SOC6302 Statistics for Sociologists

Monica Alexander

Week 6: Introduction to statsitical inference

Announcements

- A2 out
- Has some practice questions for exam (not to hand in)
- Mid-term after reading week
  - Paper and pen
  - No books, computers, calculators
  - Will be provided any formulas you need
  - Short answer and multiple choice
  - Interpretation of code, graphs
  - Short questions based on formal concepts covered in class

Sampling distributions

A **sampling distribution** is a probability distribution for a statistic based on repeated samples.

Say we are interested in taking a random sample of people's heights, X and calculating the mean height for that sample. So our statistic of interest is the mean height, $\bar{X}$.

Randomly sampling heights

We first take a random sample of 12 people and get the following heights

```
##  [1] 168.5 166.7 169.5 181.3 173.7 171.3 169.0 156.3 179.6 164.6 169.4 153.4
```

The observed mean height of this sample is 168.6.

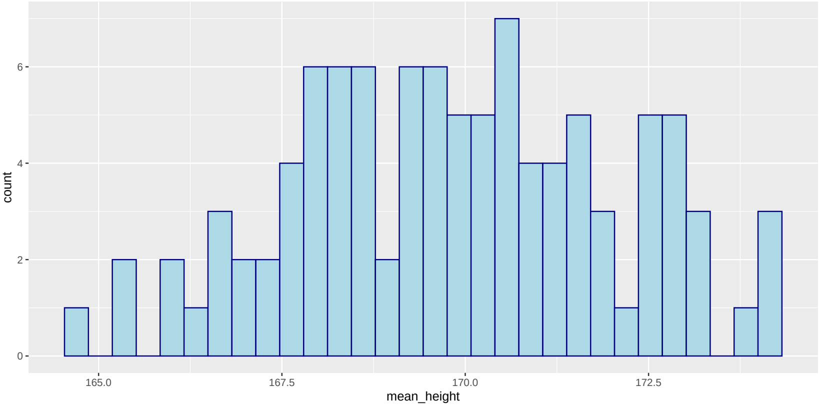
We take a random sample of another 12 people and get the following heights:

```
##  [1] 166.3 160.9 168.9 178.1 163.1 178.1 181.5 173.4 164.5 174.4 165.0 175.5
```

The observed mean height of this sample is 170.8.

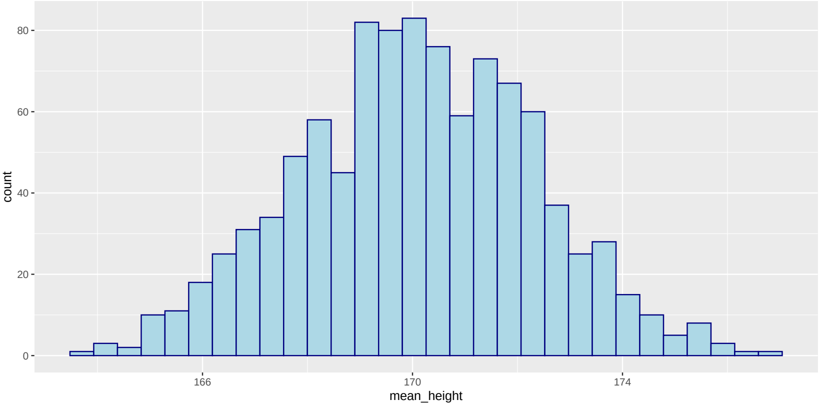
Randomly sampling heights

Say that I keep doing this process again and again and again, and end up with 100 observations of mean height. I can plot a histogram of these means:



Randomly sampling heights

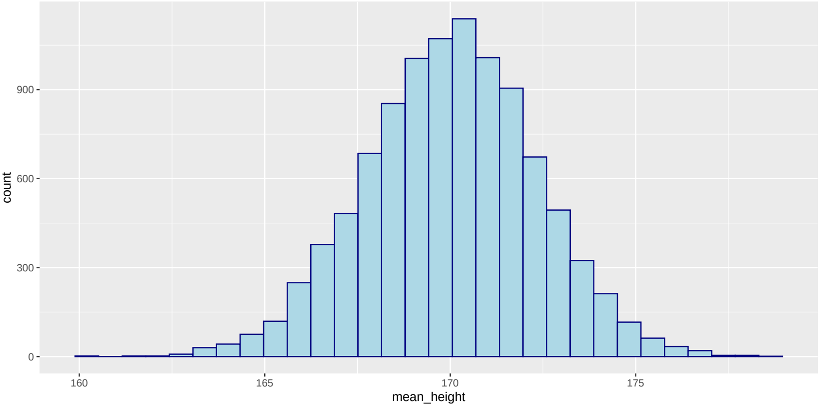
Okay, what if I take 1000 samples. I plot a histogram of the means again



What do you notice?

Randomly sampling heights

What about 10,000 samples?



The central limit theorem

The distribution of the sum (or mean) of a set of independent random variables will **tend towards** a normal distribution.

- "tend towards" means as the number of observations of the sum or mean gets larger, the distribution will become more normal
- The central limit theorem holds even if the original variables themselves are not normally distributed.

The central limit theorem

For a random variable X with $E(X) = \mu$ and $Var(X) = \sigma^2$, the central limit theorem results in the following distribution for the mean $\bar{X}$:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

What does this mean?

- The mean $\bar{X}$ will be centered at the same value as X
- The variance of $\bar{X}$ depends on the variance of the original random variable X and also the number of samples of the mean we have, n.

The quantity $\frac{\sigma}{\sqrt{n}}$ is also called the **standard error of the mean**.

Sampling distributions for proportions

Sampling distribution for proportions

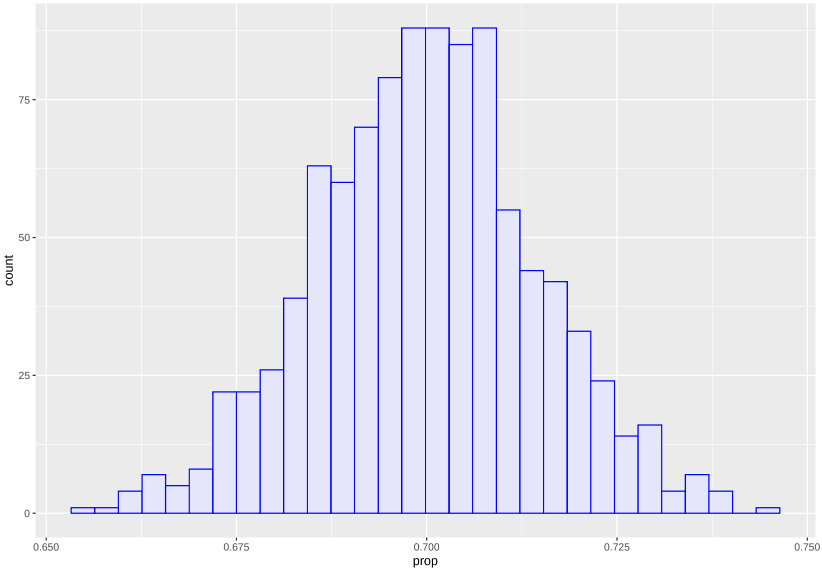
Say I'm interested in estimating the proportion of University of Toronto students who like ice cream.

- I sample 100 students and ask them whether they like ice cream (Y/N). Here are the results:

```
##   [1] "Y" "N" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "N" "Y" "N" "N" "N"
##  [19] "Y" "Y" "Y" "Y" "Y" "Y" "Y" "N" "Y" "N" "Y" "Y" "N" "Y" "Y" "Y" "Y" "Y"
##  [37] "Y" "N" "N" "Y" "N" "Y" "Y" "N" "Y" "N" "Y" "Y" "Y" "Y" "Y" "N" "Y" "Y"
##  [55] "Y" "Y" "Y" "N" "Y" "Y" "N" "Y" "Y" "Y" "Y" "N" "Y" "Y" "Y" "Y" "Y" "Y"
##  [73] "Y" "N" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "Y" "N" "Y" "Y" "Y" "Y"
##  [91] "Y" "Y" "N" "Y" "Y" "Y" "Y" "Y" "Y" "Y"
```

- How many people said yes? 80. So an estimate of the proportion is 0.8
- Notice that this a mean. The implication is that estimates of the proportion of students who like ice cream will follow the central limit theorem.
- Let's take 1000 more samples of students and plot the resulting proportion estimates

Sampling distribution for the proportion



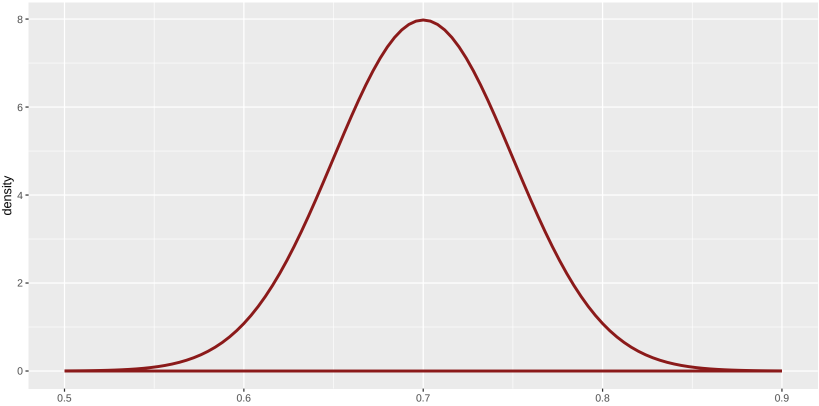
CLT for proportions

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error:

- Mean $\mu = p$
- Standard error $SE = \sqrt{\frac{p(1-p)}{n}}$

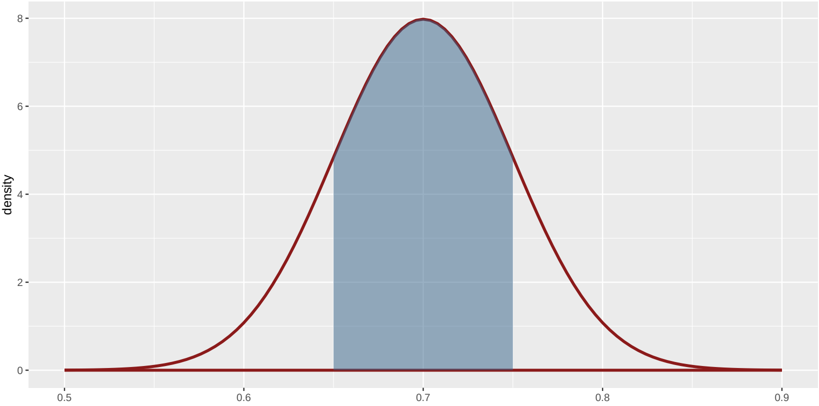
In the previous example, I generated sample proportions using $p = 0.7$ and $n = 100$. So the standard error of the proportion is 0.05.

Sample distribution for proportions



Question: How often will the proportion be within 0.05 of the truth (0.7)?

Sample distribution for proportions



Back to Z-scores

$$Z_{0.65} = \frac{0.65 - 0.7}{0.05} = -1$$

$$Z_{0.75} = \frac{0.75 - 0.7}{0.05} = 1$$

Recall how much of a probability distribution is within 1 standard deviation? $\sim$ 68%

Standard error for sample proportion

- In practice we don't know the true population proportion p
- We can use the substitution principle and use our estimate $\hat{p}$ in place of p in the standard error calculations, i.e.

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

In the ice cream example, our mean estimate of the sample proportions was 0.7, so the estimate of the standard error is the same as before.

Confidence intervals

Confidence intervals

- The sample proportion $\hat{p}$ provides a single plausible value for the population proportion p.
- However, the sample proportion isn't perfect and will have some standard error associated with it.
- We can think about, instead of providing just one estimate of the population proportion, providing a range of likely values
- This captures the measure of central tendency, but also spread (standard error) based on our sampling

95% Confidence interval

- In a normal distribution, 95% of the data is within 1.96 standard deviations of the mean (check you know how to calculate this!)
- As such we can construct a confidence interval that extends 1.96 standard errors from the sample proportion to be 95% confident that the interval captures the population proportion:

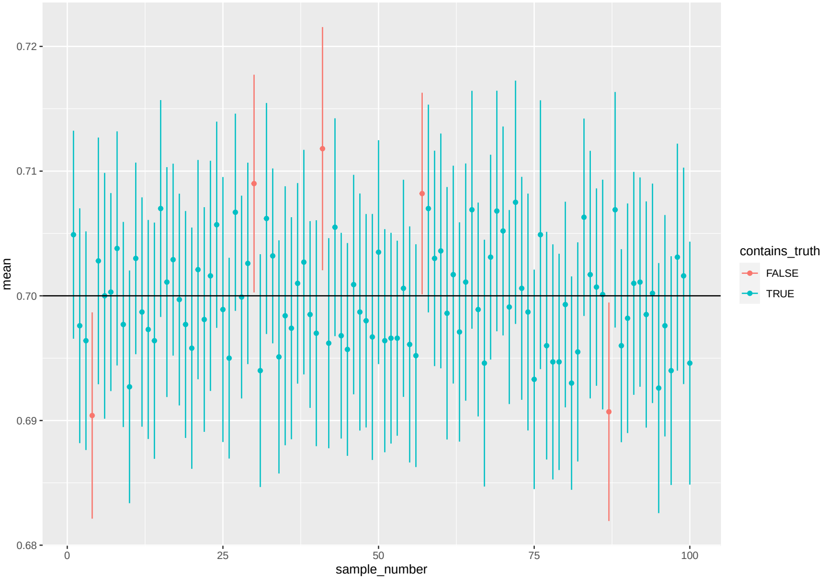
$$\text{point estimate } \pm 1.96 \times SE$$

$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

What does a 95% confidence interval actually mean?

Suppose we took many samples and built a 95% confidence interval from each. Then about 95% of those intervals would contain the parameter, p

What does a 95% confidence interval actually mean?



95% confidence interval for ice-cream

We estimate that the proportion of students who like ice-cream is $0.7 \pm 1.96 * 0.05 \implies 0.07(0.602, 0.798)$.

Interpretation: if we repeatedly sampled students over and over, 95% of the time the interval (0.602, 0.798) would contain the true proportion of people who liked ice cream.

Error rate

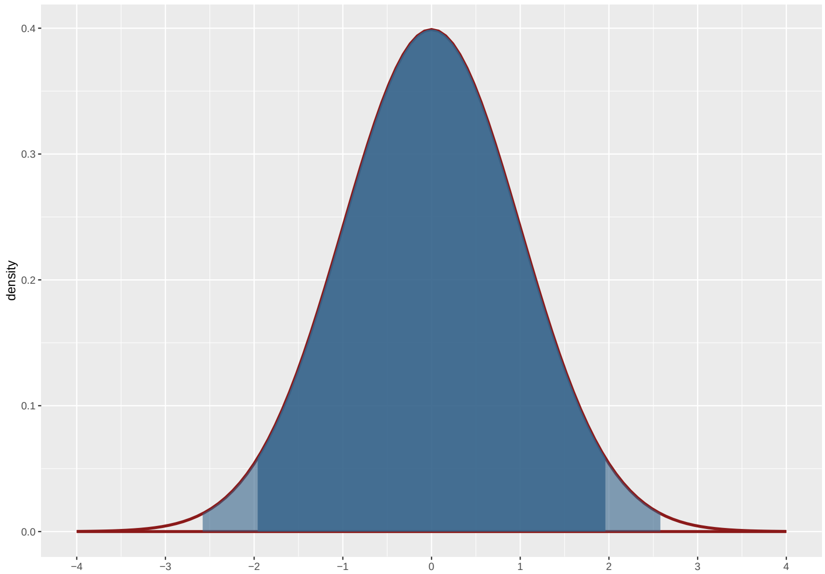
- "Interpretation: if we repeatedly sampled students over and over, 95% of the time the interval (0.602, 0.798) would contain the true proportion of people who liked ice cream."
- The implication is that 5% of the time we would be wrong (more later).

Changing the confidence level

A 95% confidence interval is

$$\text{point estimate } \pm 1.96 \times SE$$
$$\hat{p} \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

What do we change to get a different level of confidence?


density
0.4
0.3
0.2
0.1
0.0
−4
−3
−2
−1
0
1
2
3
4

Changing the confidence level

If a point estimate closely follows a normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate } \pm z^{\star} \times SE$$

where $z^{\star}$ corresponds to the confidence level selected.

Example

Calculate a 99% confidence interval for the proportion of people who like ice cream.

- point estimate $= 0.7$
- SE $= 0.05$
- $z^\star =?$

```
qnorm(0.025)
```

```
## [1] -1.959964
```

```
qnorm(0.005)
```

```
## [1] -2.575829
```

t-distribution

Back to sample means

Recall that the CLT for samples means suggests that we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

- Mean: μ
- Standard error: $\sqrt{\frac{\sigma}{n}}$

An approximation

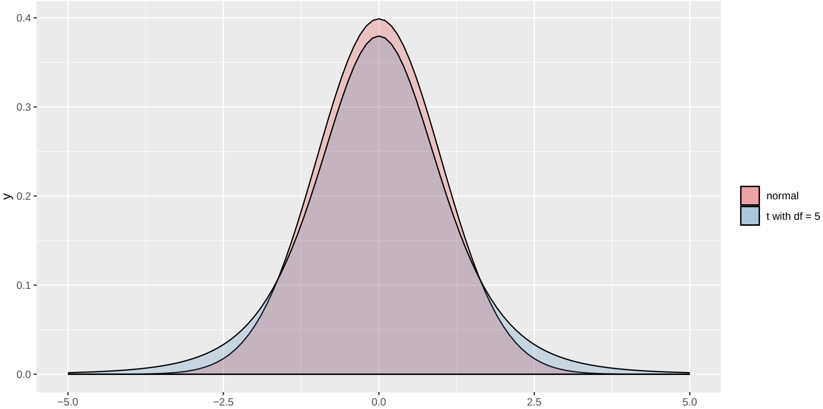
The standard error is dependent on the population standard deviation, σ. However, we rarely know σ, and instead we must estimate it. In particular we use

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation. This strategy tends to work well when we have a lot of data and can estimate σ using s accurately. However, the estimate is less precise with smaller samples.

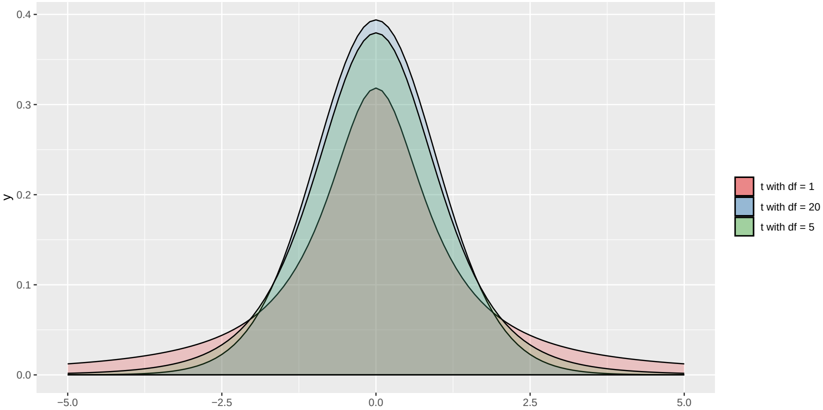
t-distribution

For inference of sample means, we'll use the t-distribution. It has a bell-shape like the normal distribution, but extra thick tails to correct for the problem of using s in place of σ in the SE calculation. (why?)



t-distribution

- The t-distribution is always centered at zero and has a single parameter: degrees of freedom.
- The degrees of freedom (df) describes the precise form of the bell-shaped t-distribution.

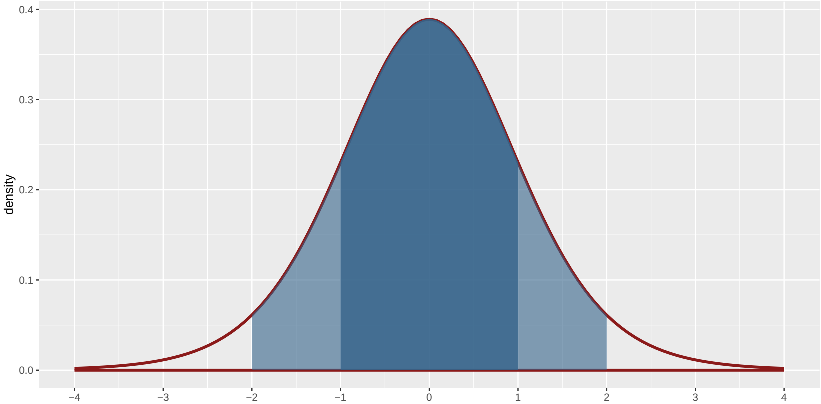


Degrees of freedom

- The bigger the df, the closer the t is to a standard normal distribution
- When the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution.

Calculating areas

- Just like we could calculate probabilities as areas under the normal distribution curve, we can do the same thing for the t-distribution



Calculating areas

```r
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

```r
pt(1, df = 5) - pt(-1, df = 5)
```

```
## [1] 0.6367825
```

E.g proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

```r
pt(-2.1, df = 18)
```

```
## [1] 0.0250452
```

One sample t confidence intervals

- Example: I'm interested in the average amount of ice-cream (by weight) that University of Toronto students have eaten in the past week.
- I randomly sample 19 students have the following information:
  - the sample mean $\bar{x}$ is 400g
  - the sample standard deviation s is 20g

Construct a 95% confidence interval for average ice-cream consumed.

$$\bar{x} \pm t^{\star}_{df} \times \frac{s}{\sqrt{n}}$$

What other information do we need?

Critical *t* value

```r
qt(0.025, df = 18)
```

```
## [1] -2.100922
```

Answer

```r
n <- 19
x_bar <- 400
se <- 20/sqrt(n)
df <- n - 1
t_star <- qt(0.025, df = 18)

x_bar - t_star*se
```

```
## [1] 409.6397
```

```r
x_bar + t_star*se
```

```
## [1] 390.3603
```

Hypothesis testing

Motivation

Say that I'm interested to know whether UofT student's consumption of ice-cream is changing over time. I know that the average amount consumed in 2019 was 380 grams. We want to determine using the data collected above whether students are consuming more or less ice cream, versus the other possibility of no change.

What do you think, intuitively?

Hypothesis testing framework

We want to test the hypothesis

- H_0: no change in ice-cream consumption

versus

- H_A: some change in ice-cream consumption

"H-naught" versus "the alternative hypothesis". The null hypothesis (H_0) often represents a skeptical perspective or a claim to be tested. The alternative hypothesis (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values. Our job as data scientists is to play the role of a skeptic: before we buy into the alternative hypothesis, we need to see strong supporting evidence.

How to test for evidence against H_0: intuition

We need to calculate the T-score (like the Z-score, but now we are using t-distributions)

$$T = \frac{\bar{x} - \mu}{SE_{\bar{x}}}$$

- Under the null hypothesis, the 2023 observation of ice-cream consumption would come from the same distribution as the 2019
- So under H_0 the difference between the two means should be close to zero, with some variation around that
- The further the difference is away from zero, the less likely that the two observations come from the same distribution

How to test for evidence against H_0: intuition

$$T = \frac{\bar{x} - \mu}{SE_{\bar{x}}}$$

This means

- **Under** H_0 the T-score will be t-distributed with a certain df (with the df depending on sample size)
- If the T-score is close to zero, it's likely the two observations come from the same distribution
- If T-score is big, it's likely they aren't

How to test for evidence against H_0: steps

We are testing $H_0 : \bar{x} = \mu$ versus $H_0 : \bar{x} \neq \mu$. This is a **two-sided hypothesis test**

- Gather up the info you need
- Calculate T-score
- Calculate p_value, represents the probability of observing such an extreme sample proportion by chance, if the null hypothesis were true.
- In this case this is just calculating $P(t \leq -T)$ and doubling it

Ice cream example

```r
mu <- 380
x_bar <- 400
se <- 20/sqrt(n)
df <- n - 1

T_score <- (x_bar-mu)/se
T_score
```

```
## [1] 4.358899
```

```r
pt(-1*T_score, df = 18)*2
```

```
## [1] 0.0003783575
```

```r
#alternatively
(1-pt(T_score, df = 18))*2
```

```
## [1] 0.0003783575
```

```r
#or
1-pt(T_score, df = 18)+pt(-1*T_score, df = 18)
```

```
## [1] 0.0003783575
```

What do we conclude?

Decision errors

- Hypothesis tests are not flawless: we can make an incorrect decision in a statistical hypothesis test based on the data.
- In particular, the **Type-I error** is rejecting the null hypothesis when H_0 is actually true
- We have to set a significance level α, which is our threshold for the rate of Type-I errors we are comfortable with
- We compare the p-value to the significance level. It is often set to $\alpha = 0.05$.

Since in the ice-cream case the p-value is less than α, we reject the null hypothesis. That is, the data provide strong evidence against H_0 and we believe instead that ice-cream consumption has changed.

Summary

- Sampling distributions for proportions
- Confidence intervals
- Calculating SE with sample standard deviation
- t-distributions
- Hypothesis testing for one sample means