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 θ 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 θ_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 Θ is the set of possible values of θ (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,

$$T = T\big((Y_1, X_1), \dots, (Y_n, X_n)\big)$$

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,

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

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

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

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

The most common approach to hypothesis testing is to pre-select a **significance level** α (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 α. Subject to this constraint, the goal is then to have high power. The power of a test depends on the true value θ (generally, power is higher for values of θ further away from θ_0) and on the sample size n (power increases with larger sample sizes)

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).

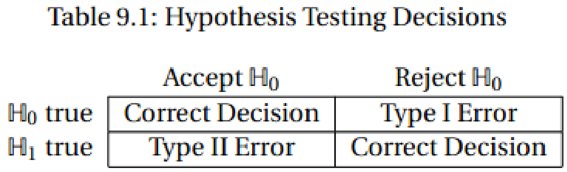


Table 9.1: Hypothesis Testing Decisions

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

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

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

where $\hat{\sigma}^2$ is an estimator of σ^2. For example, if $\theta = \text{E}[g(x)]$, then, it it is natural to use $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (g(X_i) - \bar{g})^2$ where $\bar{g} := \frac{1}{n}\sum_{i=1}^n g(X_i)$.

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

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

where $\hat{\mathbf{V}}$ is an estimator of $\mathbf{V}$. For example, if $\theta = \text{E}[g(X)]$, then a natural way to estimate the variance is $\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^n (g(X_i) - \bar{g})(g(X_i) - \bar{g})'$ which is a $k \times k$ matrix. And, $\hat{\mathbf{V}}_{jj}$ is the (j, j) element of $\hat{\mathbf{V}}$ (which is an element along the diagonal of $\hat{\mathbf{V}}$)

t-statistic

PSE: 13.9

When θ 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. To understand why, it is helpful to re-write it as

$$\begin{aligned} t &= \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{\mathbf{V}}_\theta}} \\ &= \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\mathbf{V}_\theta}} + o_p(1) \end{aligned}$$

where the first equality holds by the definition of s.e.$(\hat{\theta})$ and the second equality holds because $\hat{\mathbf{V}}_\theta$ is consistent for $\mathbf{V}_\theta$ and by the continuous mapping theorem.

Next, consider how t behaves in two different scenarios:

- If $\mathbb{H}_0$ is true, then $\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathbf{V}_\theta)$. This implies that $t \xrightarrow{d} N(0, 1)$. 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.

- Now consider the case where $\mathbb{H}_0$ is false. In this case $\hat{\theta} - \theta_0$ does not converge to 0 (rather, it converges to $\theta - \theta_0 \neq 0$). Thus, $\sqrt{n}(\hat{\theta} - \theta_0)$ diverges (i.e., goes to either positive or negative infinity, depending no the sign of $\theta - \theta_0$). This further implies that t diverges in this case.

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 will not reject $\mathbb{H}_0$. Otherwise (particularly if the magnitude of t is large), then 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 (that is typically chosen *ex ante*) and denoted by α. 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). Thus, notice that

$$\begin{aligned}
\mathrm{P}(|t| > c | \mathbb{H}_0) &\rightarrow \mathrm{P}(|Z| > c) \\
&= \mathrm{P}(Z < -c \text{ or } Z > c) \\
&= \mathrm{P}(Z < -c) + \mathrm{P}(Z > c) \\
&= 2\mathrm{P}(Z > c) \\
&= 2(1 - \Phi(c)) = \alpha
\end{aligned}$$

where $Z \sim N(0,1)$, the third equality holds because Z is symmetric. This suggests a way to choose a critical value, given a value of α. For example, when $\alpha = 0.05$, $c = 1.96$ or when $\alpha = 0.01$, then $c = 2.58$.

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 (asympotic) 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 α for which the test would reject $\mathbb{H}_0$). Given that $t \xrightarrow{d} Z \sim N(0,1)$,

$$\begin{aligned}
p &= \mathrm{P}(Z < -|t|) + \mathrm{P}(Z > |t|) \\
&= 2(1 - \Phi(|t|))
\end{aligned}$$

where the second line follows by symmetry of Z. 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

Let's continue with the case where θ is a scalar. $\hat{\theta}$ is **point estimator** for θ; that is, it is a single value. An alternative approach is to provide an **interval estimator** for θ. This is typically of the form, $\hat{C} = [\hat{L}, \hat{U}]$ (here: L stands for "lower" and U stands for "upper").

Notice that $\hat{C}$ itself is random because it is a function of the data. The **coverage probability** of $\hat{C}$ is $\mathrm{P}(\theta \in \hat{C})$. The randomness comes from $\hat{C}$ because θ is a fixed, though unknown, population parameter. $\hat{C}$ is called a $(1-\alpha)$ confidence interval if, at least in large samples, $\mathrm{P}(\theta \in \hat{C}) = 1 - \alpha$.

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

$$\begin{aligned}
\mathrm{P}(\theta \in \hat{C}) &= \mathrm{P}(\hat{\theta} - c_{1-\alpha/2}\text{s.e.}(\hat{\theta}) < \theta < \hat{\theta} + c_{1-\alpha/2}\text{s.e.}(\hat{\theta})) \\
&= \mathrm{P}\left(-c_{1-\alpha/2} < \frac{\theta - \hat{\theta}}{\text{s.e.}(\hat{\theta})} < c_{1-\alpha/2}\right) \\
&= \mathrm{P}\left(c_{1-\alpha/2} > \frac{\hat{\theta} - \theta}{\text{s.e.}(\hat{\theta})} > -c_{1-\alpha/2}\right) \\
&= \mathrm{P}\left(-c_{1-\alpha/2} < t < c_{1-\alpha/2}\right) \\
&= \mathrm{P}(|t| < c_{1-\alpha/2}) \\
&\rightarrow \mathrm{P}(|Z| < c_{1-\alpha/2}) \\
&= 1 - \alpha
\end{aligned}$$

where the first equality holds by the definition of the confidence interval, the second equality holds by rearranging terms, the third equality holds by multiplying each term by -1, the fourth equality holds by the definition of t and by re-ordering terms, the fifth equality holds by the definition of absolute value, the sixth equality holds because $|t| \xrightarrow{d} |Z|$ and the last equality holds by the definition of $c_{1-\alpha/2}$.

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, its 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". For example, most economic policies involve some sort of cost. There could very well be many policies that have a positive effect on some outcome of interest, but where the positive effect is smaller than the cost of the policy.

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 θ is a scalar. Now suppose that θ is a $k \times 1$ vector. Further, suppose that we are interested in $\mathbb{H}_0 : \theta = \theta_0$. (Note the arguments below go through for the case where we are interested in conducting inference with respect to $\beta = h(\theta)$ and where β is a vector). It is hard to operationalize the t-statistic that we talked about above. Instead, we will consider a Wald statistic:

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

Notice that this is a number that we can compute (given a value of θ_0) and that it is a scalar. As for t above, let's consider the behavior of W under $\mathbb{H}_0$ and under $\mathbb{H}_1 : \theta \neq \theta_0$.

Before doing this, it is useful to recall the following result (see Theorem 5.3, part 4 in the textbook): If $Z \sim N(0, \mathbf{I}_k)$, then $Z'Z \sim \chi^2_k$ (that is, $Z'Z$ follows a chi-square distribution with k degrees of freedom).

- If $\mathbb{H}_0$ is true, then

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

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

- 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. This means that, if $\mathbb{H}_0$ is not true, then W will be "large".

As for the t-statistic, this different behavior under $\mathbb{H}_0$ relative to $\mathbb{H}_1$ provides an approach to inference. 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

$$\text{p-value} = 1 - G_k(W)$$

where G_q is the cdf of a chi-square random variable with k degrees of freedom.

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)
```

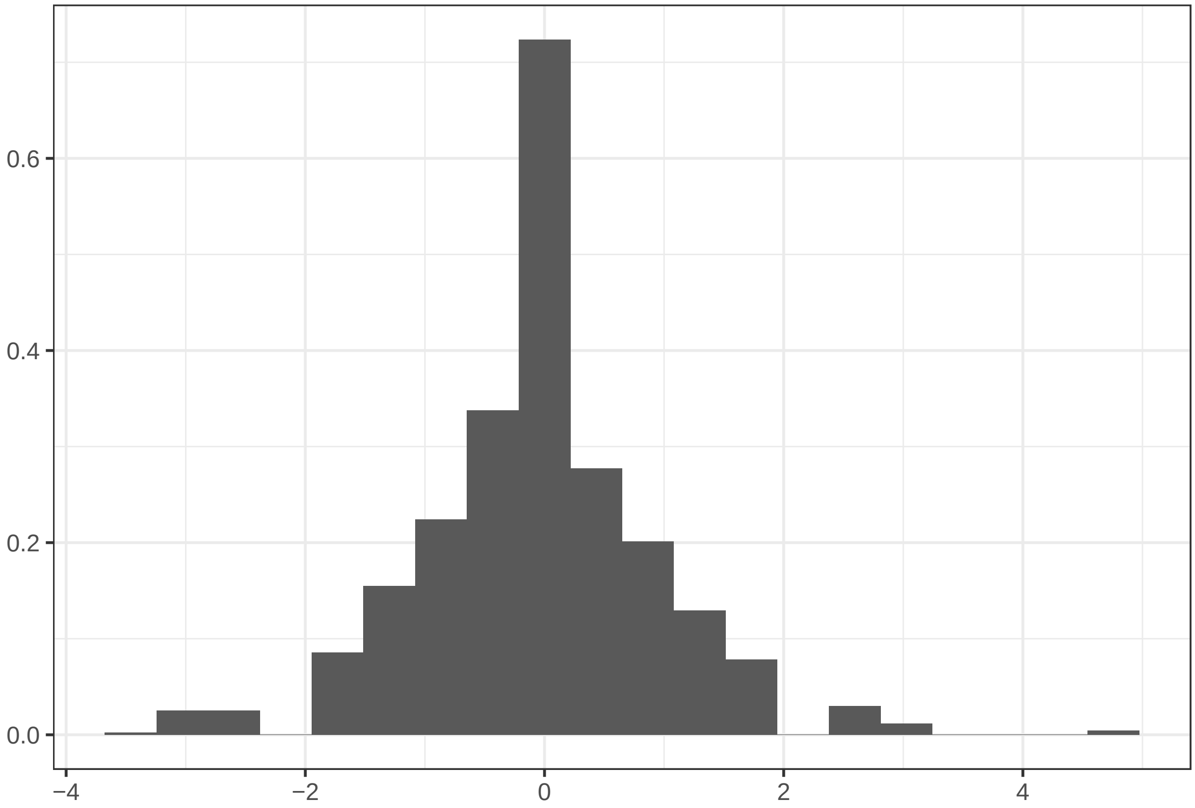
```
## [1] 0.0302
```

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)
```

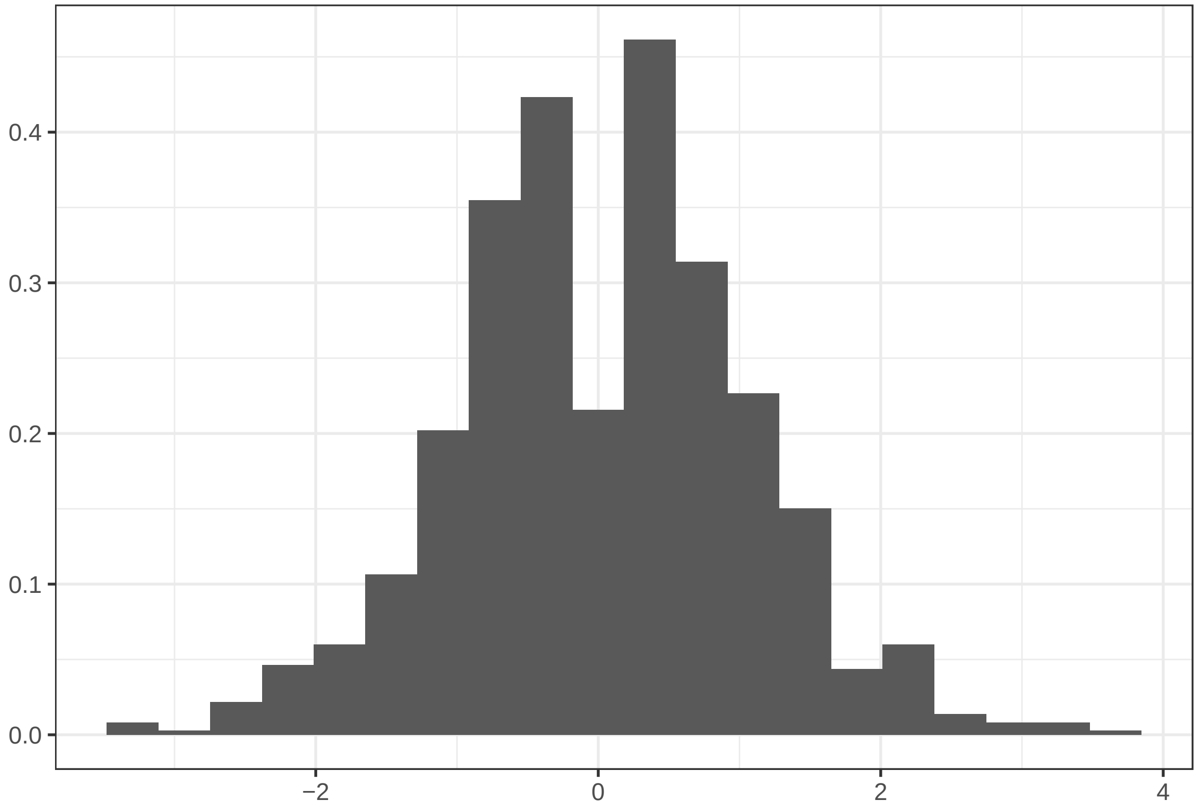
```
mc2 <- function(n, H0=0.5, nsims=1000) {
  phat <- c()  # vector to hold estimated p
  tstat <- c() # vector to hold estimated beta
  rej <- c()   # vector to hold whether or not we reject H0
  for (i in 1:nsims) {
    Y <- generate_sample(n,p)
    phat[i] <- mean(Y)
    tstat[i] <- sqrt(n)*(phat[i]-H0)/sqrt(var(Y))
    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
```

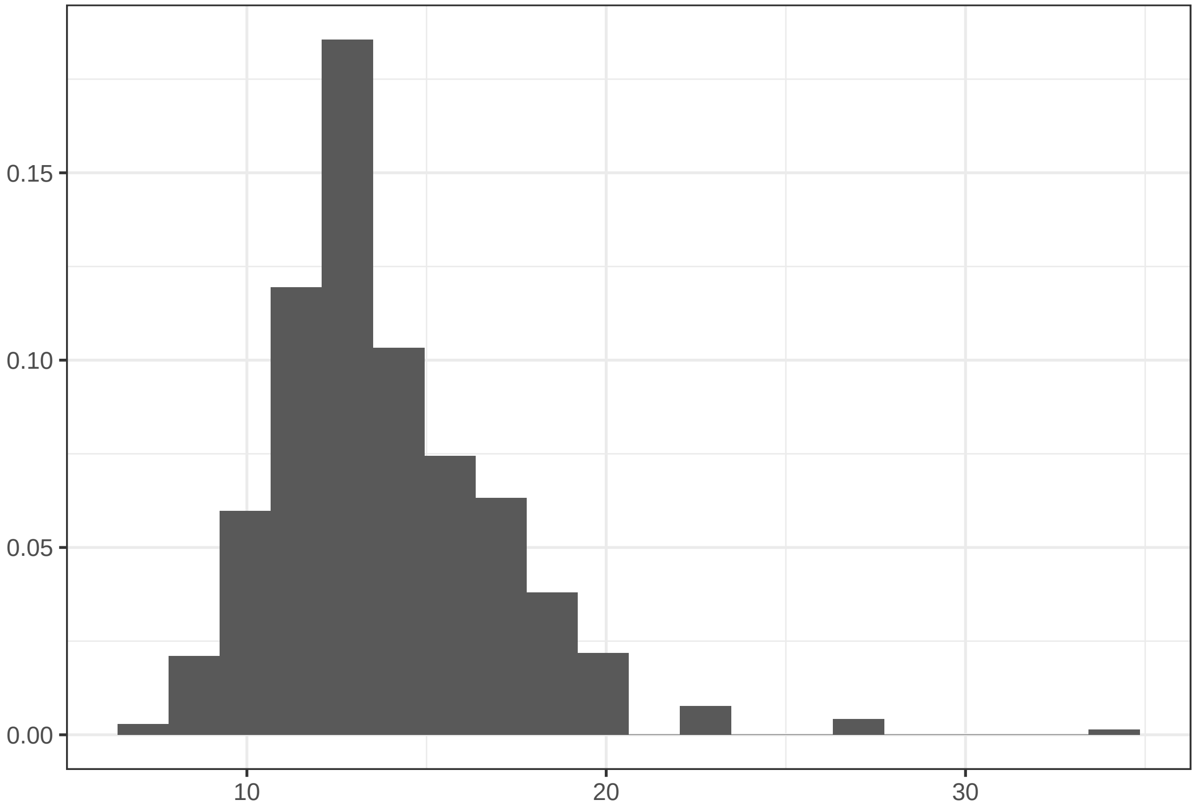
```
# n=100
mc_res2 <- mc2(100)
plot_t_results(mc_res2)
```



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

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

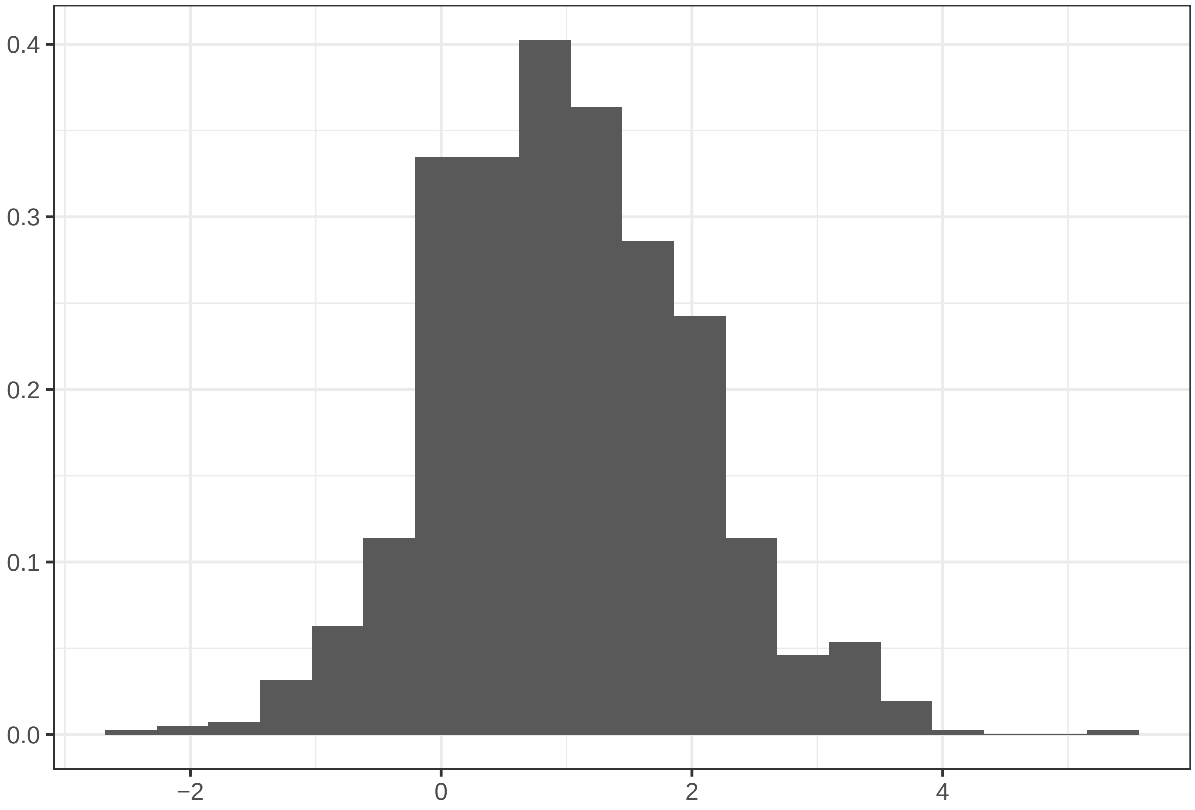
```
# change p to be 0.9
p <- 0.9
mc_res3 <- mc2(100)
plot_t_results(mc_res3)
```



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

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

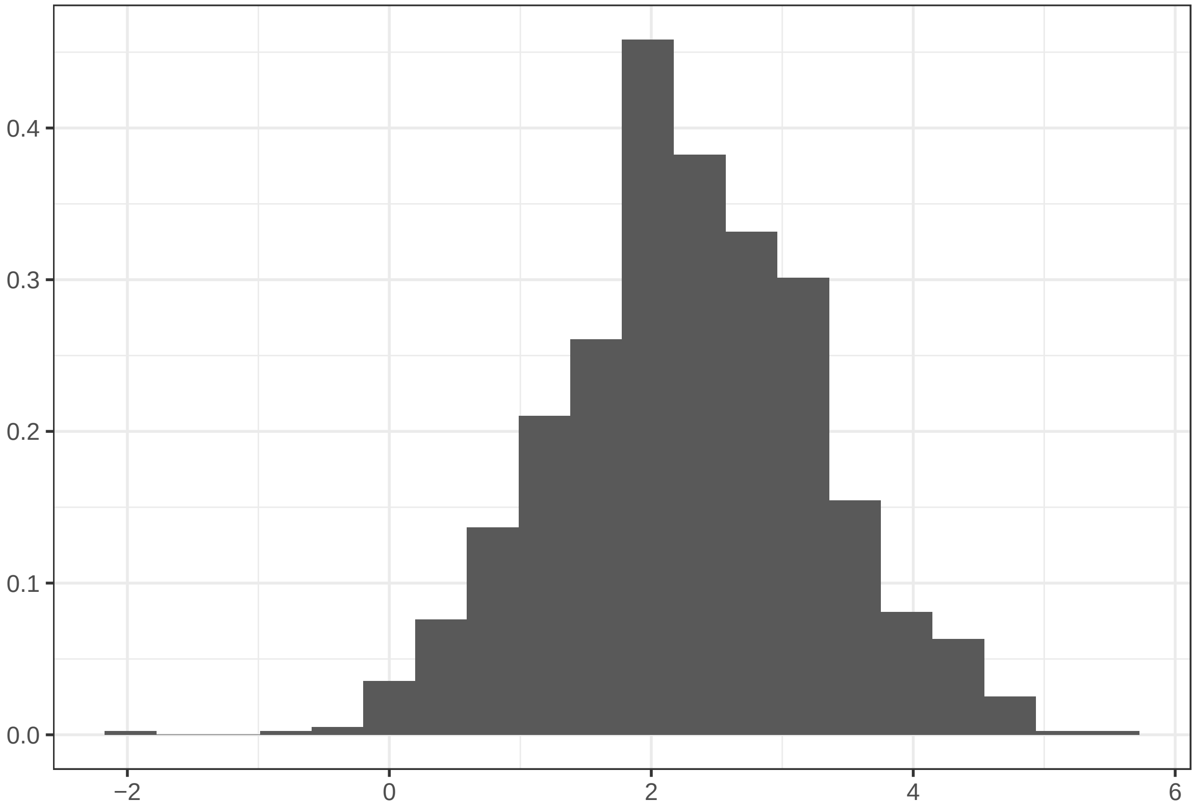
```
# change p to be 0.55
p <- 0.55
mc_res4 <- mc2(100)
plot_t_results(mc_res4)
```



```
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
```