Statistical Inference

Learning about a population

Download the section 11 .Rmd handout to
`STAT240/lecture/sect11-inference`.

Download `ggprob.R` to
`STAT240/scripts`
if you have not already.

Random variables represent theoretical populations.

In statistics, we study a real population.

- We can't measure every single object
- We don't know the true probability distribution

We have to take a sample, which adds *uncertainty*.

A population has a **parameter of interest**.

- What % of citizens support a ballot measure?
- What is the true average price of apples?
- What is the median age of homeowners?

We can't measure these, so we find a corresponding value from a sample.

"53 of 100 Madisonians surveyed support the ballot measure."

- 53% is a good guess for the population level of support (p)

This is only true if the 100 people are *representative* of the city.

- Should be completely random
- Bigger sample is better

Two issues: **Sampling bias**.

- Cameron's shirts example

Sampling error.

- Bigger samples are more stable

We generated different guesses of p based on who happened to appear in our random sample.

The statistic $\hat{p}$ (sample proportion) **estimates** p (population proportion).

What are all the possible values of $\hat{p}$?

A **sampling distribution** is the probability distribution of a statisitc.

Our inference is more rigorous:

- Does $\hat{p}$ do a good job of estimating p?
- What is the error we incur from sampling?
- Can we get a range of guesses for p rather than a single number?

Different statistics are used in the context of different parameters of interest.

- Sample proportion $\hat{p}$ estimates p
- Sample mean $\bar{X}$ estimates μ
- Sample variance S^2 estimates σ^2

Each sample quantity has its own sampling distribution.

0.53 is a **point estimate** for the true proportion.

It can be useful to instead find an **interval estimate** that represents a range of “guesses”.

We want our interval to be precise, but we also want to actually cover the true parameter.

We want our interval estimate to be centered at $\hat{p} = 0.53$.

$$(0.53 \ - \ \text{Margin}, \ 0.53 \ + \ \text{Margin})$$

What should the margin be?

What if we used a margin of 0.2? We guess that the true proportion of supporters is

$$0.53 \pm 0.2 \ = \ (0.33,\ 0.73)$$

More generally:

$$\hat{p} \pm 0.2 \ = \ (\hat{p} - 0.2,\ \hat{p} + 0.2)$$

We can find the probability that this "margin 0.2" interval covers p.

$$P(\hat{p} - 0.2 \ < \ p \ < \ \hat{p} + 0.2) \ = ?$$

What if the probability is small? We a large coverage probability, and an arbitrary margin of 0.2 won't always work.

Instead, we *choose* a coverage probability, and solve for the margin.

$$P(\hat{p} - ? \ < \ p \ < \ \hat{p} + ?) \ = \ 1 - \alpha$$

A **confidence interval (CI)** is an interval estimate with a specific coverage probability $1 - \alpha$.

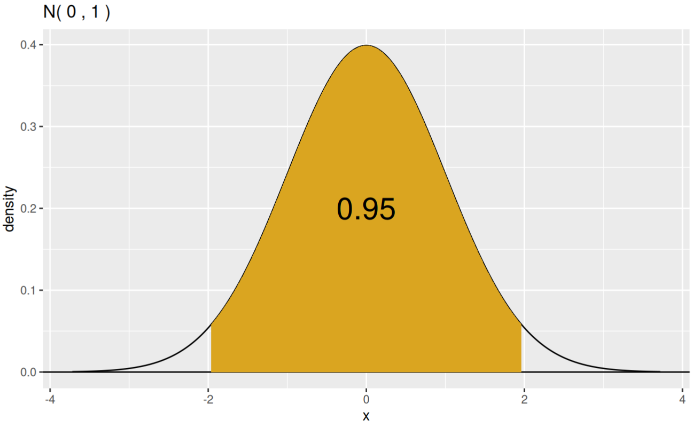
Let's make a 95% CI, with $\alpha = 0.05$. Our interval will look like

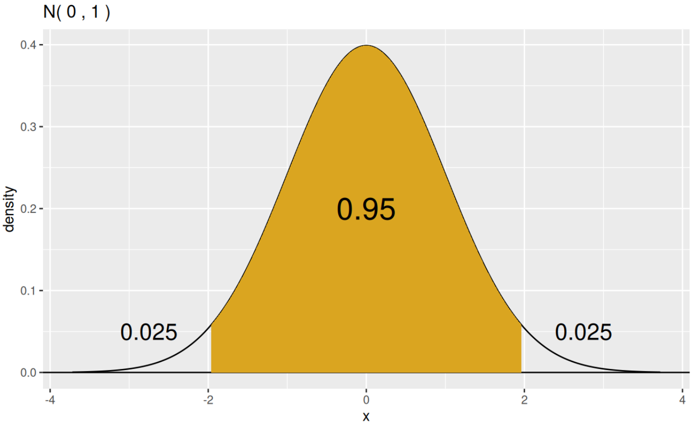
$\hat{p} \pm$ (value related to 0.95) $\times$ (estimation error of $\hat{p}$)

We'll see in a future unit that the sampling distribution of $\hat{p}$ is approximately

$$\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

We can use a value on the normal distribution to guarantee a coverage probability of 0.95.





The area in the middle is 0.95, and the area outside is 0.05. Each tail has area 0.025.

So, the values that guarantee a coverage probability of 0.95 are the 2.5 and 97.5 percentiles of $N(0, 1)$.

With `qnorm`, we find these to be -1.96 and 1.96. These are the **critical values**.

Due to the symmetric nature of $N(0, 1)$, we can just use 1.96.

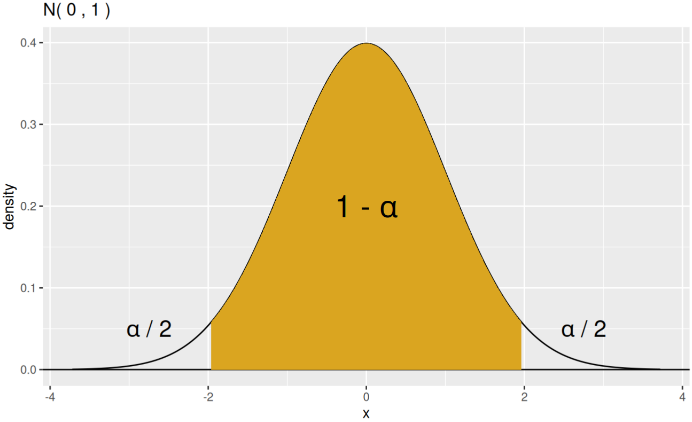
Our 95% CI for p is

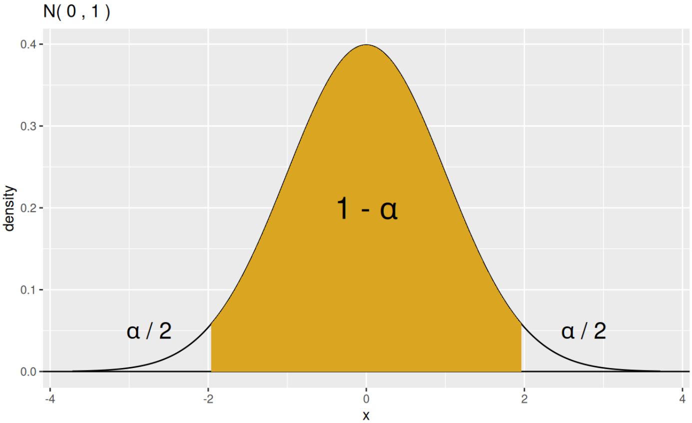
$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad = \quad 0.53 \pm 1.96\sqrt{\frac{0.53(0.47)}{100}}$$

$$(0.432, 0.628)$$

We are 95% confident that the true proportion of supporters is within (0.432, 0.628).

We can generalize this to other α values.





The center is our coverage probability, and each tail has area $\alpha/2$. What if we wanted 90% confidence?

The general formula for a CI is

point estimate $\pm$ critical value $\times$ standard error

The critical value comes from α, and the other terms come from our data.

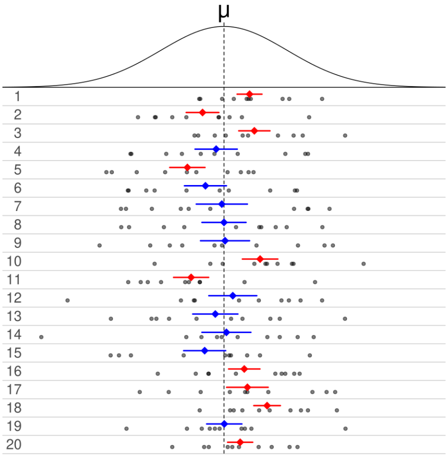
The margin is the **margin of error**.

Once a CI is realized, it either covers p, or not.

Across all samples, 95% of the CIs would cover p.

The CI procedure has a 0.95 success rate. We say we have 95% *confidence* ($100(1-\alpha)\%$ confidence) that an interval covers p.

50% CIs, from Wikipedia:



What is a good confidence level?

A standard choice is $\alpha = 0.05$, which results in a 95% confidence level. Other common choices are 90%, 98%, and 99%.

A smaller α results in a wider interval.

In statistical **testing**, we make a guess about the value of a parameter.

Then, we collect data and see if the data is consistent with our guess.

This is different from CIs, which use the data to form an estimate.

Researcher Muriel Bristol claimed she could tell whether tea had been added before or after milk to her cup, just by tasting it.

Is her ability to identify milk-first versus tea-first cups better than random guessing? ($p = 0.5$).

Background: Lady tasting tea

Let's make some assumptions:

- She gets each cup correct or incorrect
- Each cup is independent
- The number of cups is specified as $n = 8$
- $P($ Correct $)$ is a fixed p

Under these assumptions, the number of correct guesses is $X \sim Binom(8, p)$.

We formalize our question with **hypotheses**. The **null hypothesis** (H_0) is the "baseline" result.

$$H_0 : p \leq 0.5$$

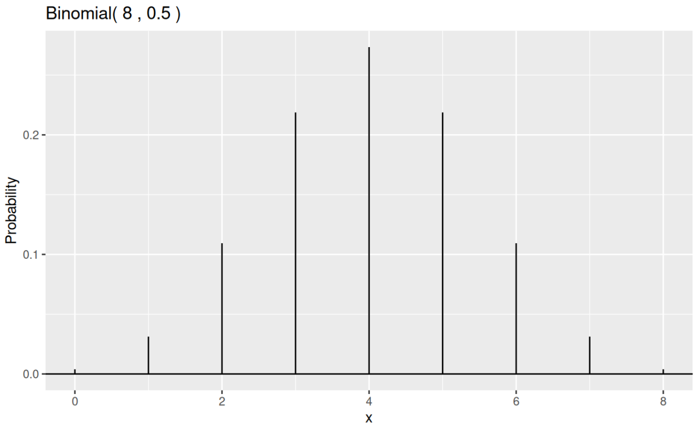
The **alternative hypothesis** (H_A) is the "interesting" result, covering all other cases.

$$H_A : p > 0.5$$

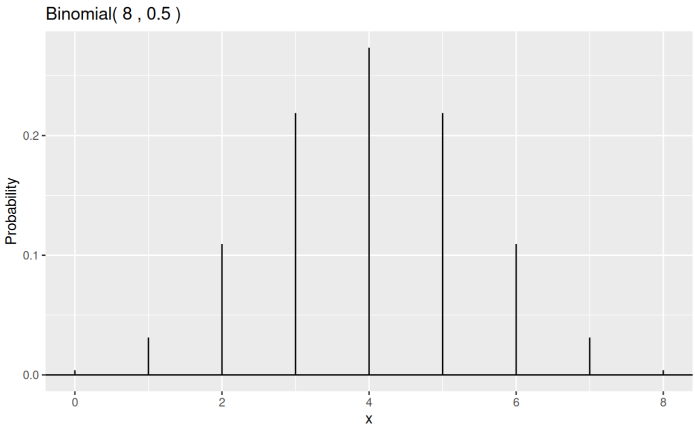
Start by assuming H_0 is true (Bristol is guessing).

We then use our data to collect evidence *against* the null. Does our data contradict H_0?

Instead of proving $p > 0.5$, show that $p \leq 0.5$ is very unlikely.



If the null is true, then the number of correct guesses is $Binom(8, 0.5)$.

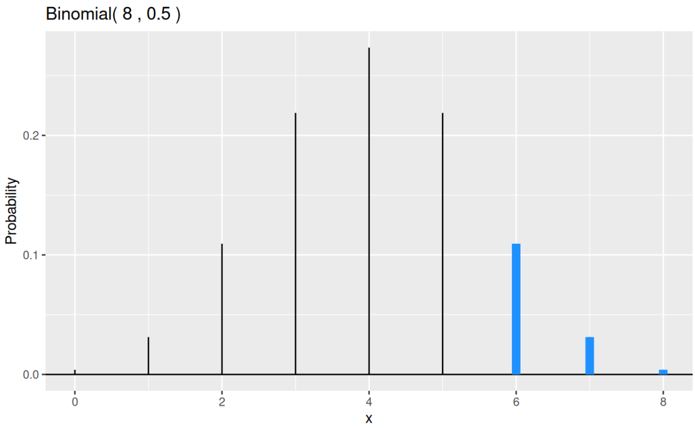


This is the **null distribution**, which represents H_0 being true.

For example, say Bristol got 6 guesses right out of 8.

What is the probability that she gets 6 or more right *by random chance*?

This is a probability on our null distribution.

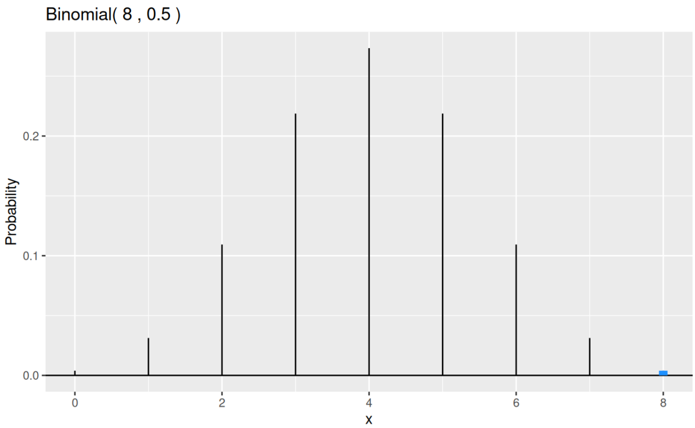


$$P(X \geq 6) = 0.145$$

This probability is called a **p-value**. What is the probability of observing our data or something more extreme, under H_0?

Smaller p-values mean stronger evidence against H_0.

We choose a threshold for rejection, the significance level α. A common choice is $\alpha = 0.05$.



In reality, Bristol got all 8 cups correct.

Our p-value is $P(X \geq 8) = 0.004$, which is much smaller than $\alpha = 0.05$.

We have strong evidence that Bristol can tell whether milk or tea is poured first at a better rate than random chance.

We have strong evidence against H_0.

If the test indicates that the null is likely false, we **reject** H_0.

If we don't have enough evidence against the null, then we **fail to reject** H_0.

It is NOT accurate to "accept" the null hypothesis, since the test does not guarantee that H_0 is true.

CIs and hypothesis tests are general methods that can be used in a variety of situations.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The parameter β_1 describes the linear relationship between X and Y.

We can study a single mean μ, or a difference in two means:

$$\mu_1 - \mu_2$$

Or a difference in proportions:

$$p_1 - p_2$$

Each parameter has its own sampling distribution and standard error.