STAT 102: Week 10

Ricky’s Section

Introductions and Attendance

Introduction: Name

Question of the Week: If you could add one consistent item to HUDS, what would it be?

Important Reminders

My Office Hours

- This week, changed to Sat, 04/05 from 2:30 to 4:30 PM
- Next week, changed to Fri, 04/11 from 8 to 10 PM
- Slack me if you have a question!

Content Review: Week 10

A (Quick) Review of Probability and Random Variables

- **Probability**: A value between 0 and 1
- **Random variable**: An unknown value that "crystallizes" to a certain number AFTER an **experiment**
    - A **discrete r.v.** can crystallize to countable numbers *(ex: 1, 2, 3)*
    - A **continuous r.v.** can crystallize to any real number in an interval *(ex: $-\sqrt{2}$, π, 102.74012…)*
- **Probability distributions**: Functions that describe a **r.v.** through its **probabilities**
    - For **discrete r.v.s**, we use **PMFs**, which give the probability of the r.v. crystallizing to any specific number
    - For **continuous r.v.s**, we use **PDFs**, which are shapes whose area represents **probability** (thus giving the **probability** of the **r.v.** crystallizing to any number within a specific interval)

What is an example of a discrete r.v.? A continuous r.v.?

Question:

What is an example of a discrete r.v.? A continuous r.v.?

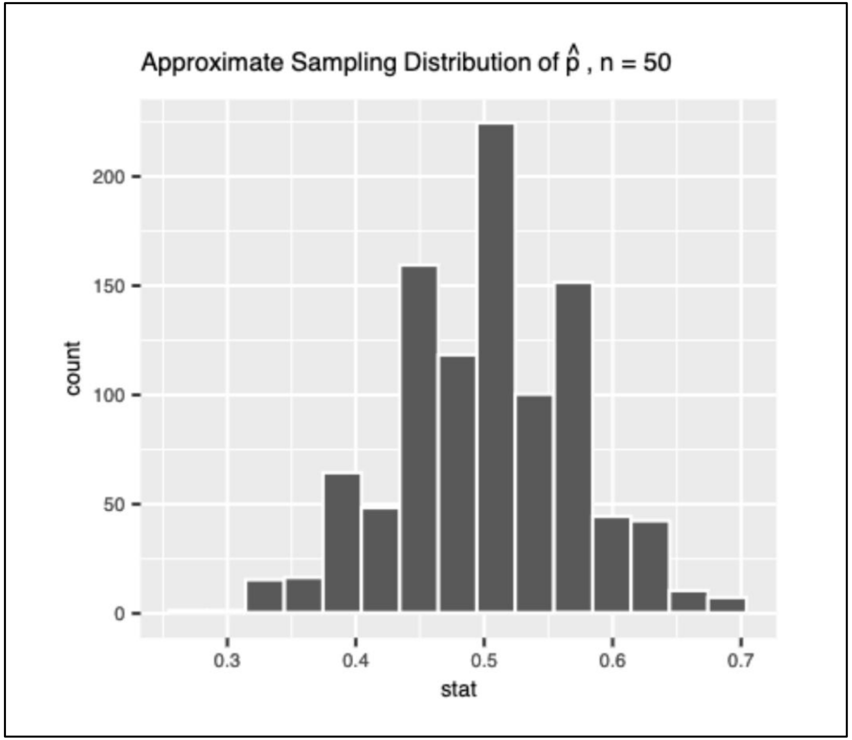
There are a bunch of different examples! The number of students in this room is **discrete** while a student's height is **continuous**.

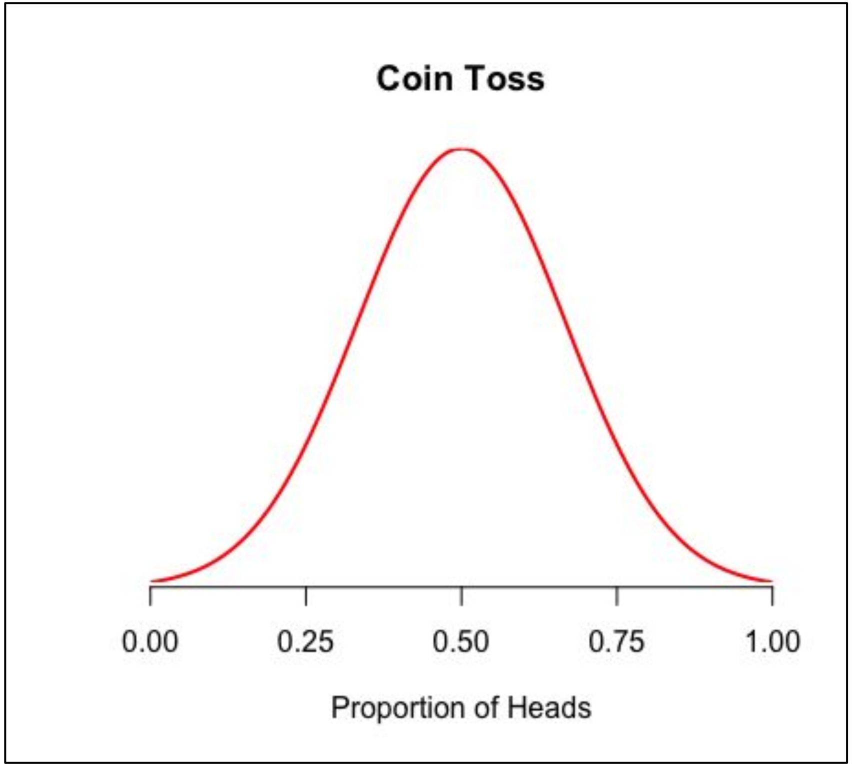
Moving forward, we'll be working mostly with **continuous r.v.s** (because we're recasting our **sample statistics** as **continuous r.v.s**).

A (Quick) Review of Probability and Random Variables

- **Normal r.v.**: X ~ N(μ, σ)
  - μ = mean, σ = SD
  - We use this for **proportions**
- **Standard Normal r.v.**: X ~ N(0, 1)
- ***t* r.v.**: X ~ t(df)
  - df = degrees of freedom
  - We use this for **means** and **linear regression**
- **Central Limit Theorem (CLT)**: For **random samples** and a **large sample size**, the **sampling distribution** of many **sample statistics** is approx. **Normal**
  - When assumptions are met, we can conduct **inference** using the **Normal distribution** as a good approximation

Null Distributions: Simulation-Based vs. Theory-based





A Visual Intuition for Central Limit Theorem

https://drive.google.com/file/d/128kvCSzPjRL7NMTtDRYlPAAY05RW7d3x/view?usp=drive_link

Theory-Based Inference

- Let's recast our **sample statistics** as **random variables**
- According to the **CLT**, when **assumptions** are met...
  - $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$, where p = population proportion
  - $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, where μ = population mean and σ = population SD
- We often **standardize** our **sample statistic** to use **z-score** as our **test statistic**
  - This is because **Standard Normal dist.** is easy to use as our **Null dist.**
  - $\text{z-score} = \frac{X - \mu}{\sigma}$, where μ = population mean and σ = population SD

As a quick sanity check, why does it make sense to recast our sample statistics as random variables? Hint: Consider sampling variability and the "mystery box" intuition.

Question:

As a quick sanity check, why does it make sense to recast our sample statistics as random variables? Hint: Consider sampling variability and the “mystery box” intuition.

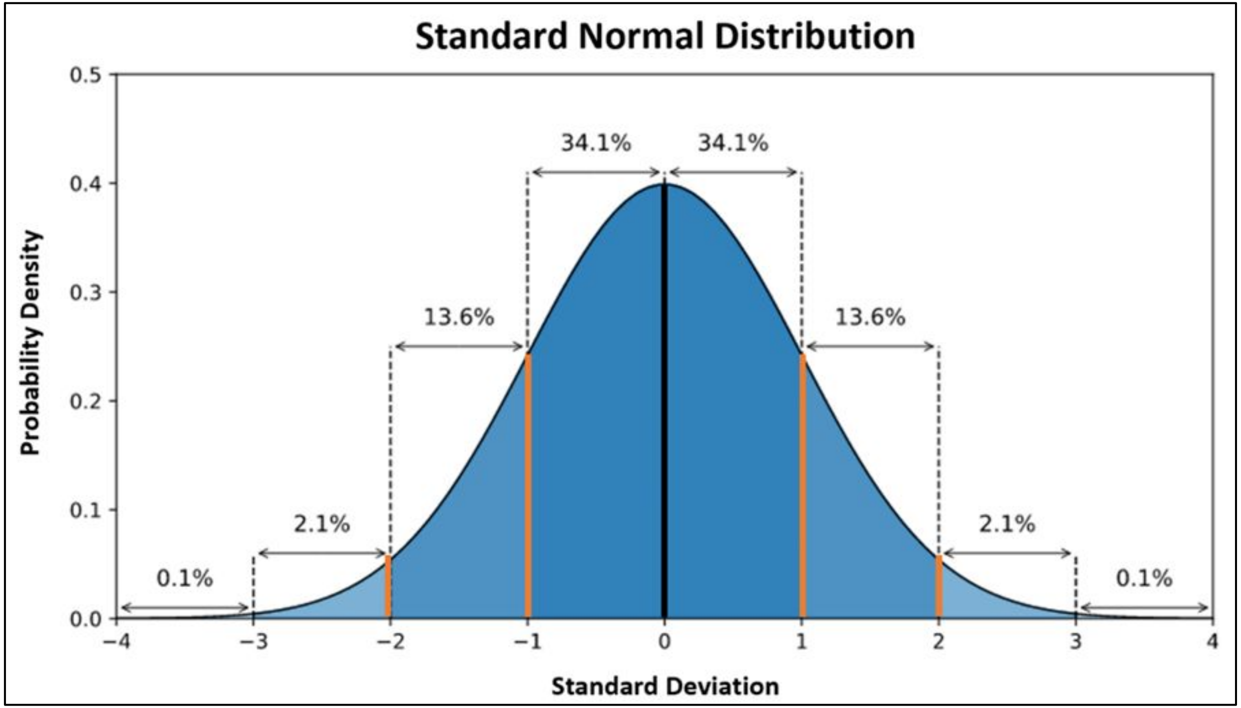
Due to **sampling variability**, **sample statistics** often differ from one another. For example, if I survey 400 people, my $\hat{p}$ would look different from yours if you surveyed 400 different people.

Thus, we can think of the **sample statistic** as a “mystery box” that will crystallize to a certain value after our sampling.

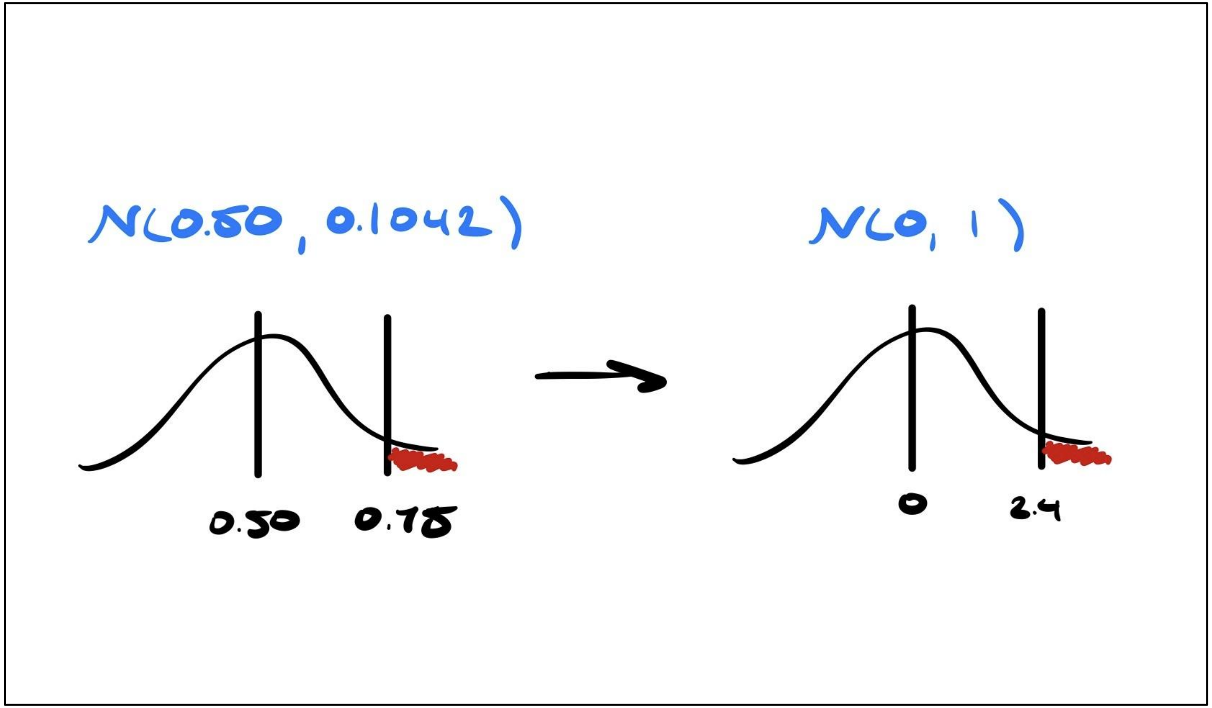
More on Test Statistic and Z-Score

- Up to now, we've been using our **(observed) sample statistic** as our **test statistic**
  - *"The prob. we get our observed test stat. of 75% heads (or more extreme) is…"*
- We can also use **z-score**, which is a standardized version of the **sample statistic**
  - *"The prob. we get a z-score of 2.4 (or more extreme) is…"*
  - It measures how many SDs the **sample statistic** is away from its **mean**
  - If **sample statistic ~ N(μ, σ)**, then **z-score ~ N(0, 1)** (**Standard Normal**)

Standard Normal Distribution

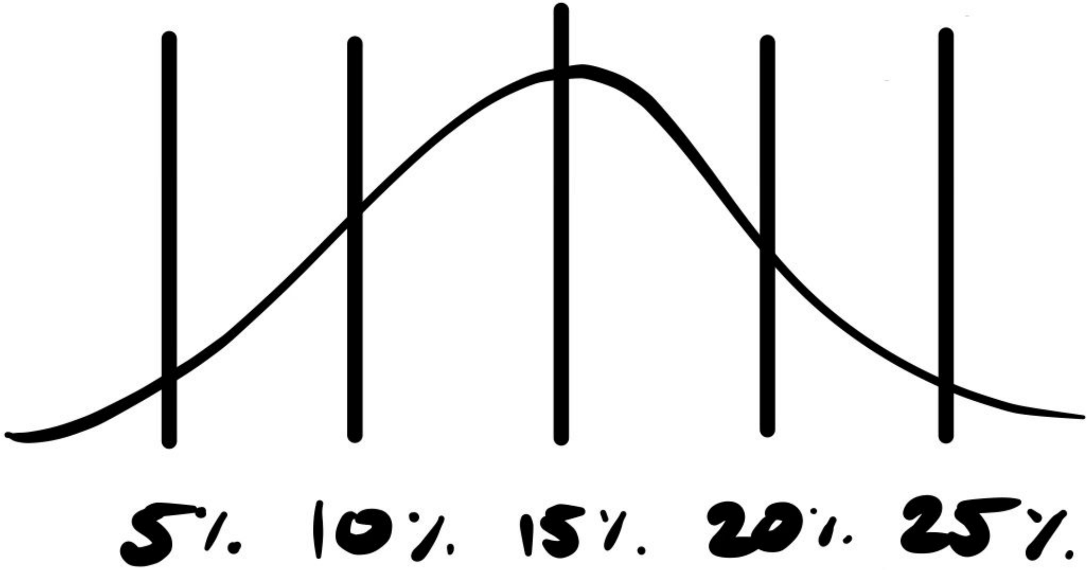


A Visual Intuition for Standardizing



If $\hat{p} \sim N(15\%, 5\%)$ and I get a sample with $\hat{p} = 25\%$, what is its z-score, and what does it mean?

$$\hat{p} \sim N(15\%, 5\%)$$



Question:

If $\hat{p} \sim N(15\%, 5\%)$ and I get a sample with $\hat{p} = 25\%$, what is its z-score, and what does it mean?

We're recasting our **sample statistic ($\hat{p}$)** as a **continuous r.v.**

We're given $\hat{p} \sim N(15\%, 5\%)$. According to **CLT**, when assumptions are met, $X \sim N(\mu, \sigma)$. Thus, mean = 15%, and SD = 5%.

z-score = $(X - \mu)/\sigma$, so z-score = 2. We see 25% is 2 SDs away from 15%.

Theory-Based Hypothesis Tests (for Proportions)

- According to **CLT**, under the **H_0**, $\hat{p} \sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}})$
  - Remember $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$
- Our **z-score (test statistic)** follows a **standard normal distribution**
  - z ~ N(0, 1)
- $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
  - Remember z-score = $(X - \mu)/\sigma$

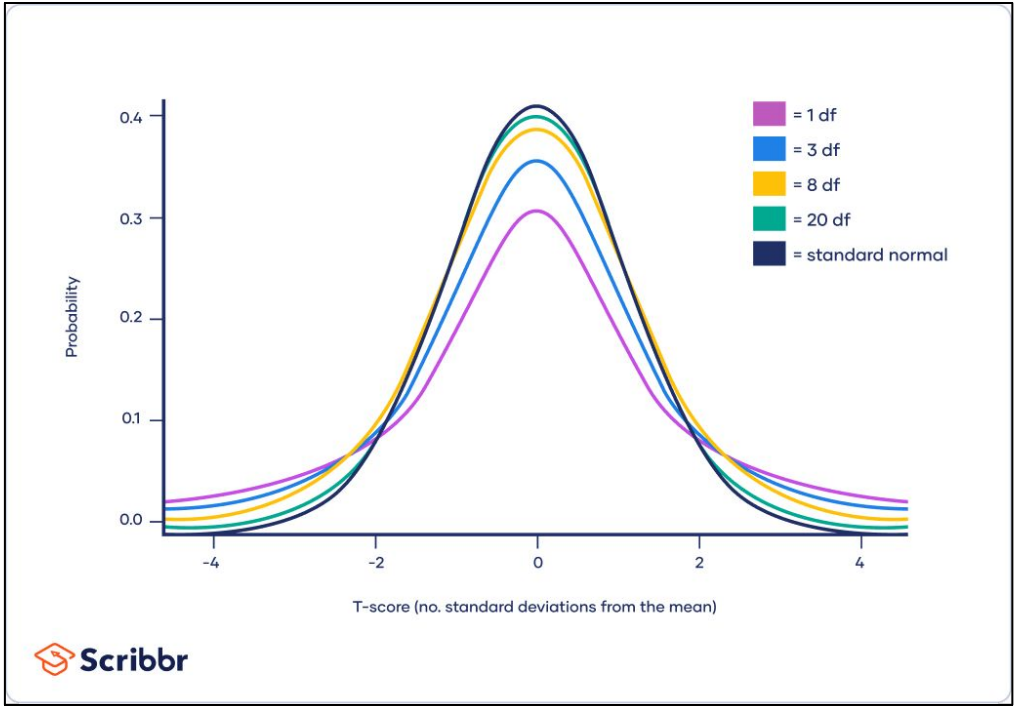
Theory-Based Confidence Intervals (for Proportions)

- A **CI** has the form of point estimate ± (critical value × SE)
  - **Critical value** is based on our desired **confidence level**
- According to **CLT**, $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$
  - **SE** is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Thus, our **CI** (substituting in **p̂** for **p**) is $\hat{p} \pm (z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$
  - **z*** is **critical value** in **norm. dist.**

For Means, We Have a Problem

- By **CLT**, $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, but we don't know **σ** (population SD), so we replace it with **s** (sample SD)
- When we use $\frac{s}{\sqrt{n}}$ as our SD, our **standardized test statistic** will follow a ***t* distribution** with df = n − 1 rather than **N(0, 1)**
  - Using the ***t* distribution** accounts for the extra variability introduced by using **s** as an estimate of **σ**
  - Our CI should be wider because we are now more uncertain

t distribution



For a t distribution, what happens as the degrees of freedoms increase?

Question:

For a t distribution, what happens as the degrees of freedoms increase?

As **degrees of freedom** increase for a ***t*** **distribution**, it looks more like a **normal distribution**.

Intuitively, as **degrees of freedom** increase, there is less uncertainty, so it becomes more appropriate to use **normal distribution**.

What Are Degrees of Freedom?

- **Degrees of freedom**: The number of values in the final calculation of a statistic that are free to vary
- With $n = 3$, if I tell you that $\bar{x} = 10$, $x_1 = 5$, $x_2 = 15$, then what must x_3 be? $x_3 = 10$!
- Thus, the is no variability/independence in that last observation, so degrees of freedom is $n - 1$

Theory-Based Hypothesis Tests (for Means)

- According to **CLT**, under the **H_0**, $\bar{x} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}})$
  - Remember we don't have **σ**, so we replace it with **s**
- Thus, $\bar{x} \sim N(\mu_0, \frac{s}{\sqrt{n}})$
- Now, our ***t*-score (standardized test statistic)** follows a ***t* distribution**
  - $t \sim t(\text{df} = n - 1)$
- $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
  - Remember z-score = $(X - \mu)/\sigma$... This is the ***t* distribution** analogue

Theory-Based Confidence Intervals (for Means)

- A **CI** has the form of point estimate ± (critical value × SE)
  - **Critical value** is based on our desired **confidence level**
- According to **CLT** and substituting in **s** for **σ**, $\bar{x} \sim N(\mu, \frac{s}{\sqrt{n}})$
  - **SE** is $\frac{s}{\sqrt{n}}$
- Thus, our **CI** is $\bar{x} \pm (t^* \times \frac{s}{\sqrt{n}})$
  - **t*** is critical value in ***t* distribution**

The Important Functions for Normal Distribution

- **pnorm()**: Used to calculate **probabilities** on a **normal distribution** (often, for **p-value** during **hypothesis test**)
    - *Ex: What is the* ***probability*** *a student scores an 1800 or less on the SAT if the scores are N(1500, 300)?*
- `pnorm(q = TEST-STAT, mean = MEAN, sd = STAN-DEV)`
    - *`Ex: pnorm(q = 1800, mean = 1500, sd = 300) = 0.8413447`*
- **qnorm()**: Used to calculate **quantiles** on a **normal distribution** (often, for **critical value** during **confidence interval**)
    - *Ex: What score on the SAT would put a student in the 99th* ***quantile*** *(percentile)?*
- `qnorm(p = QUANTILE, mean = MEAN, sd = STAN-DEV)`
    - *`Ex: qnorm(p = 0.99, mean = 1500, sd = 300) = 2197.904`*

The Important Functions for *t* distribution

- **pt()**: Used to calculate **probabilities** on a ***t* distribution** (often, for **p-value** during **hypothesis test**)
  - *Ex: What is the **probability** a student scores a 3 or less on an exam if the scores are ~ t(301 - 1)?*
- `pt(q = TEST-STAT, df = DEGREES-OF-FREEDOM)`
  - *Ex:* `pt(q = 3, df = 301 - 1) = 0.9985369`
- **qt()**: Used to calculate **quantiles** on a ***t* distribution** (often, for **critical value** during **confidence interval**)
  - *Ex: What score would put a student in the 99th **quantile** (percentile)?*
- `qt(p = QUANTILE, df = DEGREES-OF-FREEDOM)`
  - *Ex:* `qt(p = 0.99, df = 301 - 1) = 2.338842`

Let's Recap

- Want **probability**?
  - Use `pnorm()`, `pt()`
  - This is often done for **p-value** in **hypothesis testing**
- Want **quantile** (i.e. percentile)?
  - Use `qnorm()`, `qt()`
  - This is often done to find **z*** or **t*** in **confidence intervals**

Important Code for Theory-Based Inference

https://drive.google.com/file/d/1I2_ySaupN7crU8EwRVY1y_PFsfQP9nem/view?usp=drive_link

“Estimate” vs. “Statistic” in R

- **Estimate** is the **observed sample statistic** (i.e., the numeric quantity calculated with the data set)
  - *Here, the dataset had a sample correlation coefficient of -0.398*
- **Statistic** is the **standardized test statistic** (i.e., z-score or *t*-score)
  - *Here, that sample statistic is 7.07 standard errors below what we'd expect if the null hypothesis were true (i.e., if there is no correlation between age and vitamin D levels)*
  - *Here, the standardized test statistic is a t-score that's distributed t(266)*

```
## # A tibble: 1 x 8
##   estimate statistic  p.value parameter conf.low conf.high method    alternative
##      <dbl>     <dbl>    <dbl>     <int>    <dbl>     <dbl> <chr>     <chr>
## 1   -0.398     -7.07 6.89e-12       266       -1    -0.309 Pearson'~ less
```

In the previous example, what values can “estimate” take on? What values can “statistic” take on?

Question:

In the previous example, what values can “estimate” take on? What values can “statistic” take on?

“Estimate,” as a sample correlation coefficient, can take on values in the interval [-1, 1].

“Statistic,” as a t-score, can take on values in the interval $(-\infty, \infty)$.

Sample Size Calculation

- This is performed **before collecting data** to determine an appropriate **sample size** to gain desired **precision** for a **CI**
  - If my CI for average amount of sleep is between 1 and 23 hours, how helpful is that?
- **CI = point estimate ± (critical value × SE)**, where **margin of error = (critical value × SE)**
  - For proportions, **margin of error = z* ×** $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - For means, **margin of error = t* ×** $\frac{s}{\sqrt{n}}$
- We want our **margin of error** to be no larger than **B**, a bound
  - For proportions, $z* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq B \implies \frac{(z*)^2\hat{p}(1-\hat{p})}{B^2} \leq n$
  - For means, $t* \times \frac{s}{\sqrt{n}} \leq B \implies \frac{(st*)^2}{B^2} \leq n$

Questions?

P-Set 6

Have a great rest of your week!