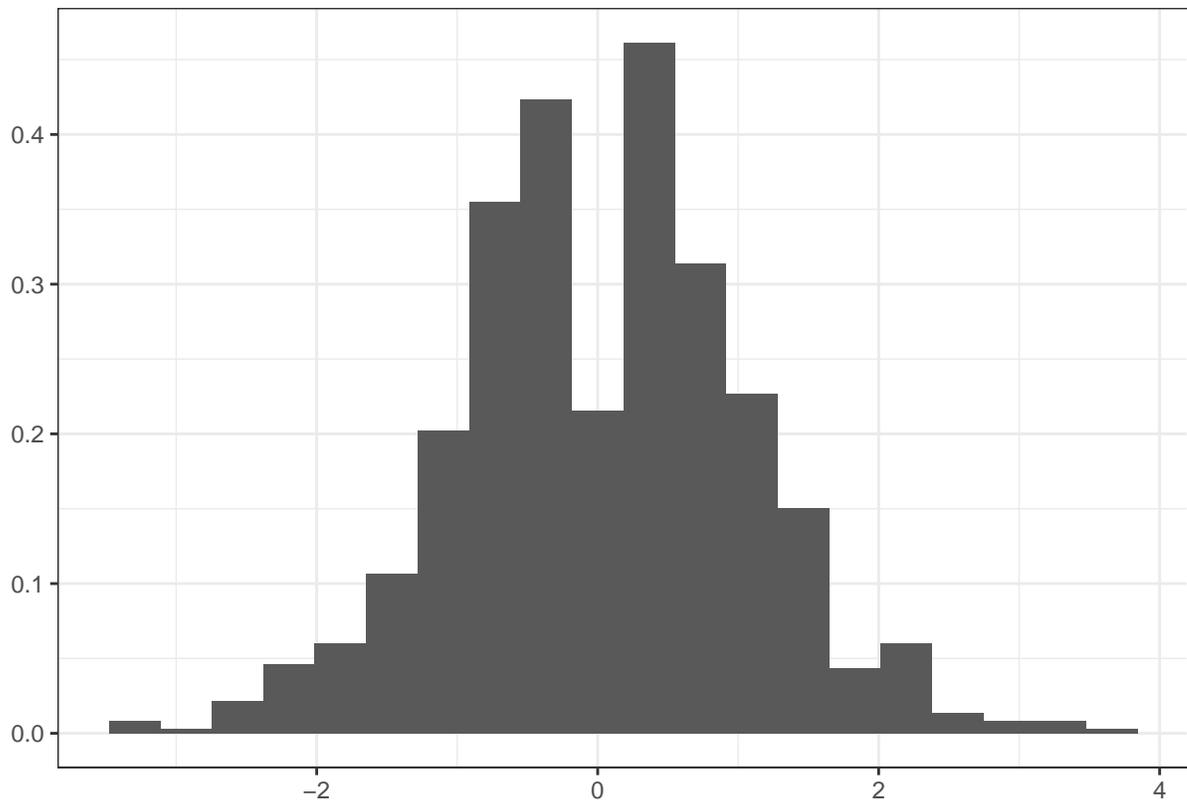
```
mean(mc_res1$rej)
```

```
## [1] 0.043
```

```
# n=100
```

```
mc_res2 <- mc2(100)
```

```
plot_t_results(mc_res2)
```



```
mean(mc_res2$rej)
```

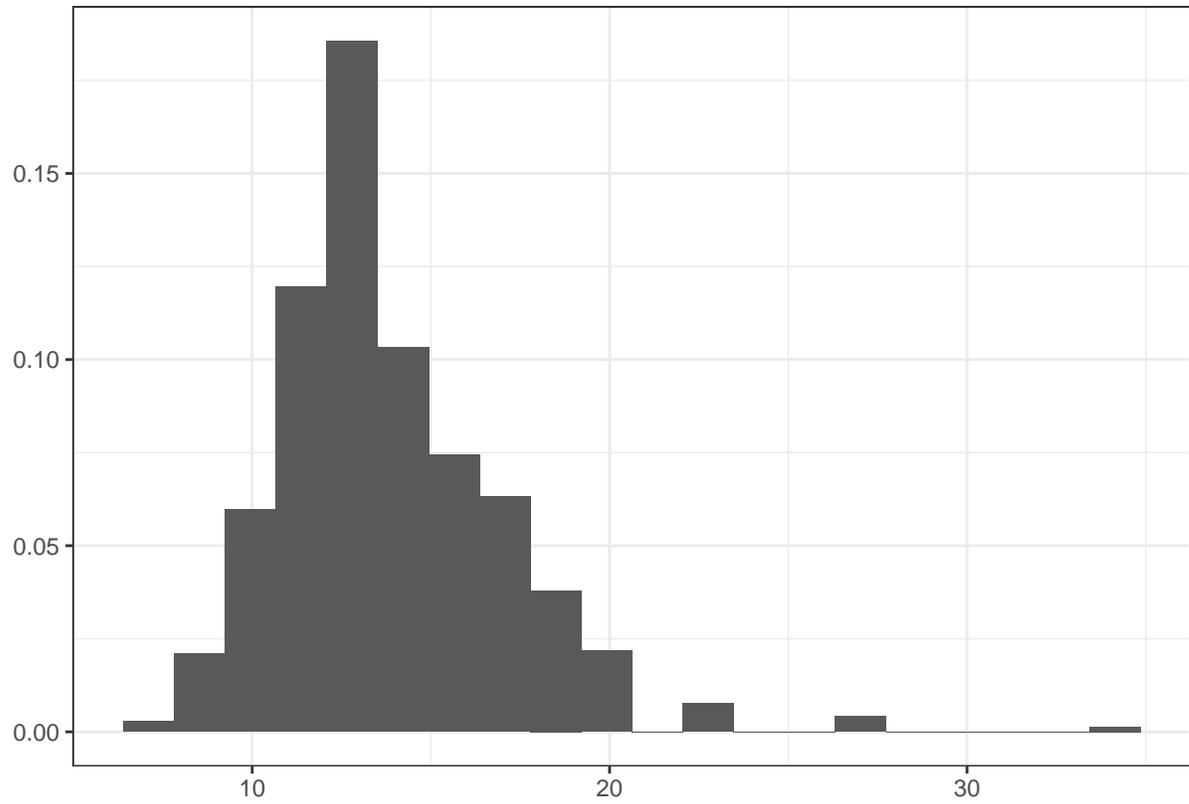
```
## [1] 0.063
```

```
# change p to be 0.9
```

```
p <- 0.9
```

```
mc_res3 <- mc2(100)
```

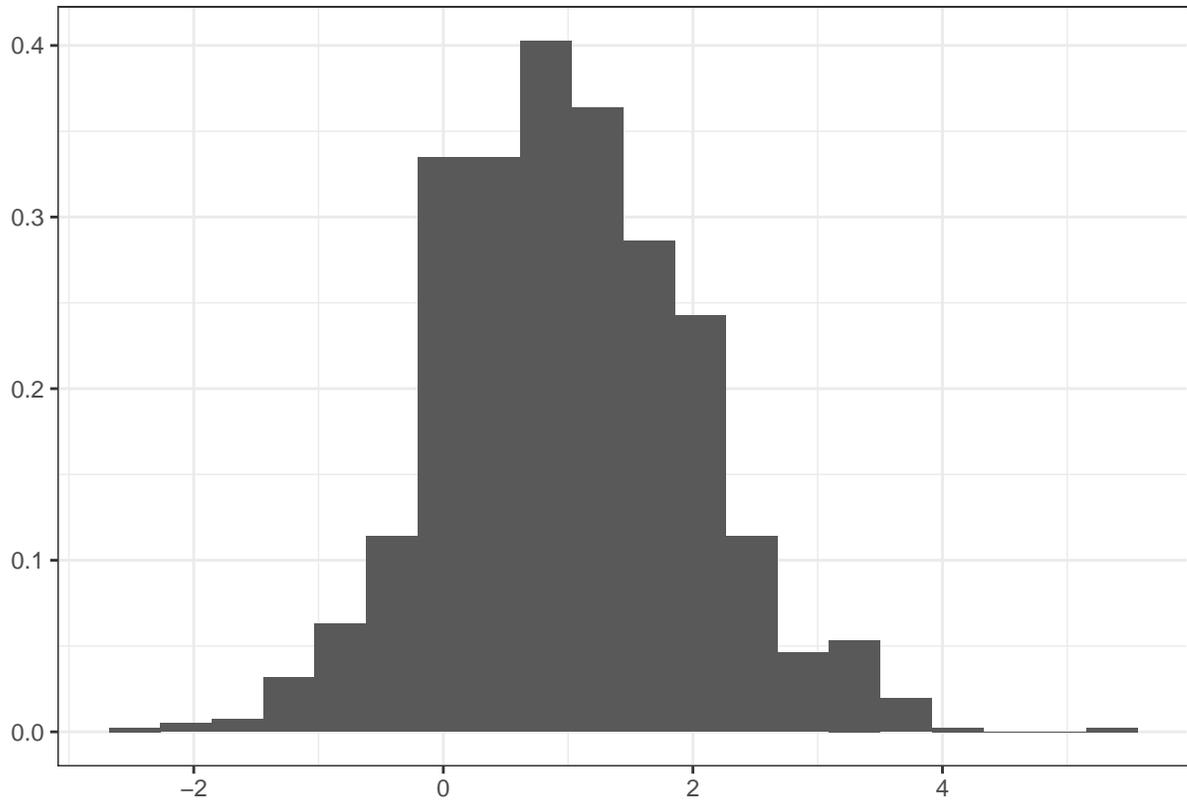
```
plot_t_results(mc_res3)
```



```
mean(mc_res3$rej)
```

```
## [1] 1
```

```
# change p to be 0.55  
p <- 0.55  
mc_res4 <- mc2(100)  
plot_t_results(mc_res4)
```



```
mean(mc_res4$rej)
```

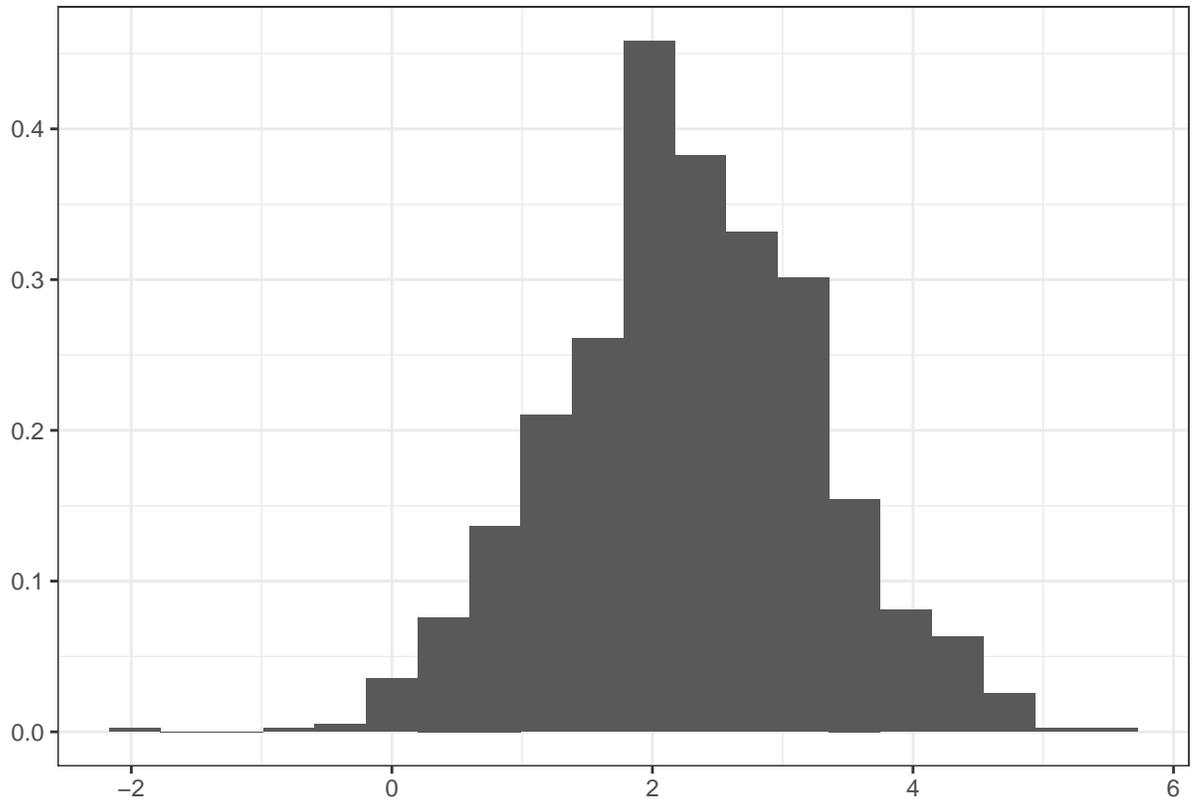
```
## [1] 0.201
```

```
# change p to be 0.55 and n=500
```

```
p <- 0.55
```

```
mc_res5 <- mc2(500)
```

```
plot_t_results(mc_res5)
```



```
mean(mc_res5$rej)
```

```
## [1] 0.635
```