Chapter 3: Inference for a Single Proportion

The Goal:

To estimate or test the true population parameter value. In this case it is a **population proportion,** p.

The Approach (sampling distribution):

In order to either estimate or test a population parameter, we need to understand the **sampling distribution**, how we expect the sample statistic to behave. There are two approaches to creating this distribution of likely values:

1. Central Limit Theorem (based on mathematical theory)
2. Simulation in R (based on the observed sample, like we have done with hypothesis tests so far)

This document focuses on using the Central Limit Theorem.

Confidence intervals

Estimating p

We can estimate p using a **confidence interval**. We use the sample to *infer* a plausible range of values the population proportion could truly be. If we want to use the CLT and Normal distribution for the sampling proportion, we MUST make sure the conditions are met. Determine if:

- the sample has at least 10 "successes" and at least 10 "failures".
- it is reasonable to assume the observations are independent from one another.

If the above checks out we can describe the expected values for sample proportions and their variation as:

$$\hat{p} \sim N\left(\text{mean} = p, \text{SE} = \sqrt{\frac{p(1-p)}{n}}\right)$$

Read the above as: We expect sample proportions ($\hat{p}$) to be Normally distributed ($\sim N$). Sample proportions should take on values similar to the population proportion (mean $= p$), but will vary to some degree (SE $= \sqrt{\frac{p(1-p)}{n}}$).

Formula

The anatomy of a confidence interval:

$$\text{point estimate} \pm \text{margin of error}$$

$$\text{point estimate} \pm \text{critical value} \times \text{SE}$$

- **Point Estimate**: taken from the sample, $\hat{p}$
- **Critical Value**: how many SE's wide should the interval be for us to achieve the desired confidence. z^*

- **Standard Error (SE)**: how much uncertainty do we have in our sample. $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (remember that we don't know what p truly is, so we must estimate the it in the SE with $\hat{p}$)

  This means for estimating a population proportion use:

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Bonus: Calculate by hand (only a little bit of R)

Confidence interval

You can find the point estimate and standard error using the formulas above. But how do you find the critical value for a confidence interval?

The **critical value** is calculated using the standard normal distribution. This is why we need to check the CLT before estimating population proportion, to make sure it is appropriate.

Whatever the confidence percentage we are looking for (90%, 95%, 97.2354%, etc.), we find the cutoffs on the standard Normal distribution for that middle percentage.

Recall that standard normal has `mean = 0`, and `sd = 1`. We can find the cutoffs using `xqnorm` when we know the percentage.

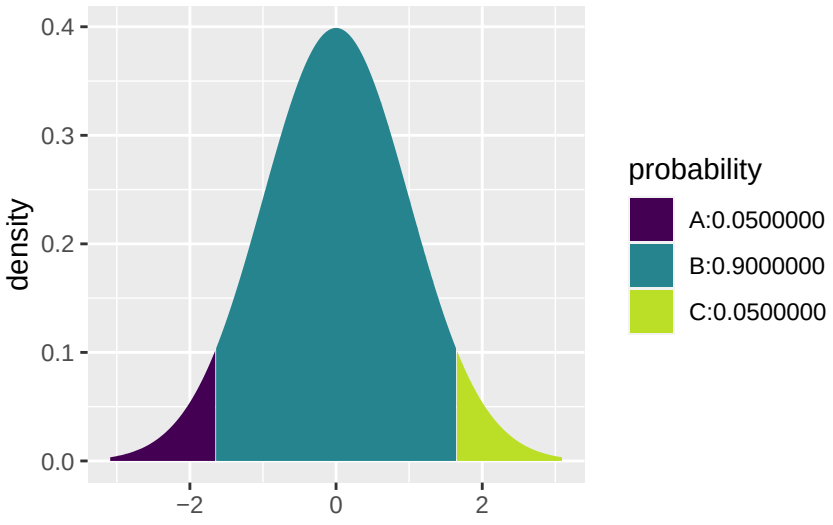
Find the cutoffs for the middle 90% on the standard normal distribution.

To get the middle 90%, we consider the 5th and 95th percentile (equal proportions in the tails).

```
# remember this function is in the mosaic package, it has already been loaded here

## the default for mean is 0 and for sd is 1, so we actually don't need to include
## these arguments
xqnorm(c(0.05, 0.95))
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -1.644854) = 0.05    P(X <=  1.644854) = 0.95
##  P(X >  -1.644854) = 0.95    P(X >   1.644854) = 0.05
##
```




```
## [1] -1.644854  1.644854
```

I can see that I need to go 1.645 standard deviation away from 0 on the standard Normal distribution to cover 90% in the middle. **So I need to go 1.645 *standard errors* away from the sample proportion to get my 90% confidence interval.**

Confidence %	Critical value
90%	$z^* = 1.645$
95%	$z^* = 1.96$
99%	$z^* =$???
94%	$z^* =$???
92.65%	$z^* =$???

Find the critical values for 99%,[1] 94%,[2] and 92.65%[3] Confidence Intervals

Example

A recent survey conducted by AT&T asked 1,200 teenagers aged 15-19 years old about their texting and driving habits. Among those surveyed, 43% admitted to texting while driving over the past three months. We are interested in estimating the true proportion of teens who text and drive in the U.S. Find a 99% confidence interval for this proportion.

Put the problem in context

parameter: $p = unknown$ - population proportion of all teens who text and drive in the U.S.

sample statistic: $\hat{p} = 0.43$ - sample proportion of teens who text and drive.

Check CLT

Can we use the Normal distribution to describe the distribution of sample proportions?

- Observations (teens in survey) were randomly contacted and surveyed, it is reasonable to assume their answers are independent from one another.

[1] 2.575
[2] 2.05
[3] 1.79

- At least 10 answered "yes" (success)? $0.43 \times 1200 = 516$
- At least 10 answered "no" (failure)? $(1 - 0.43) \times 1200 = 684$
  - Note that $1 - p$ is often referred to as the proportion of "failure".

Plug it in

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Plug in the rest of the pieces for the example:

Point Estimate: $\hat{p} = 0.43$

Critical Value: $z^* = 2.576$ (for 99% confidence)

Standard Error (SE): $\sqrt{\frac{0.43(0.57)}{1200}} = 0.0143$

$$0.43 \pm 2.576 \times 0.0143$$

$$0.43 \pm 0.0368$$

(margin of error is 0.0368)

$$[0.3932, 0.4668]$$

> We are 99% confident that the true proportion of all U.S. teens who text and drive is between 39% and 47%.

Our confidence comes from the fact that 99% of samples like this ($n = 1200$) should have $\hat{p}$'s close enough to the population parameter to capture the true value. There is a 1% chance that our observed statistic ($\hat{p}$) is extremely unusual which would make our interval inaccurate.

Calculating Sample Size

Sometimes we know the **margin of error (ME)** we would like to achieve in our confidence interval, but we don't know what sample size we should collect. We know:

$$ME = \text{critical value} \times \text{SE}$$

So for a single proportion, to calculate sample size for a desired ME:

$$ME = z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\frac{ME}{z^*} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\frac{ME^2}{(z^*)^2} = \frac{\hat{p}(1-\hat{p})}{n}$$

$$\frac{(z^*)^2}{ME^2} = \frac{n}{\hat{p}(1-\hat{p})}$$

$$n = \frac{(z^*)^2 \times \hat{p}(1-\hat{p})}{ME^2}$$

You can use a value from a previous study (if available) for $\hat{p}$, or use $\hat{p} = 0.5$ as a conservative guess (resulting in the largest possible sample size needed for desired ME).

Practice

What sample size would you need if you wanted a 95% confidence interval to be within 2% of the true population proportion of teen drivers who text?

You can use R to help with some of the calculations.

Answer

You should get $n = 2401$

$$n = \frac{1.96^2 \times 0.5(0.5)}{0.02^2} = 2401$$

When your sample size calculations are decimals, always round up to the next whole number to ensure you have enough.

In R

Much of this can be performed easily in R using the `prop.test` function. Simply use the following arguments:

```
prop.test(x, n, conf.level, correct)
```

Argument	Purpose	Default
`x`	the number of observed "successes"	no default
`n`	the sample size	no default
`conf.level`	the confidence you would like for your interval as a proportion `[0.0 to 1.0]`	`conf.level = 0.95`
`correct`	a special "correction" for when we don't have 10 success or 10 failures.	`correct = TRUE`

Note we will only calculate confidence intervals for samples that have enough successes and failures, so we will use `correct = FALSE`.

Estimate and Interpret

Since the CLT checks out alright, we can proceed with our estimation. In R, remember we need the number of successes (calculated in our CLT check), the sample size, and the percentage of confidence.

```
prop.test(x = 516, n = 1200, conf.level = 0.99, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  516 out of 1200
## X-squared = 23.52, df = 1, p-value = 1.236e-06
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.3936715 0.4670983
## sample estimates:
##    p
## 0.43
```

Interpretation

This provides a lot of output, but look for the `99 percent confidence interval:` section. Our confidence interval is [0.39, 0.47].

Very similar to the interval we calculated by hand (only off due to rounding).

Hypothesis tests

Simulation approach (how we have been doing it)

Consider a claim that 50% of teenage drivers text and drive. How would you test that the true rate is lower than this using simulation in R?

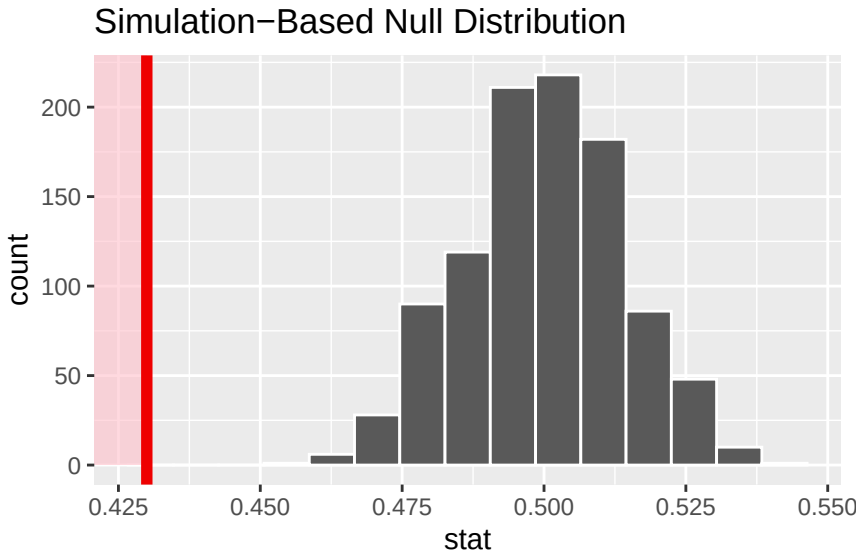
```
## remember these functions need the infer package, it has already been loaded

textingteens <- data.frame(survey_results = c(rep("text", 516), rep("do not text", 684)))

null_distn <- textingteens %>%
  specify(response = survey_results, success = "text") %>%
  hypothesize(null = "point", p = 0.5) %>%
  generate(reps = 1000, type = "simulate") %>%
  calculate(stat = "prop")

obs.prop = 0.43

null_distn %>%
  visualize(method = "simulation") +
  shade_p_value(obs.prop, direction = "left")
```




```
null_distn %>%
  get_p_value(obs_stat = obs.prop, direction = "left")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Theoretical Approach - in R

How would you test this claim using the Normal approximation method? In other words, instead of *simulating* the null distribution, we will use the CLT to *approximate* our null distribution.

Null distribution:

$$\hat{p} \sim N\left(\text{mean} = p, \text{SE} = \sqrt{\frac{p(1-p)}{n}}\right)$$

We can do this with the same code we used for the confidence interval. We just need to specify the direction of the alternative hypothesis.

`prop.test(..., alt = "less", "greater", or "two.sided")` (choose one of the 3 options)

```
prop.test(x = 516, n = 1200, conf.level = 0.99, alt = "less", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  516 out of 1200
## X-squared = 23.52, df = 1, p-value = 6.181e-07
## alternative hypothesis: true p is less than 0.5
## 99 percent confidence interval:
##  0.0000000 0.4634883
## sample estimates:
##    p
## 0.43
```

- Note the scientific notation of the p-value (6.1811e-07). This means the actual value is 6.1811×10^{-7} or 0.00000061811. We can just say the p-value ≈ 0.

Interpret

In both cases we observe a very small p-value (< 0.0001). Therefore we have very strong evidence to reject the null hypothesis. The sample indicates that the proportion of teen texters is significantly less than 50%.

Connection to Confidence Intervals

Notice that we could have also compared the claim with the confidence interval. The interval estimated the proportion between 39% and 47%. Since our claim falls outside of this interval, there is little chance that 50% is a plausible value for the population. **Comparing a claim to a confidence interval is equivalent to a *two-sided* hypothesis test of that claim, where the level of significance is 100 - (CI%).**

For example, 50% teen texting and driving is outside the 99% confidence interval we found, so we reject the hypothesis that $p = 0.50$ with level of significance, $\alpha = 0.01$ (since 100% - 99% = 1%) .

Bonus: Calculate by hand (only a little bit of R)

Hypothesis test

How would we conduct the hypothesis test by hand?

Hypotheses

$$H_0 : p = 0.50$$

$$H_a : p < 0.50$$

Check conditions of CLT - Under the Null Hypothesis.

Independence and would we **expect** at least 10 successes and 10 failures in the sample sample size if the null hypothesis were true ($p = 0.5$)?

- Independent observations is still reasonable
- $0.50 \times 1200 = 600$
- $0.50 \times 1200 = 600$

Standardize the observed statistic

Using the null proportion to see how unusual it is.

Under the null hypothesis:

$$\hat{p} \sim N\left(\text{mean} = p_0, \text{SE} = \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

In our example:

$$\hat{p} \sim N\left(\text{mean} = 0.5,\ \text{SE}_{p_0} = \sqrt{\frac{0.5(0.5)}{1200}} = 0.0144\right)$$

So to standardize the observed sample proportion into a test statistic, take the z-score:

$$Z - score = \frac{\text{obs-mean}}{sd}$$

$$\text{test stat } Z = \frac{\hat{p} - p_0}{SE_{p_0}} = \frac{0.43 - 0.5}{0.0144} = -4.861$$

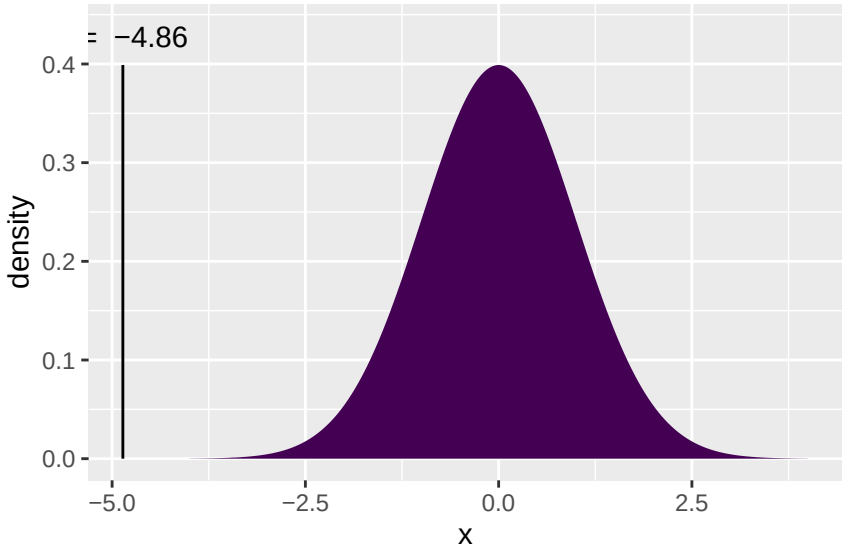
This means that our observed sample proportion, 43% of teen drivers text, **would be 4.861 standard deviations below the expected proportion of 50% if the null hypothesis was true**. From what we know about the normal distribution, anything that is more than 3 standard deviations away from the mean is *extremely* unlikely. But lets calculate the p-value to get a measure of the strength of our evidence.

Find p-value

Since the alternative hypothesis used "<", this is a left tailed test and we want to find the probability of getting another sample that is more than 4.861 standard deviation **below** the mean on the standard Normal distribution. $P(Z < -4.861)$

```
## Since the default mean = 0 and sd = 1, we don't need to specify these
xpnorm(-4.861, lower.tail = TRUE)
```

```
##
## If X ~ N(0, 1), then
##  P(X <= -4.861) = P(Z <= -4.861) = 5.84e-07
##  P(X >  -4.861) = P(Z >  -4.861) = 1
##
```




```
## [1] 5.839713e-07
```

Our test statistic is off the chart! Again, the p-value is very small (we expect the same result as previous methods), and we have strong evidence that the true proportion of teen drivers who text is less than 50%.