STATS 250 Lab 10

Confidence Intervals and Hypothesis Tests for Proportions

Nick Seewald
nseewald@umich.edu

Week of 11/2/2020

Reminders

Your tasks for the week running Friday 10/30 - Friday 11/6:

Task	Due Date	Submission
Vote (if eligible)	Tuesday 11/3 8:00PM ET	Your Election Precinct
M-Write 2 Initial Submission	Thursday 11/5 4:59PM ET	Canvas
Lab 10	Friday 11/6 8:00AM ET	Canvas
Homework 7	Friday 11/6 8:00AM ET	course.work

Lab Demo: ISRS Problem 3.9

Life after college. We're interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

Part 1: What are we trying to find? What do we know?

Life after college. We're interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

Part 1: What are we trying to find? What do we know?

Life after college. We're interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

What is the population parameter of interest?

Part 1: What are we trying to find? What do we know?

Life after college. We're interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

What is the population parameter of interest?

We want to find p, the proportion of all graduates at a mid-sized university who found a job within one year of completing their undergraduate degree.

Lab Demo: ISRS Problem 3.9

Life after college. We're interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

Lab Demo: ISRS Problem 3.9

Life after college. We're interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

What is our *point estimate* of p?

Lab Demo: ISRS Problem 3.9

Life after college. We're interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

What is our *point estimate* of p?

$$\hat{p} = \frac{348}{400} = 0.87$$

Part 2: Check Conditions

Before we can make a confidence interval using the normal distribution, we want to make sure that our data meet certain conditions.

What conditions do we need to check?

Part 2: Check Conditions

Before we can make a confidence interval using the normal distribution, we want to make sure that our data meet certain conditions.

What conditions do we need to check?

1. **Independent observations:** graduates in the sample can't be related to each other
2. **Large enough sample:** $np \geq 10$ and $n(1-p) \geq 10$ (at least 10 "successes" and 10 "failures")

Part 2: Check Conditions

Check Independence

Part 2: Check Conditions

Check Independence

Our sample size of 400 is less than 10% of the population size of 4500. ✅

Part 2: Check Conditions

Check Independence

Our sample size of 400 is less than 10% of the population size of 4500. ✅

Check sample size

Part 2: Check Conditions

Check Independence

Our sample size of 400 is less than 10% of the population size of 4500. ✅

Check sample size

We don't know p, so we'll check this condition with $\hat{p}$, our best guess of p:

$$n\hat{p} = 400 \times 0.87 = \mathbf{348} \geq 10$$

$$n(1-\hat{p}) = 400 \times 0.13 = \mathbf{52} \geq 10$$

Both are at least 10 ✅

Step 3: Compute a confidence interval

Calculate a 95% confidence interval for p, the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

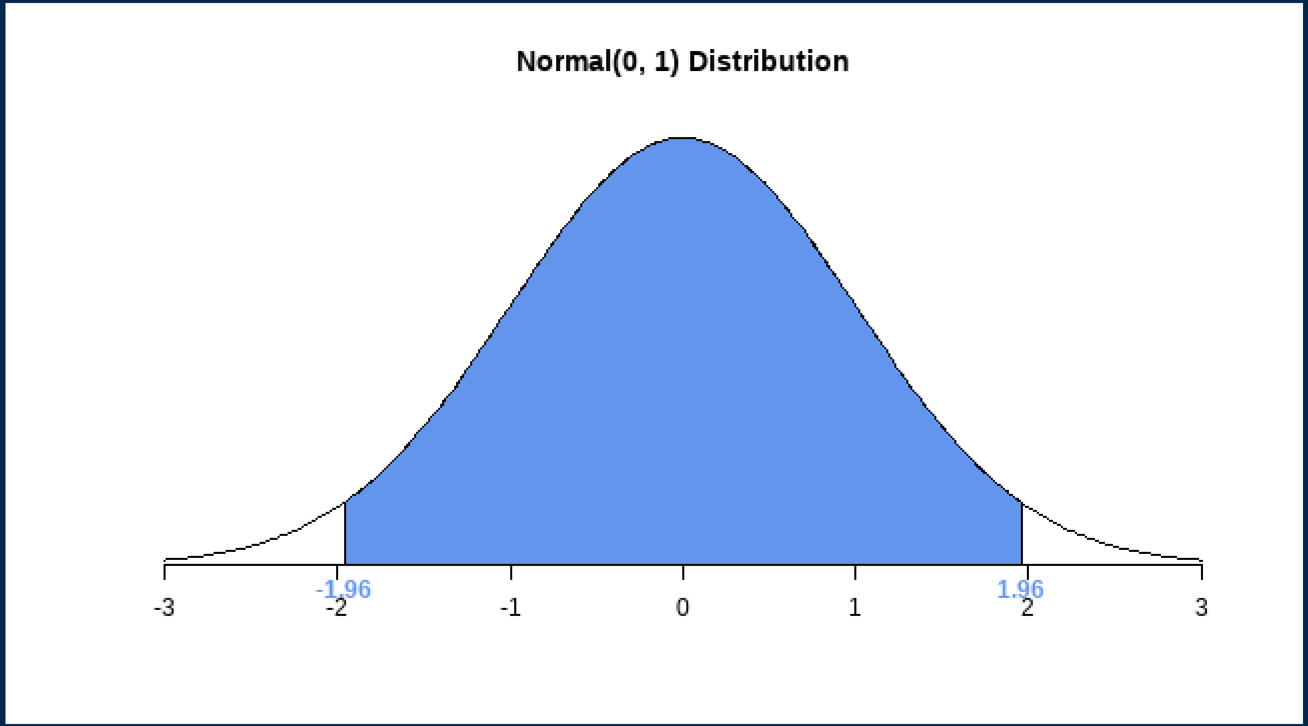
Remember that a confidence interval generally looks like

$$\text{estimate} \pm (\text{a few}) \times \text{SE}_{\text{estimate}}$$

Step 3: Compute a confidence interval

$$\text{estimate} \pm (\text{a few}) \times \text{SE}_{\text{estimate}}$$

Using a multiplier of 1.96 will give us a 95% confidence interval:



Step 3: Compute a confidence interval

We know from section 3.1 that

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

but since we don't know p, we'll use $\hat{p}$.

Use R as a calculator to compute $\mathrm{SE}_{\hat{p}}$, using $\hat{p} = 0.87$.

Step 3: Compute a confidence interval

We know from section 3.1 that

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

but since we don't know p, we'll use $\hat{p}$.

Use R as a calculator to compute $\mathrm{SE}_{\hat{p}}$, using $\hat{p} = 0.87$.

```
se <- sqrt(0.87 * (1 - 0.87) / 400)
se
```

```
[1] 0.01681517
```

Step 3: Compute a confidence interval

Now let's compute the **margin of error**: the term that's added to and subtracted from the estimate to get the limits of the confidence interval.

$$\text{estimate} \pm \underbrace{(\text{a few}) \times \text{SE}_{\text{estimate}}}_{\text{margin of error}}$$

Remember that "a few" here means 1.96 (for a 95% confidence interval)

Use R as a calculator to compute the margin of error.

Step 3: Compute a confidence interval

Now let's compute the **margin of error**: the term that's added to and subtracted from the estimate to get the limits of the confidence interval.

$$\text{estimate} \pm \underbrace{(\text{a few}) \times \text{SE}_{\text{estimate}}}_{\text{margin of error}}$$

Remember that "a few" here means 1.96 (for a 95% confidence interval)

Use R as a calculator to compute the margin of error.

```
moe <- 1.96 * se
moe
```

```
[1] 0.03295774
```

Step 3: Compute a Confidence Interval

Our confidence interval, therefore, is

$$0.87 \pm 0.033.$$

or

$$(0.837, 0.903)$$

How do we interpret this confidence interval?

Step 3: Compute a Confidence Interval

Our confidence interval, therefore, is

$$0.87 \pm 0.033.$$

or

$$(0.837, 0.903)$$

How do we interpret this confidence interval?

We are 95% confident that the population proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree is between .837 and .903.

Step 4: Interpreting a Confidence Level

What does "95% confidence" mean?

- Imagine that we know p is 0.85.
- Take repeated samples from this population, and make a confidence interval using each sample
- We expect about 95% of the resulting confidence intervals to contain $p = 0.85$

Step 4: Interpreting a Confidence Level

```r
set.seed(5902)

# LINE ~120 OR SO
ci <- replicate(50, {
  s <- sample(0:1, size = 400,
              replace = TRUE,
              prob = c(0.15, 0.85))

  pHat <- sum(s) / 400
  se <- sqrt(pHat * (1 - pHat) / 400)
  marginOfError <- 1.96 * se

  lowerLimit <- pHat - marginOfError
  upperLimit <- pHat + marginOfError

  c(lowerLimit, upperLimit)
})

ci <- t(ci)
```

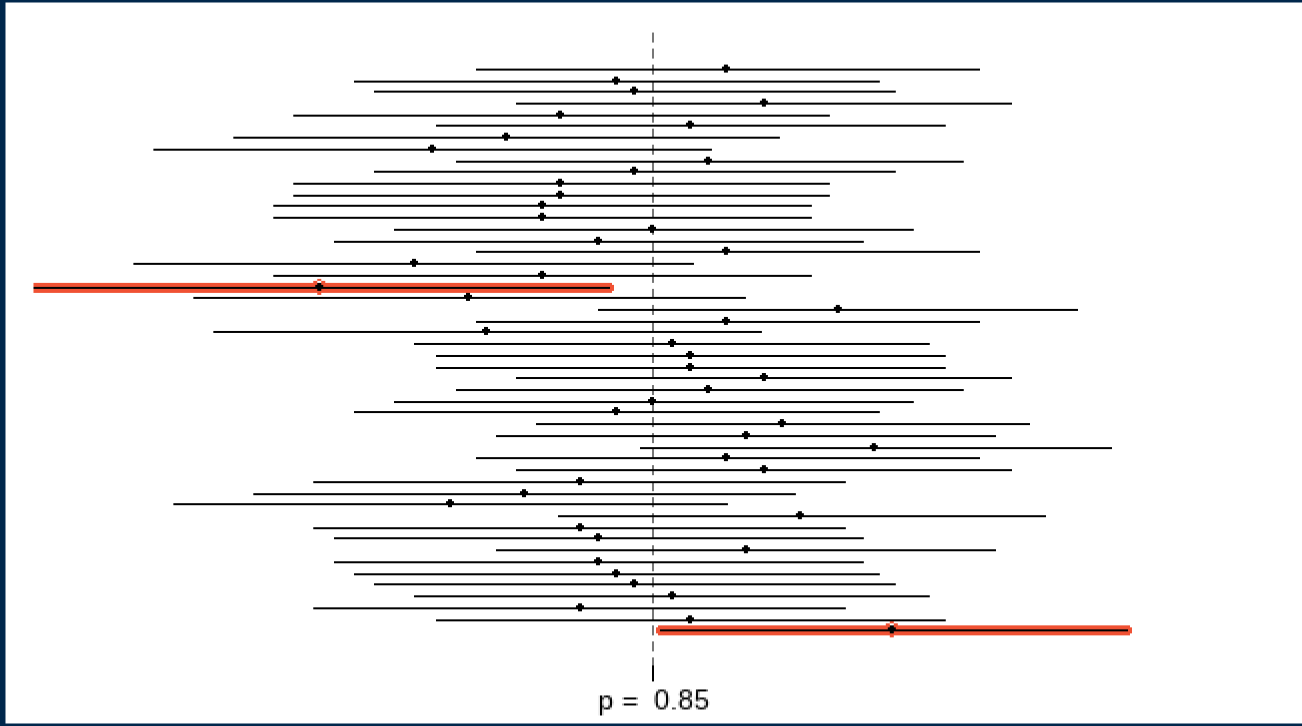
```r
head(ci)
```

```
          [,1]      [,2]
[1,] 0.8509425 0.9140575
[2,] 0.8204941 0.8895059
[3,] 0.8040726 0.8759274
[4,] 0.8177488 0.8872512
[5,] 0.8122685 0.8827315
[6,] 0.8095333 0.8804667
```

Step 4: Interpreting a Confidence Level

48/50 = 96% of the intervals contain $p = 0.85$.

```
plot_ci(lo = ci[, 1], hi = ci[, 2], m = 0.85)
```



Step 4: Interpreting a Confidence Level

How would you interpret the 95% confidence level?

Step 4: Interpreting a Confidence Level

How would you interpret the 95% confidence level?

If we repeated our sampling procedure many times, we would expect 95% of our resulting 95% confidence intervals to contain p, the true proportion of graduates who get a job within one year of finishing their undergraduate degrees.

R can do this for us (line ~156)

We can have R make confidence intervals for us:

```
prop_test(x = 348, n = 400, conf.level = 0.95)
```

```

    1-sample proportions test without continuity correction

data:  x out of n, null probability 0.5
Z = 14.8, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.8370429 0.9029571
sample estimates:
   p
0.87
```

Switch it up: 99% CI (line ~165)

Modify the code below to make a 99% confidence interval instead.

```
prop_test(x = 348, n = 400, conf.level = 0.95)
```

Switch it up: 99% CI (line ~165)

Modify the code below to make a 99% confidence interval instead.

```
prop_test(x = 348, n = 400, conf.level = 0.95)
```

```
prop_test(x = 348, n = 400, conf.level = 0.99)
```

```

	1-sample proportions test without continuity correction

data:  x out of n, null probability 0.5
Z = 14.8, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.826687 0.913313
sample estimates:
   p 
0.87 
```

How does the width of this interval compare to the 95% CI?

Hypothesis Testing with `prop_test()`

`prop_test()` creates a confidence interval **and** performs a hypothesis test. Let's test the following hypotheses:

$$H_0: p = 0.5 \quad \text{vs.} \quad H_a: p < 0.5$$

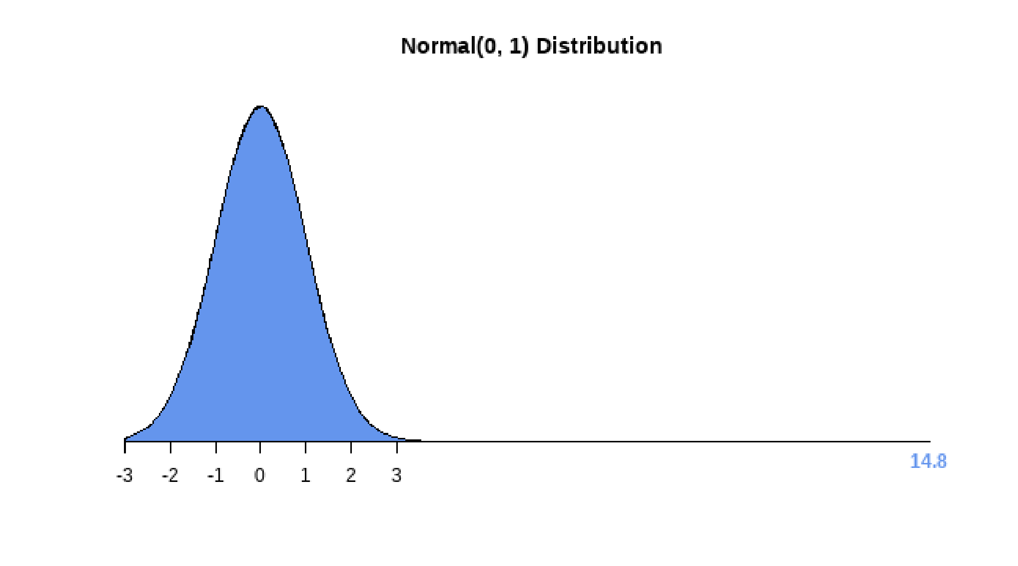
```
prop_test(x = 348, n = 400,
          p = 0.5, alternative = "less")
```

```

	1-sample proportions test without continuity correction

data:  x out of n, null probability p
Z = 14.8, p-value = 1
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.8976585
sample estimates:
   p
0.87
```

Hypothesis Testing with `prop_test()`

```
prop_test(x = 348, n = 400,
          p = 0.5, alternative = "less")
```

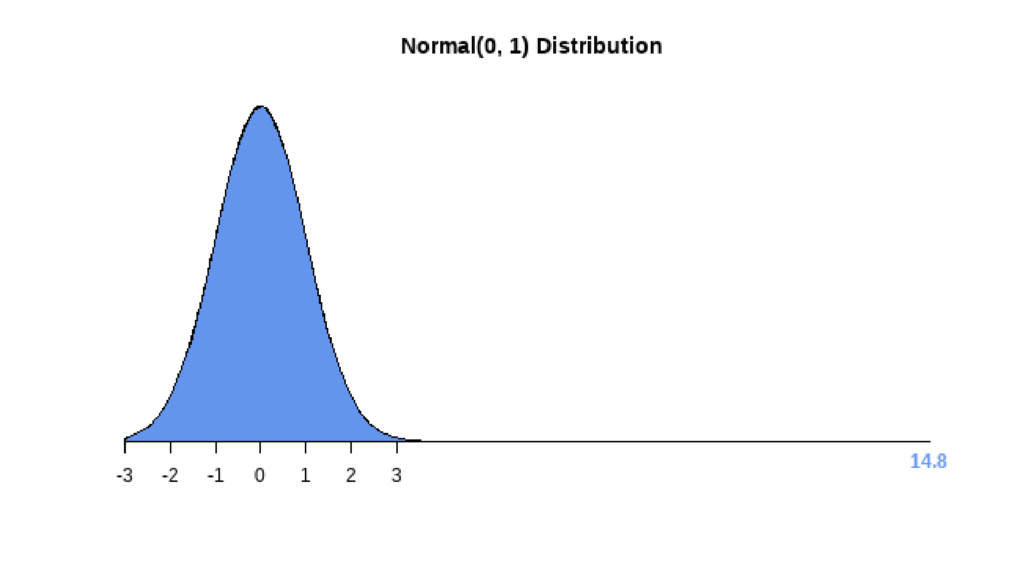
```
	1-sample proportions test without continuity corr

data:  x out of n, null probability p
Z = 14.8, p-value = 1
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.8976585
sample estimates:
   p
0.87
```



Hypothesis Testing with `prop_test()`

```
prop_test(x = 348, n = 400,
          p = 0.5, alternative = "less")
```

```
    1-sample proportions test without continuity corr

data:  x out of n, null probability p
Z = 14.8, p-value = 1
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.8976585
sample estimates:
   p
0.87
```

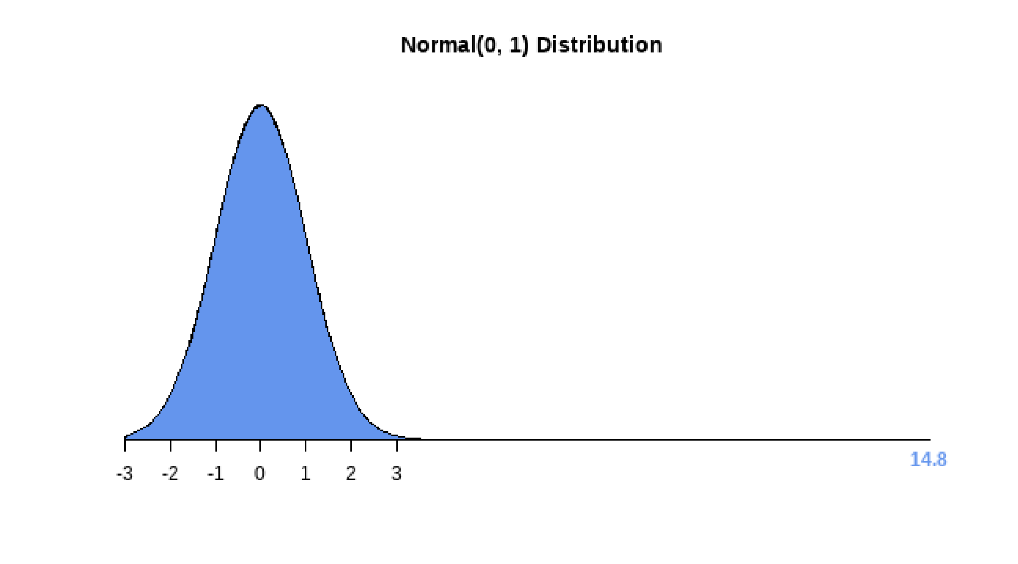


Why is that p-value 1?

Hypothesis Testing with `prop_test()`

```r
prop_test(x = 348, n = 400,
          p = 0.5, alternative = "less")
```

```
    1-sample proportions test without continuity corr

data:  x out of n, null probability p
Z = 14.8, p-value = 1
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.8976585
sample estimates:
   p
0.87
```



Why is that p-value 1?

We're testing to see if $p < 0.5$, but our data have $\hat{p} = 0.87$! Our data provide almost no evidence that $p < 0.5$, so we get a high p-value.

Careful with `alternative`!

```r
prop_test(x = 348, n = 400, conf.level = 0.95)
```

```

	1-sample proportions test without continuity correction

data:  x out of n, null probability 0.5
Z = 14.8, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.8370429 0.9029571
sample estimates:
   p 
0.87 
```

```r
prop_test(x = 348, n = 400,
          p = 0.5, alternative = "less")
```

```

	1-sample proportions test without continuity corr

data:  x out of n, null probability p
Z = 14.8, p-value = 1
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.8976585
sample estimates:
   p 
0.87 
```

If you want to make a confidence interval, you *must* do a two-sided test. Set `alternative = "two.sided"` or leave it blank.

`prop_test()` for Two Proportions

Pass a vector of the numbers of successes x and a vector of sample sizes n.

	Successes	Failures	Total
Group 1	28	2	30
Group 2	34	16	50
Total	62	18	80

```
prop_test(x = c(28, 34),
          n = c(30, 50),
          conf.level = 0.9)
```

```

	2-sample test for equality of proportions without
	correction

data:  x out of n
Z = 2.6269, p-value = 0.008616
alternative hypothesis: two.sided
90 percent confidence interval:
 0.1214773 0.3851894
sample estimates:
   prop 1    prop 2 
0.9333333 0.6800000 
```

Code Cheat Sheet 💻

`pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)`

- **`q`** refers to the value you want to find the area above or below
  - `pnorm(q, 0, 1)` gives $P(Z < q)$ where Z is $N(0, 1)$
- **`mean`** refers to μ, defaults to 0
- **`sd`** refers to σ, defaults to 1
- **`lower.tail`** controls which direction to "shade": `lower.tail = TRUE` goes less than q, `lower.tail = FALSE` goes greater than q; defaults to TRUE

Code Cheat Sheet 💻

```
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)
```

- `p` refers to the area under the curve
  - `qnorm(p, 0, 1)` is the number such that the area to the left of it is p
- `mean` refers to μ, defaults to 0
- `sd` refers to σ, defaults to 1
- `lower.tail` controls which direction to "shade": `lower.tail = TRUE` goes less than q, `lower.tail = FALSE` goes greater than q; defaults to TRUE

Code Cheat Sheet 💻

`plotNorm(mean = 0, sd = 1, shadeValues, direction, col.shade, ...)`

- **`mean`** refers to μ, defaults to 0
- **`sd`** refers to σ, defaults to 1
- **`shadeValues`** is a vector of up to 2 numbers that define the region you want to shade
- **`direction`** can be one of `less`, `greater`, `outside`, or `inside`, and controls the direction of shading between `shadeValues`. Must be `less` or `greater` if `shadeValues` has only one element; `outside` or `inside` if two
- **`col.shade`** controls the color of the shaded region, defaults to `"cornflowerblue"`
- **`...`** lets you specify other graphical parameters to control the appearance of the normal curve (e.g., `lwd`, `lty`, `col`, etc.)

Code Cheat Sheet 💻

```
prop_test(x, n, p = NULL, alternative =
c("two.sided", "less", "greater"), conf.level =
0.95)
```

- `x` is a vector of numbers of successes
- `n` is a vector of sample sizes
- `p` is is the null hypothesis value of p or the hypothesized difference in proportions
- `alternative` can be one of `less`, `greater`, or `two.sided`, and controls the direction of the alternative hypothesis. Defaults to `two.sided`, which must be used to make a confidence interval
- `conf.level` controls the confidence level used to make the confidence interval, must be a single number between 0 and 1.

Lab Project ⌨

Your tasks

- Complete the "Try It!" and "Dive Deeper" portions of the lab assignment by copy/pasting and modifying appropriate code from earlier in the document.

How to get help

- Use the "lab" tag on Piazza
- Email your lab instructor

Reminders

Your tasks for the week running Friday 10/30 - Friday 11/6:

Task	Due Date	Submission
M-Write 2 Initial Submission	Thursday 11/5 4:59PM ET	Canvas
Lab 10	Friday 11/6 8:00AM ET	Canvas
Homework 7	Friday 11/6 8:00AM ET	course.work