Inference

Introduction

We know that b is a good *estimate* of β

It is unbiased: $E(b) = \beta$

It is consistent: $b \underset{n \to \infty}{\to} \beta$

But it still only an estimate

We want a better idea of what we can say about β given our estimate b

The normal distribution

Our model of y is

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

The error term ε represents the randomness in y – all of the factors besides x that explain y

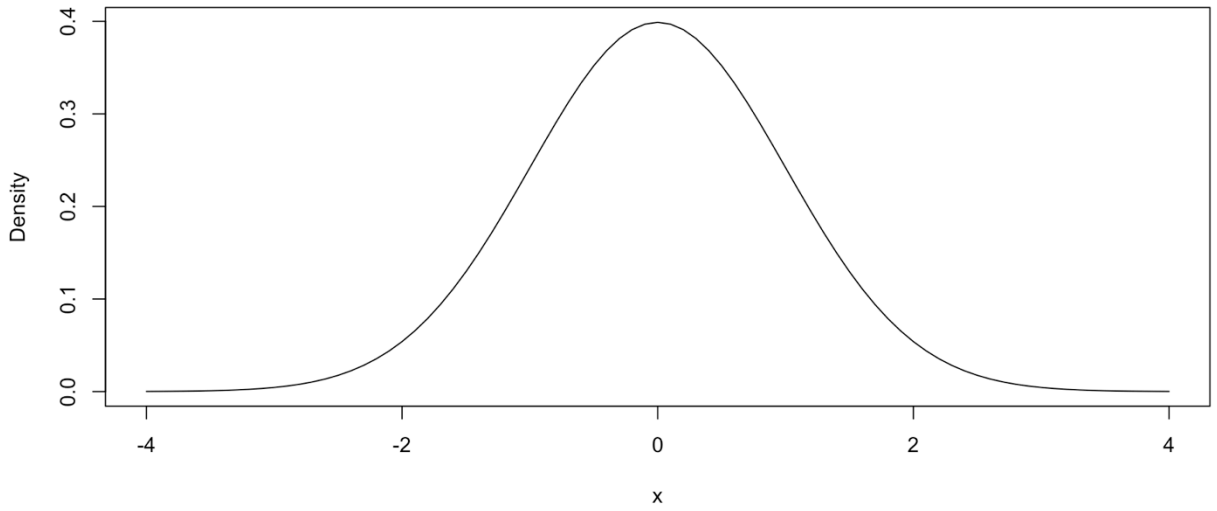
To introduce **inference**, we will make an additional assumption about ε

Assumption 5. ε is *normally distributed*

The normal distribution is the “bell curve” that you learned about in your stats class

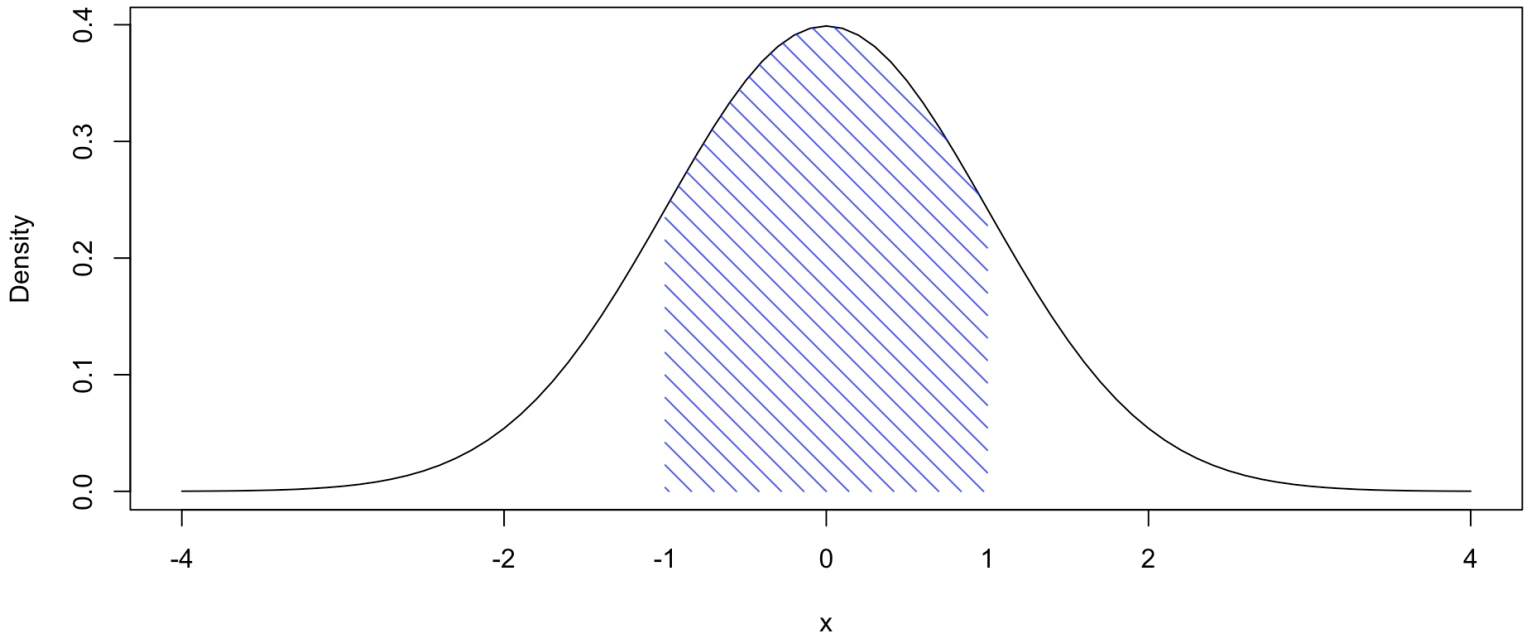
Here is a picture of it:

```r
x <- seq(-4, 4, by = 0.1)
y <- dnorm(x)
plot(x, y, type = "l", xlab = "x", ylab = "Density")
```



To interpret this, we need to recall a fact from stats

For a continuous RV, the probability that x lies between a and b is the *area under the* ***density*** *of* x *between* a *and* b



Now, since (as we proved before, assuming deviations-from-means form),

$$b = \beta + \frac{\sum_i x_i \varepsilon_i}{\sum x_i \varepsilon_i}$$

Since ε is the only source of randomness, if $\varepsilon \sim N$, $b \sim N$ as well (remember that b is also a RV)

Is the assumption that $\varepsilon \sim N$ strong?

It turns out that b is always approximately normally distributed as the sample size grows

This is because of the **Central Limit Theorem**, which says that sample averages are normally distributed with large samples[1]

1. And $b = (\sum_i x_i y_i)/(\sum_i x_i^2)$ is essentially the ratio of two sample averages

Hypothesis testing basics

We want to know whether we can draw certain conclusions about β from knowledge of b

For example, suppose we want to know whether β is different from zero (i.e., whether x has an effect on y)

Our **null hypothesis** is:

$$H_0 : \beta = 0$$

Our **alternative hypothesis** is just the opposite:[1]

$$H_1 : \beta \neq 0$$

1. some people write H_A instead of H_1

We want to design a test to see whether the data are consistent with the null hypothesis

If the data *are not* consistent, we will *reject* the null hypothesis

If they are, we will *retain* or *fail to reject* the null[1]

Quirky interpretation: If we're regressing y on x, we probably think x affects y, so our null hypothesis is the *opposite* of what we think is probably true, and we're seeing if we can reject that

1. You could even say "accept" the null, but some people don't like this

Now, we're doing statistics, so we know we're going to make mistakes sometimes (that's just how statistics goes)

A **Type I** error is the probability that we reject the null hypothesis when it's actually true (i.e., $\beta = 0$, but we conclude otherwise)

A **Type II** error is the probability that we fail to reject the null when it is false (i.e., $\beta \neq 0$, but we conclude otherwise)

If H_0 is "innocent", a Type I error sends an innocent person to jail, while a Type II error let's a guilty person go free

Type I errors are much worse, so we want to design a test where we have control over the probability that we incorrectly reject H_0

Ahead of time, we choose what we want the probability of a Type I error to be

We usually choose either .01, .05, or .1, depending on how comfortable we are with committing Type I errors

This is known as the α level, *significance level*, or *size* of the test[1]

1. The probability of a Type II error is out of our control, depending on the sample size and how large the true effect is

Hypothesis testing with regression

So far, our null hypothesis is that $\beta = 0$

We want to test this using b

We know that b is normally distributed

Since b is unbiased, under the null, the mean of b is 0

Let's also take b and *standardize it* by diving by the standard deviation of b

This makes it so that the normalized quantity $b/sd(b)$ has a *standard normal* distribution (a mean of 0 and a variance of 1)

Estimating $sd(b)$

Previously, we showed that

$$Var(b) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

The standard deviation of b is just $sd(b) = \sqrt{Var(b)}$

Problem: We can figure out $\sum_i (x_i - \bar{x})^2$ but we don't know σ^2 (this is the variance of ε, which we don't have data on)

Solution: The population line is

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

and our estimated sample line is

$$y_i = a + b x_i + e_i$$

This suggests we can use e_i as an estimate of ε_i

Our estimate of σ^2 will be

$$s^2 = \frac{\sum_i e_i^2}{n-2}$$

This is like the sample variance of e_i

We don't subtract $\bar{e}$ because it always equals zero

We divide by $n-2$ because e_i depends on our estimates of two coefficients: a and b

Our estimate of $Var(b)$ is

$$\frac{s^2}{\sum_i (x_i - \bar{x})^2}$$

Our estimate of $sd(b)$ (aka the standard error of b) is just the square root of this, or

$$se(b) = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}$$

The test statistic

Our **test statistic** or **t statistic** is

$$t = \frac{b}{se(b)}$$

Replacing $sd(b)$ with its *estimate* $se(b)$ slightly changes the distribution of our test statistic

Now, instead of being normally distributed, it has a t *distribution with* $n - 2$ *degrees of freedom*[1]

1. The $n - 2$ comes from the fact that we estimated two coefficients, a and b

The t distribution is like the normal distribution, but it has fatter tails

As n grows, the t distribution gets closer and closer to a normal distribution

If the null hypothesis is true, we know that this test statistic has a t distribution that is centered around zero

Intuition: If t is really far from zero, that makes it look like the null is false, so we should reject the null hyptohesis

Critical values

Question: How far from zero is “really far from zero”?

Answer: We have to use the **critical values** for a t-distribution

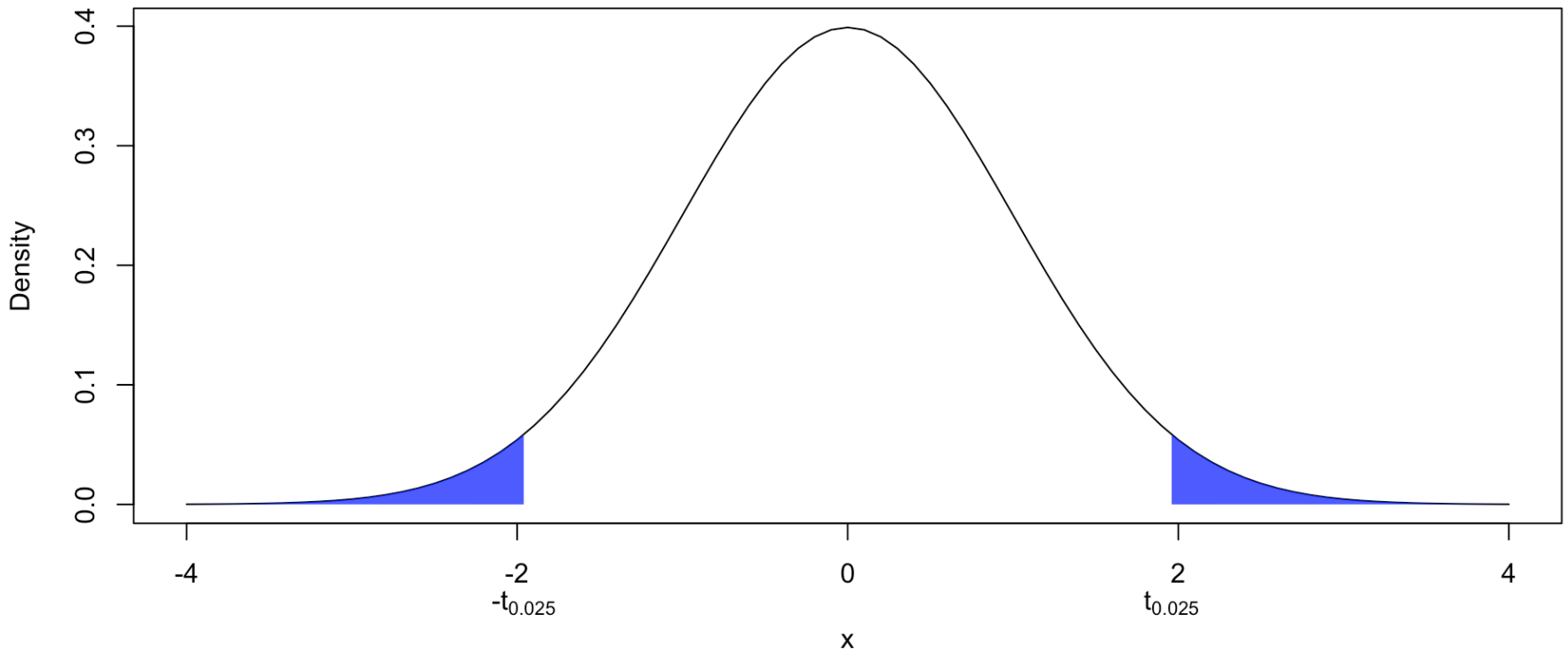
The critical values “trap” a certain percentage of the distribution

E.g.: Suppose that $\alpha = .05$. Then if the null is true, there is a 95% chance that

$$-t_{.05/2}^{(n-2)} \leq t \leq t_{.05/2}^{(n-2)}$$

If we wanted to use a different α level, we would find $t_{\alpha/2}^{(n-2)}$

This is best understood graphically:



Testing the hypothesis

When the null hypothesis is true, 2.5% of the distribution is to the left of $-t_{.025}^{(n-2)}$ and 2.5% of the distribution is to the right of $t_{.025}^{(n-2)}$

Therefore, under the null, the probability of getting a test statistic that is *as or more extreme* as $t_{.025}^{(n-2)}$ is 5%

Since there is only a 5% chance of this happening when the null is true, we reject the null if $t > t_{.025}^{(n-2)}$ or $t < -t_{.025}^{(n-2)}$:

$$\text{reject } H_0 \text{ if } |t| > t_{.025}^{(n-2)}, \text{ otherwise retain } H_0$$

To summarize, the test procedure is

1. Set the α level
2. Form the test-statistic $t = b/se(b)$
3. Find the critical value $t_{\alpha/2}^{(n-2)}$
4. If $|t| > t_{\alpha/2}^{(n-2)}$, reject the null. Otherwise, retain the null

How do we figure out the critical values?

We just look them up in a table. Practically every statistics and econometrics book has tables of critical values for the t (and other) distributions

Our you can just look them up online, for example here

E.g.:

```r
library(tidyverse)
gril <- read.csv("griliches.csv")
gril$wage <- exp(gril$lw)
our.model <- lm(wage ~ s, gril)
summary(our.model)
```

```
Call:
lm(formula = wage ~ s, data = gril)

Residuals:
    Min      1Q  Median      3Q     Max 
-344.94  -84.91  -20.59   60.90  676.12 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -125.295     28.458  -4.403 1.22e-05 ***
s             33.511      2.094  16.002  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 128.6 on 756 degrees of freedom
```

Let's test the null hypothesis that schooling has no effect on the wage

The t-stat is

$$t = \frac{33.51}{2.1} \approx 15.95$$

Note that this is close to what R reports as the "t value" (difference due to rounding)

Let's assume our $\alpha = .05$

Our dataset has $n = 758$ observations, so our degrees of freedom is $n - 2 = 756$

The t-table at the link above gives critical values in terms of the probability of being to the *left* of the critical value (instead of the probability of being to the right, as we have defined them)

So we need to look up the $1 - .05/2 = .975$ critical value (i.e., there is a 97.5% chance of being to the left of the critical value) for 756 DF

The closest we can get is 100 DF, and for that the value is 1.96

Since $15.95 > 1.96$, we (easily) reject the null

“Statistical significance”

When we can reject the null that $\beta = 0$, we say that a coefficient is “statistically significant” at the given α level

Most software automatically gives indications of whether coefficients are statistically significant at different levels

R uses different numbers of asterisks to denote signfiicance at different levels

Beware: Just because you can say that something *isn’t* zero doesn’t mean that it’s necessarily economically significant

E.g.: Let's illustrate that this testing procedure works by creating a model where $\beta = 2$ and testing $H_0 : \beta = 2$

```r
t <- rep(0,1000)
for (i in 1:1000) {
  x <- rnorm(100)
  e <- rnorm(100)
  y <- 1 + 2*x + e
  b <- lm(y ~ x)$coef[2]
  se <- sqrt(vcov(lm(y ~ x))[2,2])
  t[i] <- (b-2)/se
}
mean(abs(t)>1.96)
```

```
[1] 0.058
```

Other null hypotheses

So far, we've discussed how to test the null hypothesis that $\beta = 0$

What if we're interested in other values?

Easy. Say we want to test the hypothesis that β takes some particular value β_0

Then the null is $H_0 : \beta = \beta_0$ and the alternative is $H_1 : \beta \neq \beta_0$

The *only* change to the testing procedure is now, the t-stat becomes:

$$t = \frac{b - \beta_0}{se(b)}$$

E.g.: Suppose we want to test the null hypothesis that $\beta = 30$

Now that t-stat is

$$t = \frac{33.5 - 30}{2.1} = \frac{3.5}{2.1} \approx 1.66$$

Since $|1.66| < 1.96$, we *fail to reject* this null hypothesis

We can easily reject the null that there is no effect, but we can't reject the null of an effect 30

One-sided hypothesis tests

What if, instead of wondering whether β equals a particular value, we're interested in whether it is above/below some value?

Now, let $H_0 : \beta < \beta_0$ and $H_1 : \beta \geq \beta_0$

How can we do hypothesis testing in this case?

Previously, if b was much smaller *or* much larger than β_0, we would reject the null hypothesis

Now, a value of b that is much smaller than β_0 is *consistent* with our null that $\beta < \beta_0$, we we're only going to reject if b is much *larger* than β_0 (i.e., if $b - \beta_0$ is large)

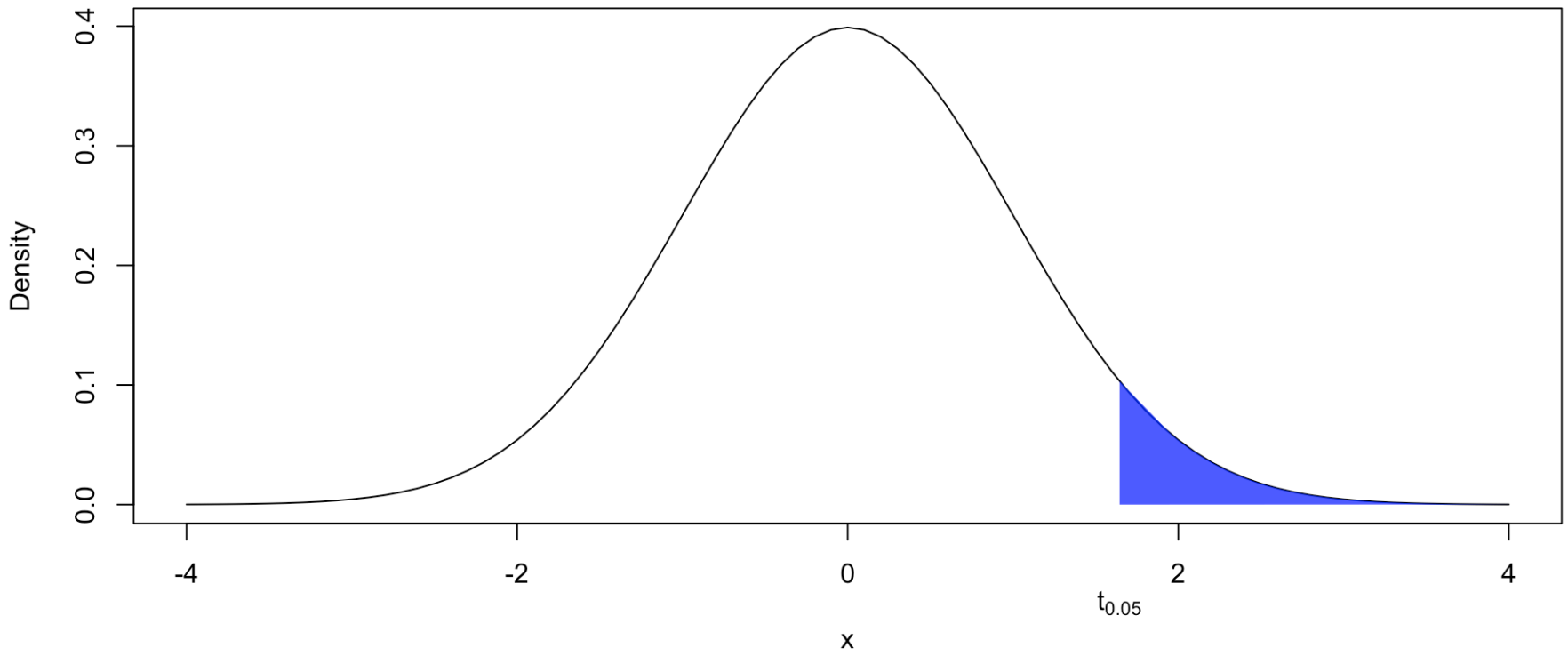
Since

$$t = \frac{b - \beta_0}{se(b)},$$

we will only reject if t is positive (in the *right* tail of the distribution)

But if we want to maintain the α-level of our test, we need the critical value that has $100 * (1 - \alpha)\%$ of the distribution to its right (before, we only had half that much to the right)

Now, the picture looks like this:



Suppose $\alpha = .05$

If the null that $\beta < \beta_0$ is true, the probability of getting a t-stat this is *greater than or equal to* $t_{.05}^{(n-2)}$ is 5%

If we get a t-stat further in the right tail than the critical value, we reject the null

Quiz: How would you change this test if instead the null was $H_0 : \beta > \beta_0$ and the alternative was $H_1 : \beta \leq \beta_0$?

E.g.: Now suppose the null is that $\beta < 25$ (so the alternative is that $\beta \geq 25$)

The t-stat becomes

$$t = \frac{33.5 - 25}{2.1} = \frac{13.5}{2.1} \approx 6.43$$

The .05 (or .95 if you use the NIST table) critical value for a t-distribution with >100 DF is 1.645

Since $6.43 > 1.645$, we reject the null (note that there is no absolute value here, we only reject if the t stat *itself* exceeds the *positive* critical value)

Confidence intervals

Recall from the definition of critical values that there is a 95% chance that

$$-t_{.025}^{(n-2)} \leq \frac{b-\beta}{se(b)} \leq t_{.025}^{(n-2)}$$

Rearranging this, there is also a 95% chance that

$$b - t_{.025}^{(n-2)} se(b) \leq \beta \leq b + t_{.025}^{(n-2)} se(b)$$

The **95% confidence interval** for β is

$$[b - t_{.025}^{(n-2)} se(b), b + t_{.025}^{(n-2)} se(b)]$$

There is a 95% chance that β lies in this interval

If we want a $100 * (1 - \alpha)$ CI instead, we just replace $t_{.025}^{(n-2)}$ with $t_{\alpha/2}^{(n-2)}$

Quirky interpretational note: Once we have a confidence interval, either it contains β or it doesn't. The right way to think about this is that if we ran many regressions on different samples, 95% of our confidence intervals would contain β

We can also use confidence intervals to conduct hypothesis tests

If our null is $H_0 : \beta = \beta_0$, we reject the null if the confidence interval does not contain β

E.g.: For the wage example, the 95% confidence interval is

$$[33.5 - 1.96 * 2.1, 33.5 + 1.96 * 2.1] = [29.38, 37.62]$$

If $H_0 : \beta = 0$, we can reject the null at the 5% level because this CI does not contain zero

Let's also illustrate the use of CIs by showing what fraction of the time the confidence interval contains the true β of 2

```
in.ci <- rep(0,1000)
for (i in 1:1000) {
  x <- rnorm(100)
  e <- rnorm(100)
  y <- 1 + 2*x + e
  b <- lm(y ~ x)$coef[2]
  se <- sqrt(vcov(lm(y ~ x))[2,2])
  low <- b - 1.96*se
  high <- b + 1.96*se
  in.ci[i] <- (low <= 2 && 2 <= high)
}
mean(in.ci)
```

```
[1] 0.943
```

While it's good to know how to create confidence intervals "by hand," we can get them automatically in R:

```r
confint(our.model, level=.95)
```

```
                 2.5 %    97.5 %
(Intercept) -181.16054 -69.42949
s             29.39969  37.62168
```

P-values

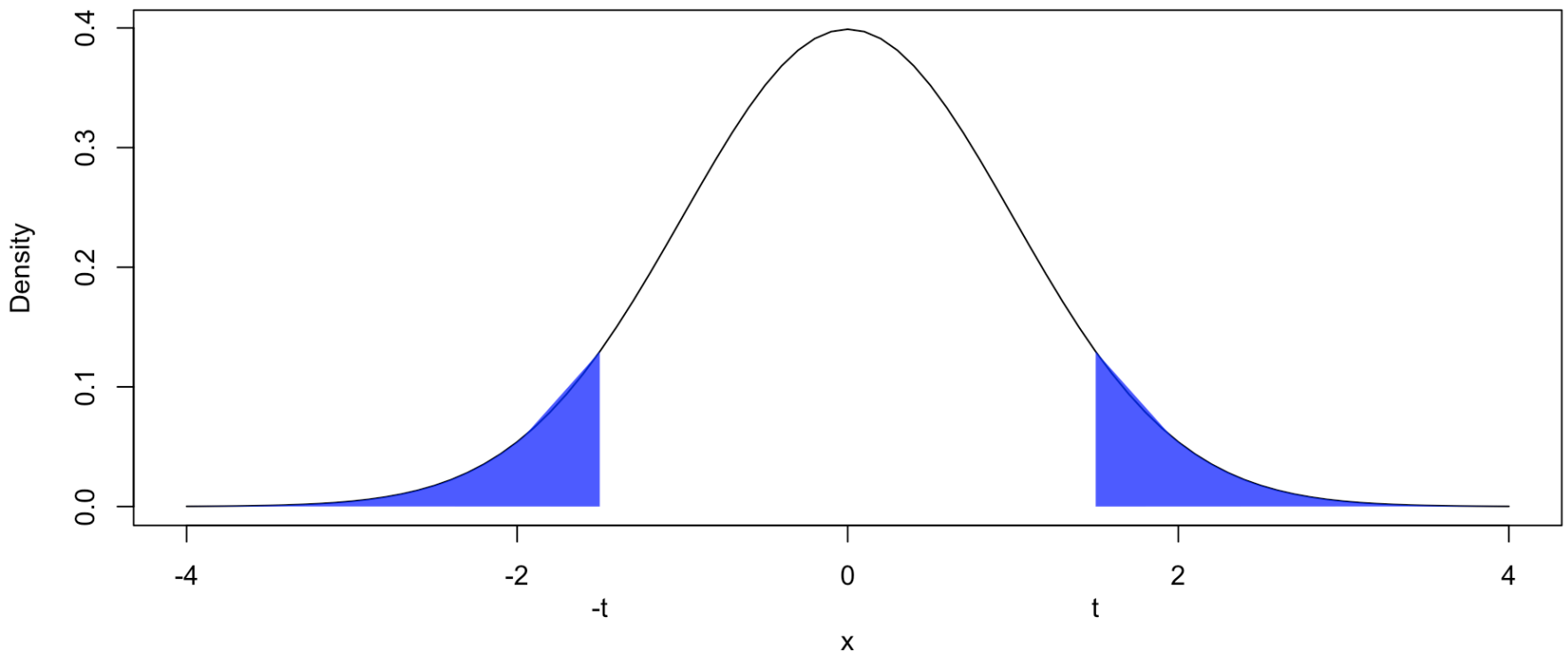
So far, our approach has been to compute the t-stat, then reject if it exceeds the critical value

Alternatively, we could compute the probability of getting a t-stat *as or more extreme* than the one we actually got

If this probability is less than our α level, we can reject the null hypothesis

The **p-value** is the probability of getting a t-stat *as or more extreme* than the one we actually got, when the null hypothesis is true

Now the picture is:



Question: Why bother with critical values if we can just use the p-value?

Answer: Historical reasons. To calculate p-values, we need to know the area to the right of *any* t-stat, whereas we only need to know a few critical values (the ones for .005, .025, and .05)

Nowadays, computers can easily calculate p-values, so we can do hypothesis testing however we want