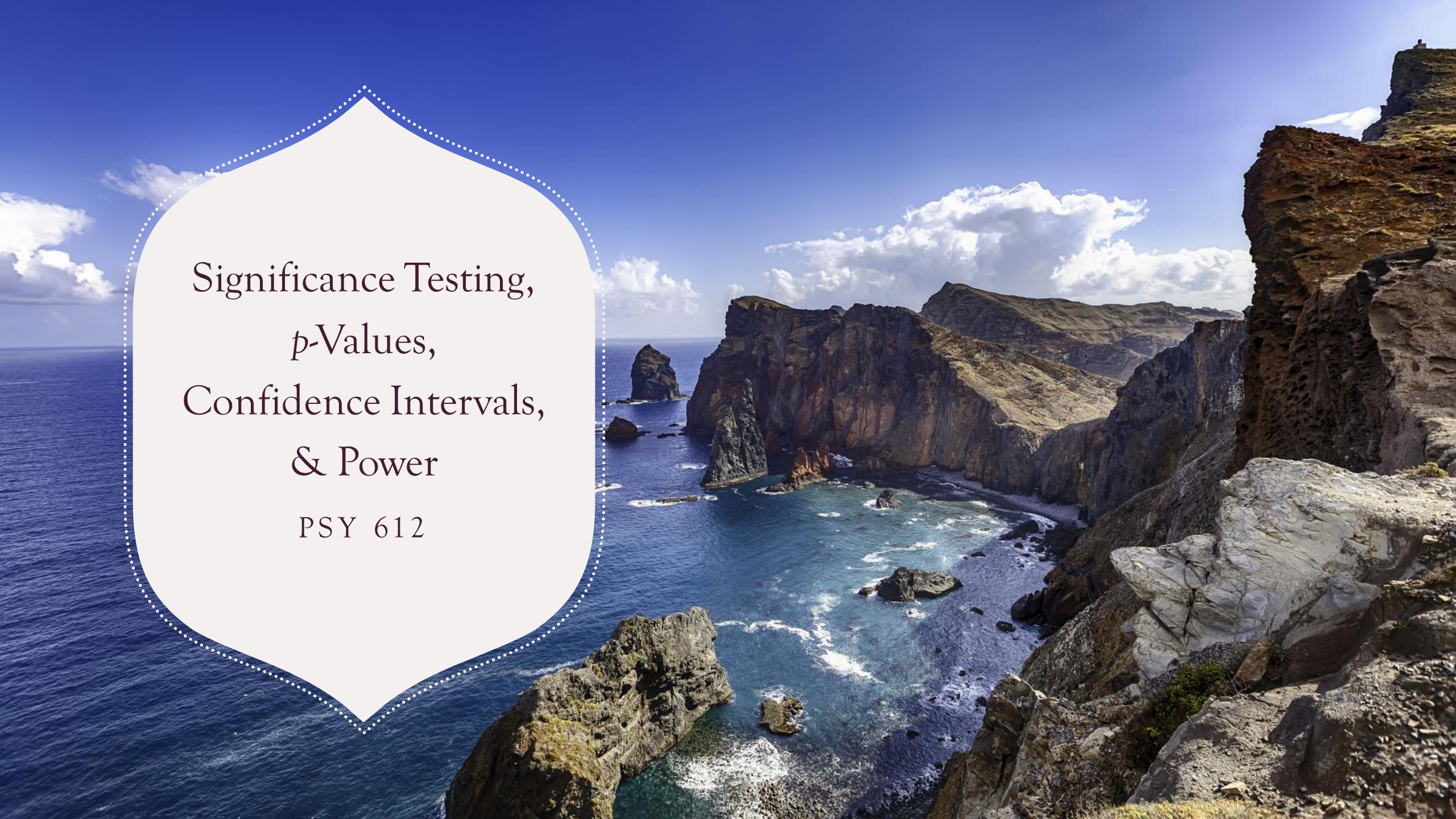


Significance Testing, *p*-Values, Confidence Intervals, & Power

PSY 612

Testing Hypotheses about Populations Using Sample Data

- Psychologists are typically interested in asking research questions about entire populations of people
  - Example: How does age predict neuroplasticity in humans?
  - Example: How does socioeconomic inequality predict voting behaviors among individuals in the US?

- However, we typically collect data from only a single sample.
  - Example: Measuring the relationship between age and neuroplasticity among 300 people (n = 500)

Q: How is it that we test hypotheses about entire populations using only a single sample of collected data?

Significance Testing

- We have focused on interpreting testing the significance of a predictor based on an ***F*-statistic** found in the Anova( ) output

- Now, let's focus on interpreting the significance of a predictor based on the ***t*-statistic** found in the summary( ) output

- Let's start with the concept of **sampling**.

Anova(model, type = 3) output:

```
Anova Table (Type III tests)

Response: final_grade
                          Sum Sq Df   F value    Pr(>F)
(Intercept)                53244  1 41069.859 3.557e-09 ***
intervention                 372  1   287.140 7.111e-05 ***
satisfaction_c                21  1    16.377 0.0155157 *
intervention:satisfaction_c  135  1   104.430 0.0005167 ***
Residuals                      5  4
```

- summary( ) output

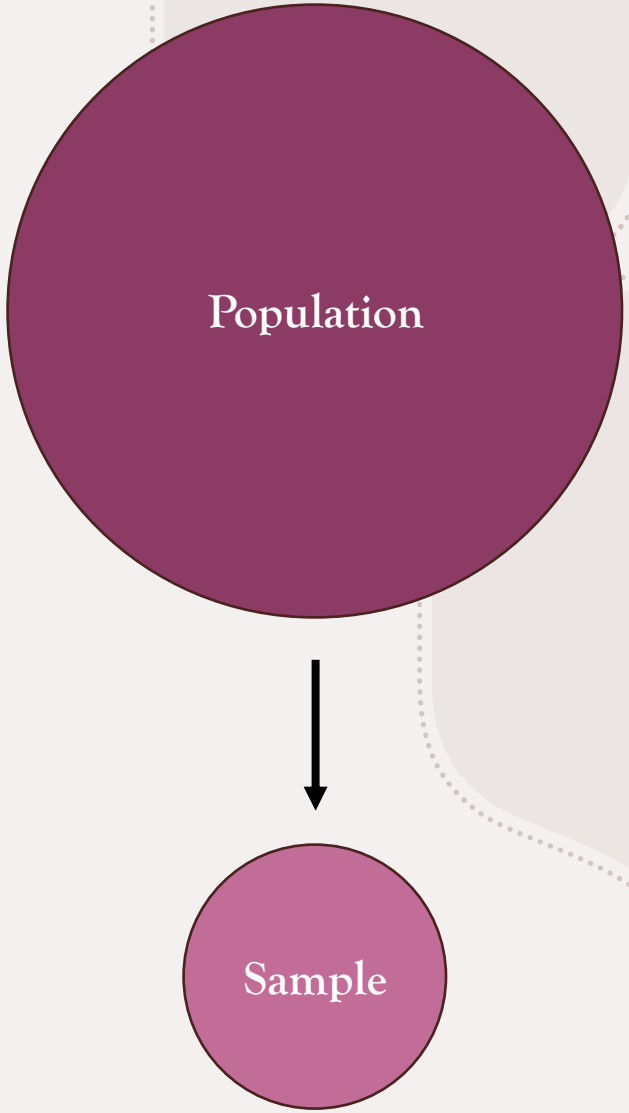
```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 83.6848     0.4129 202.657 3.56e-09 ***
intervention1               13.9946     0.8259  16.945 7.11e-05 ***
satisfaction_c              -0.9929     0.2453  -4.047 0.015516 *
intervention1:satisfaction_c  5.0143     0.4907  10.219 0.000517 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.139 on 4 degrees of freedom
Multiple R-squared:  0.9905,    Adjusted R-squared:  0.9834
F-statistic: 139.5 on 3 and 4 DF,  p-value: 0.0001674
```

Sampling

- Although researchers are interested in studying **populations**, we typically collect data from only a single **sample**.

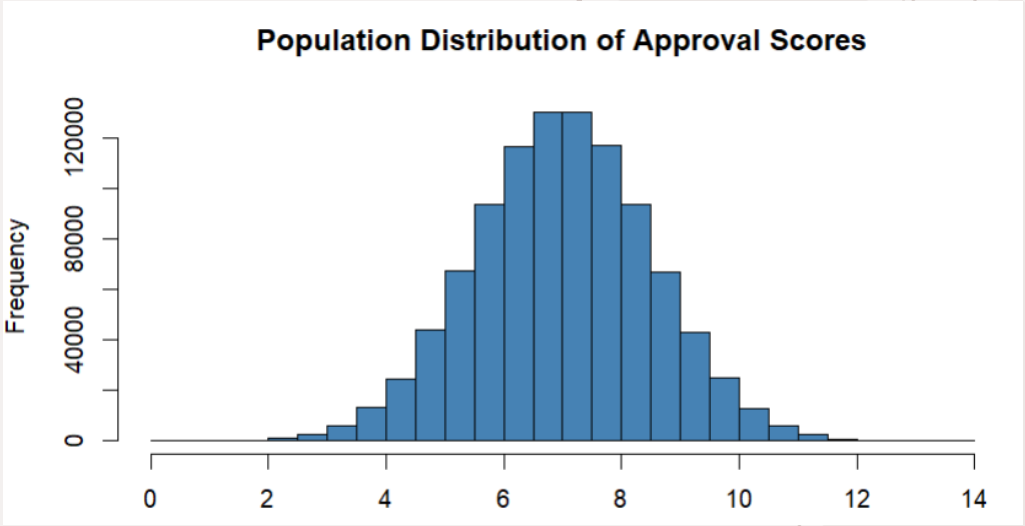
- **Population:** entire group of people we would like to study
- **Sample:** the smaller subset of participants recruited from the population that we collect data from



Population

- Example: Let's say we're interested in how strongly people in the US approve of forgiving student loans (1 = *strongly disapprove* to 13 = *strongly approve*). The distribution on the right represents everyone's approval scores in the population.

- We wouldn't know the mean of the population without collecting population-level data.
  - For this thought exercise, let's say we know that the population mean is 7 ($\mu = 7.00$).



Q: What do you notice about people's scores in the population?

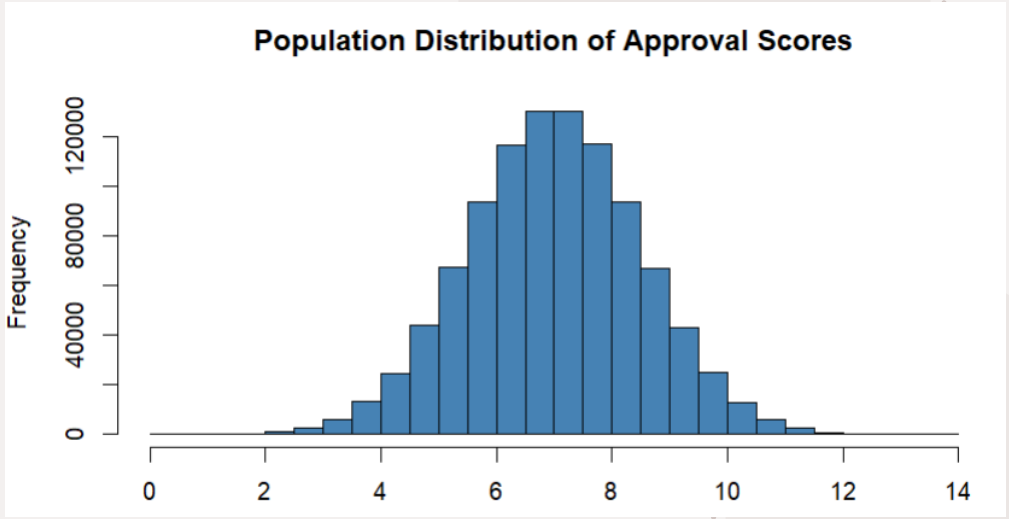
- There's natural variation in how people score in the population.

Sample

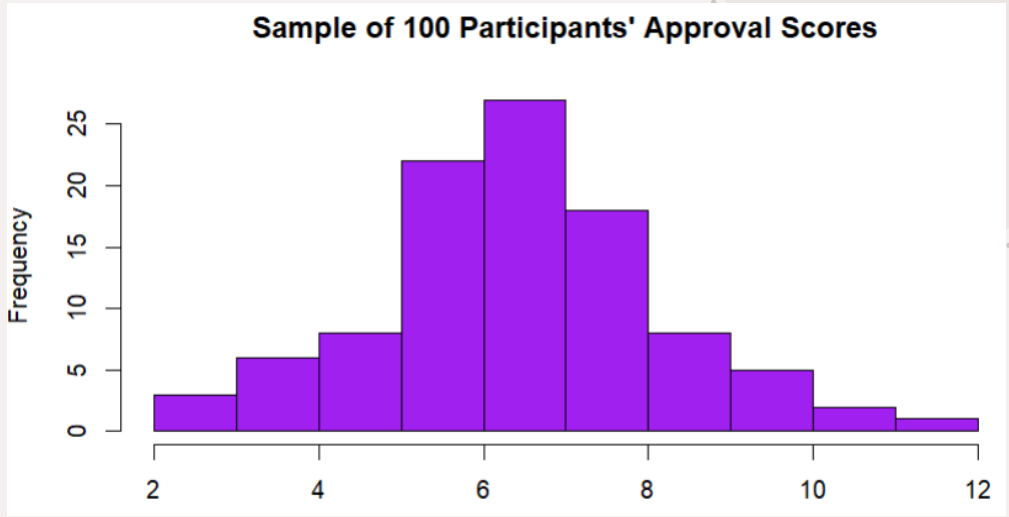
- Example: Let's say a researcher collects a random sample of 100 participants (n = 100) from the US population and measures their approval ratings for forgiving student loans.
  - The mean approval rating of this sample is equal to 6.70 (M = 6.70, SD = 1.28).

Q: Why is the mean of the sample not equal to the mean of the population?

- **Sampling error:** the difference between an estimate of a population parameter calculated using sample data and the true value of the population parameter.



μ = 7.00



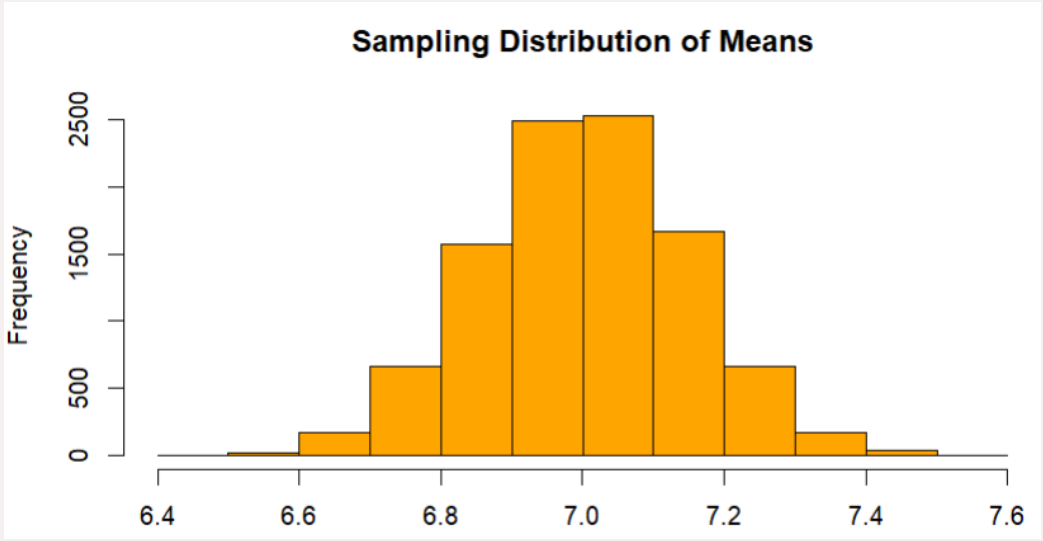
M = 6.70

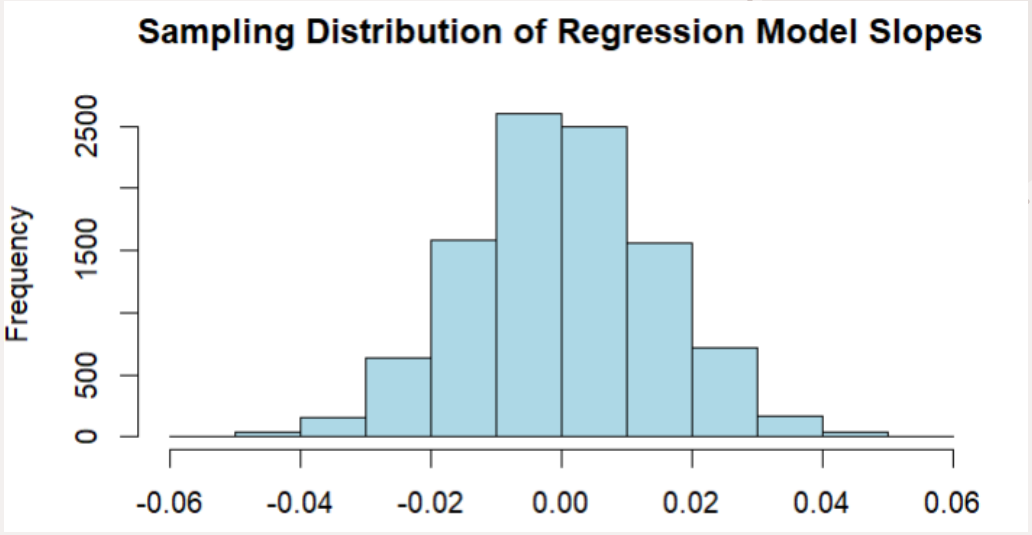
Limitations of Conducting Research with Samples

- Statistics calculated based on a single sample are not perfect estimates for the corresponding population parameters.

Sampling Distributions

- A **sampling distribution** is a distribution of the values of a sample statistic that would occur when repeatedly taking samples of a given size, *n*, from a particular population
  - Can be constructed for any statistic

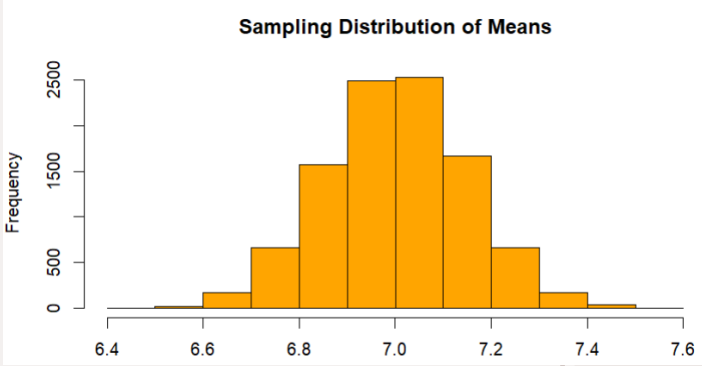


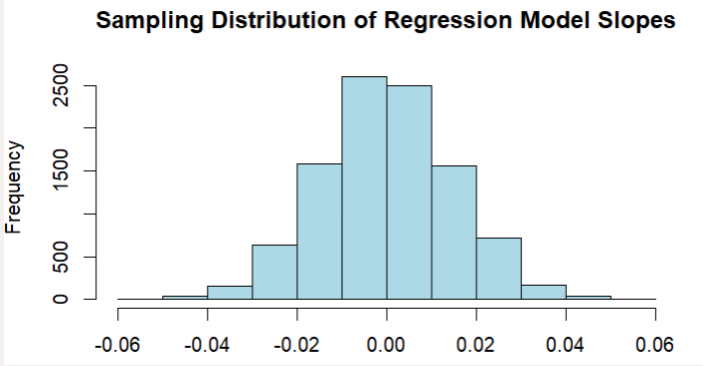


Types of Sampling Distributions

- Sampling Distribution of Means
- Sampling Distribution of Regression Model Slopes

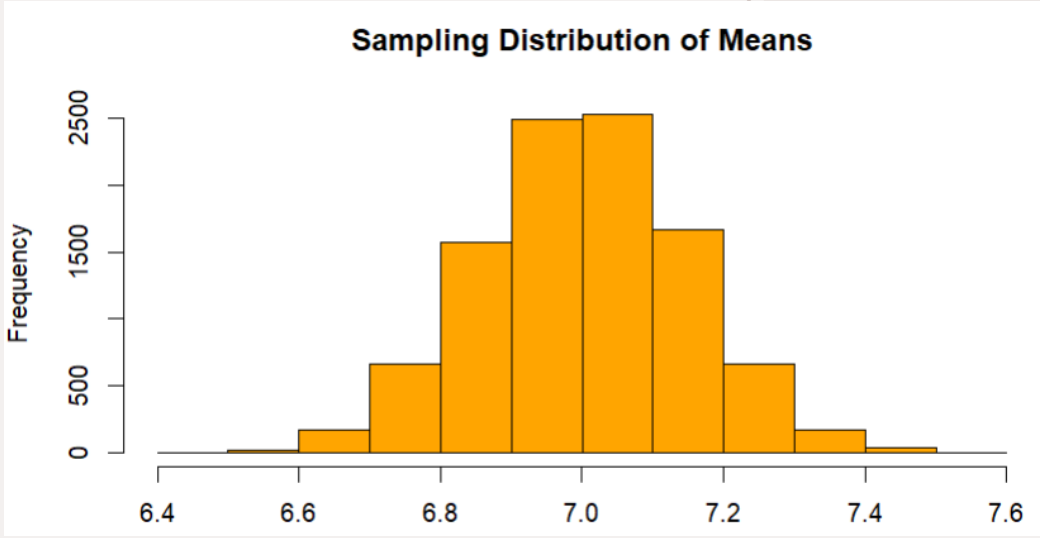
....and more! But these are the ones we will refer to today.





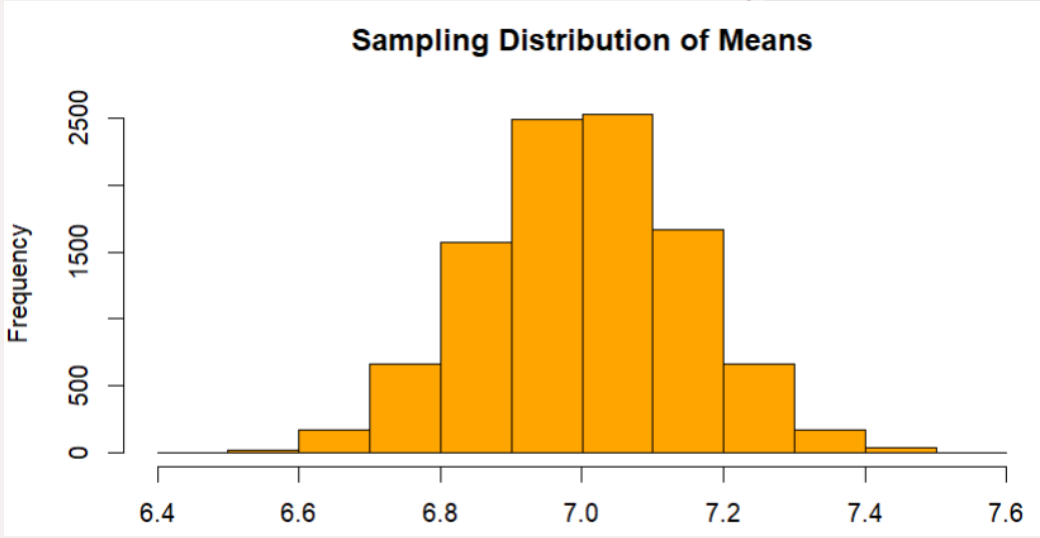
Sampling Distribution of Means

- The simplest sampling distribution is the **Sampling Distribution of Means**
  - For our example, this sampling distribution is constructed from *all* the means that could be calculated from *all possible samples of size n = 100* taken from the population (μ = 7.00)
  - Let's say our particular sample had a mean of 6.70 (M = 6.70) and a standard deviation of 1.28 (*s* = 1.28).



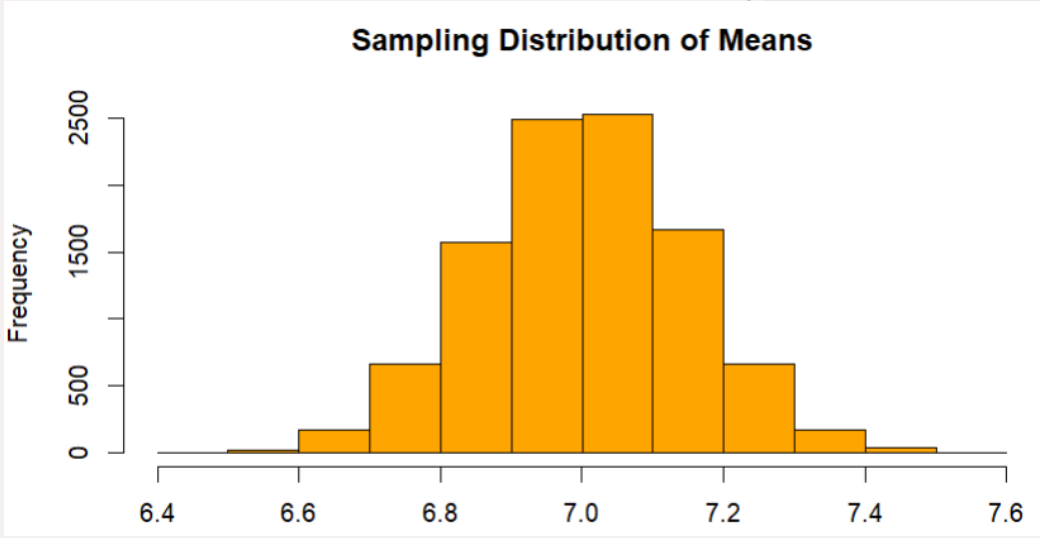
Sampling Distribution of Means

- **Characteristics of Sampling Distribution of Means**
  - Mean = the mean of the population that the samples were obtained from
  - Standard error = $\sqrt{\frac{\sigma^2}{n}}$ or $\sqrt{\frac{s^2}{n}}$
  - Shape = normal distribution if σ is known, *t*-distribution if σ is estimated



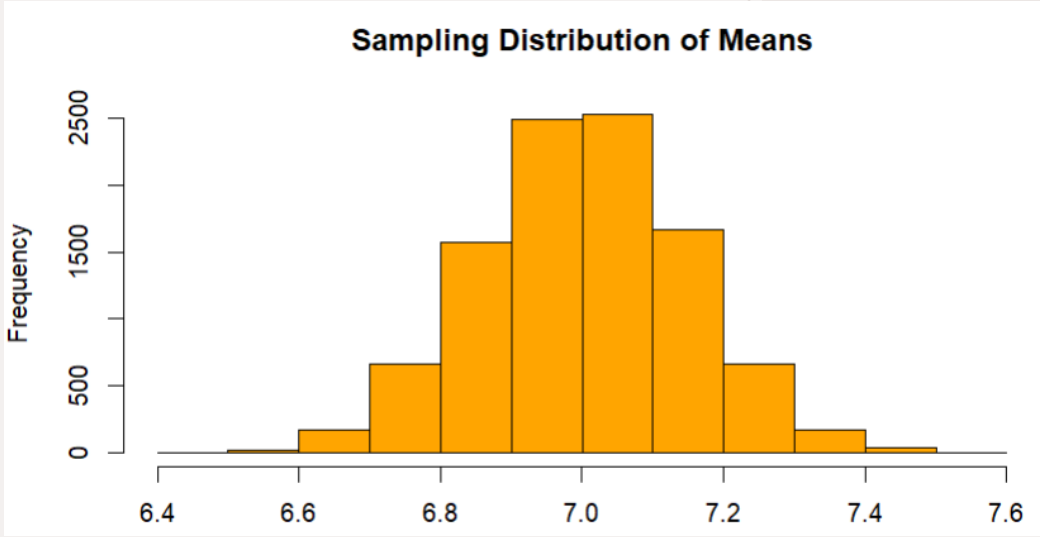
Sampling Distribution of Means

- For our example's **sampling distribution of means:**
  - Mean = μ_M = 7.00
  - Standard error = $\sqrt{\frac{1.28^2}{100}}$ = 0.13
  - Shape = *t*-distribution



Sampling Distribution of Means

- **Standard error** is a measure of the average variation in a statistic across samples.
  - For our example, the average mean across all the samples was equal to 7, and
  - The standard error, or average deviation from this overall mean, is ± 0.13
    - The sample means that fall within ± 1 SE of the overall mean lie in the range from 6.87 to 7.13.
  - This is what **standard error bars** around a sample mean represent.



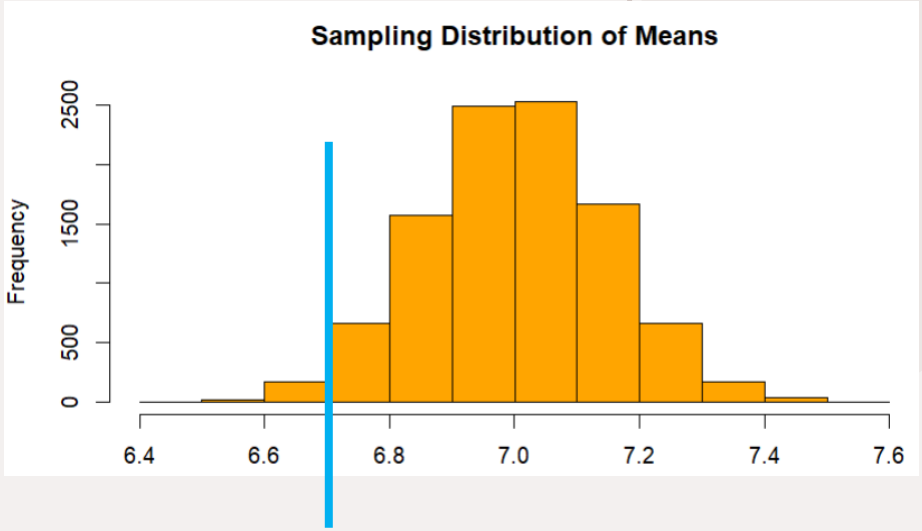
Test Statistic

- A **test statistic** can be calculated to locate a *specific statistic* on a sampling distribution relative to the mean of the distribution in standard error units.

- $t = \frac{\text{parameter estimate} - \text{mean of the sampling distribution}}{\text{standard error}}$

- Where does our sample mean of 6.7 land on the sampling distribution?

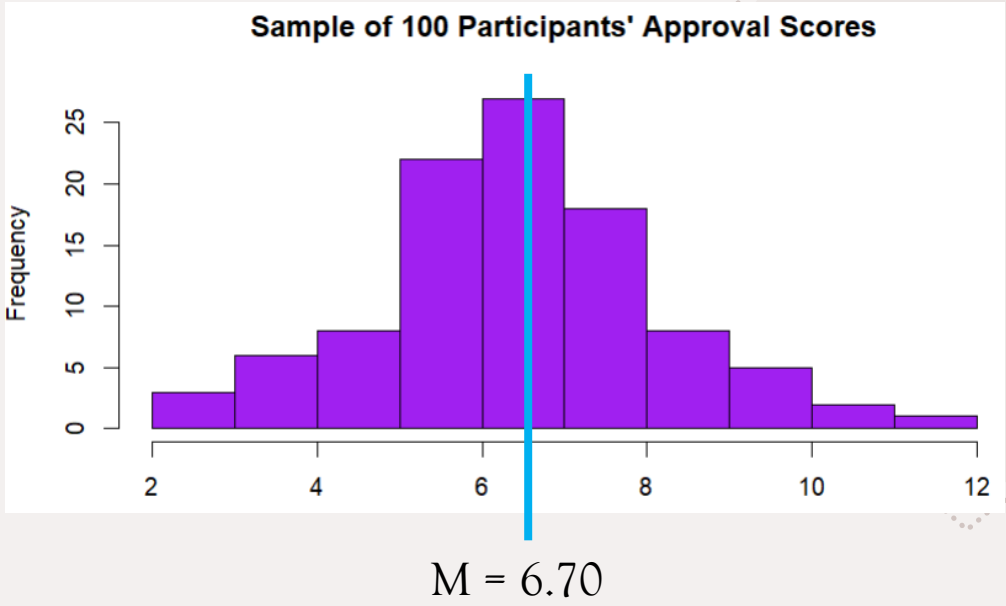
- $t = \frac{6.7 - 7}{0.13} = -2.31$, our sample mean is 2.31 standard errors below the average mean



M = 6.70,
t = -2.31

Confidence Interval

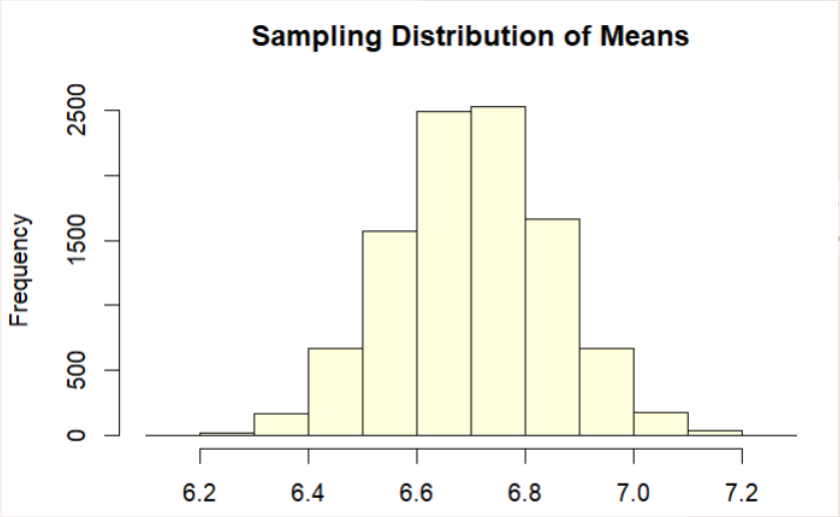
- The sample mean is an example of a **point estimate**, or a single numerical value calculated using sample data to estimate a population parameter
  - For the reasons discussed, point estimates are unlikely to be exactly equal to the true population parameter they're attempting to estimate (ex: M → μ)
- A **confidence interval** is a range of estimates around our particular point estimate that is more likely to contain the true population parameter
  - Specifically, a 95%CI contains 95% of the sample means that occur around our particular sample mean

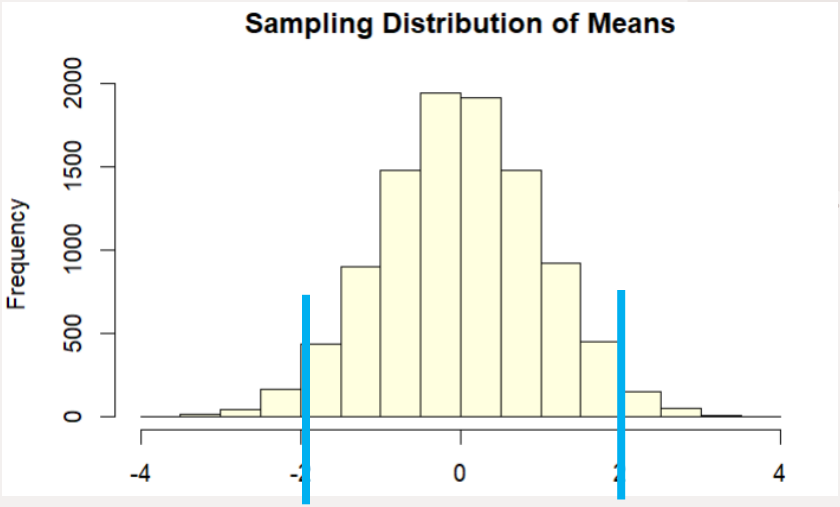


95% Confidence Interval

- A 95%CI is like calculating a sampling distribution with a mean equal to your point estimate and calculating the range for the middle 95% of the sample means

- 95%CI = point estimate ± t_{CV}*SE
  - The *t*-critical values represent how far away the sample means partitioning 5% (when α = .05) of the sample means in the tails of the distribution in standard error units
  - Which means they are also the values that partition 95% of the sample means in the center of the distribution
  - For example (https://www.criticalvaluecalculator.com/):
    - *df* = 99, α = .05
    - t_{CV} = ± 1.9842





t_{CV} = -1.9842 t_{CV} = +1.9842

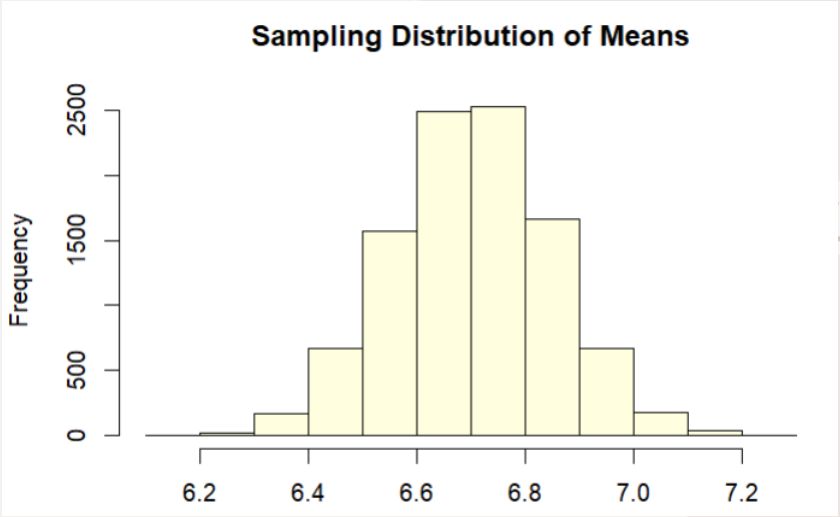
95% Confidence Interval

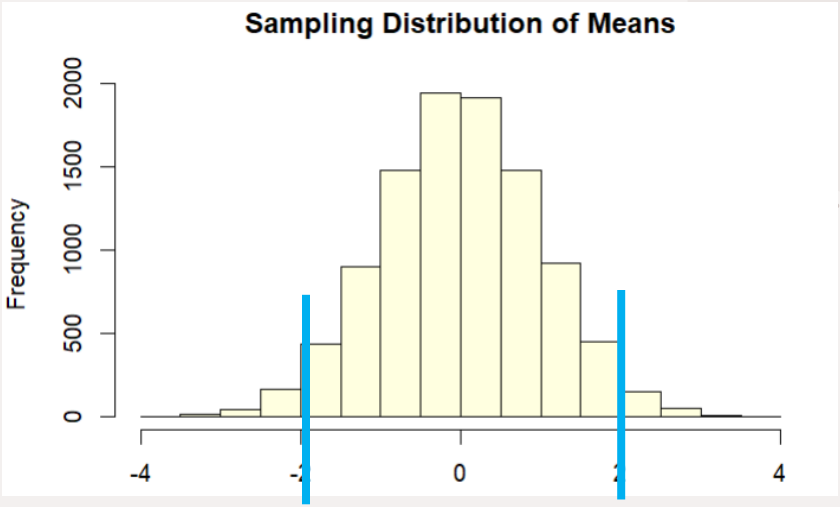
- Our *t*-critical values of ± 1.9842 indicate that the sample means between which 95% of the sample means on our sampling distribution lie are 1.9842 standard errors above and below the mean:

- 95%CI = point estimate ± t_{CV}*SE
  - 95%CI = 6.7 ± (1.9842*0.13)
  - 95%CI = 6.7 ± 0.2579
  - 95%CI = [6.44, 6.96]

95% of the time, confidence intervals calculated around sample means contain the true population mean.

- https://www.statcrunch.com/applets/type3&cimean



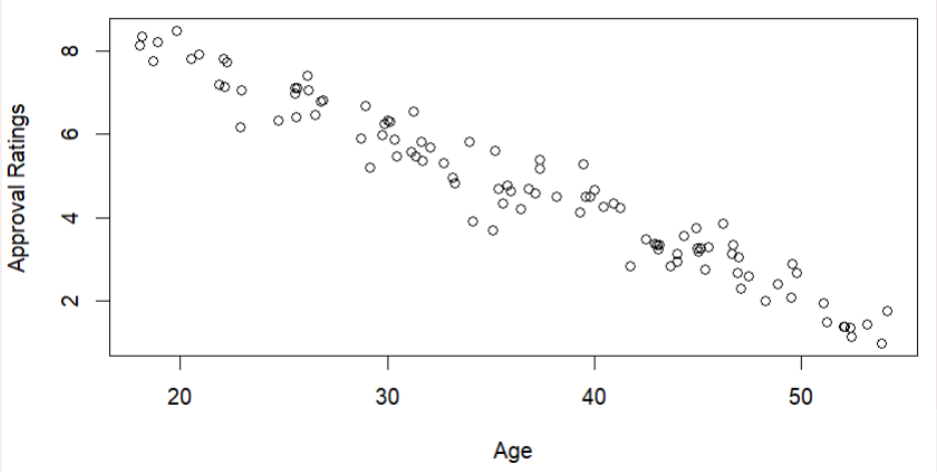


t_{CV} = -1.9842 t_{CV} = +1.9842

Example

- Let's say we're also interested in whether age predicts how strongly people approve of forgiving student loans. This corresponds to the model:
  - $Approval_i = \beta_0 + \beta_1 Age_i + \varepsilon_i$

- When we estimate the model's parameters using our sample of 100 participants, we get:
  - $Approval_i = 11.95 + -0.02*Age + e_i$

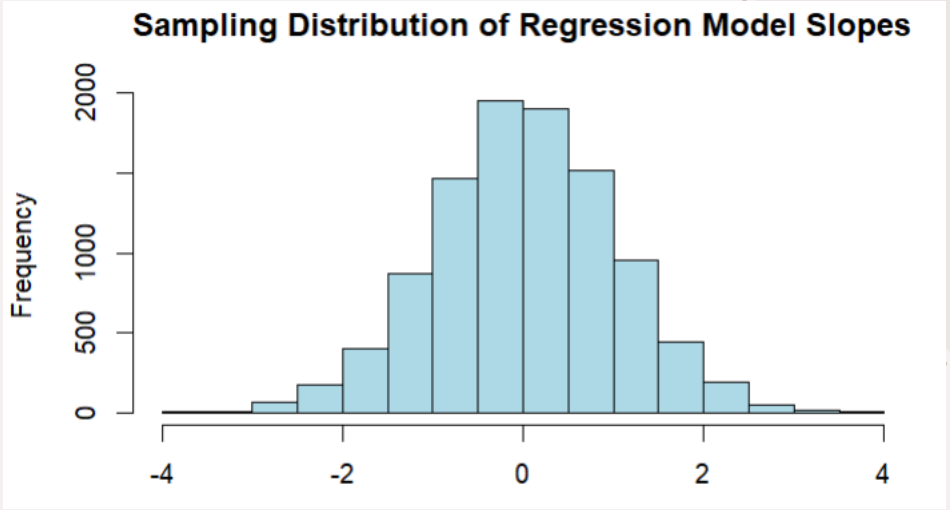


Our results suggest that for every 1-unit increase in age, approval ratings decrease by 0.02 units.

- But, remember that this value of b_1 is just based on a single sample.

Sampling Distribution of Regression Model Slopes

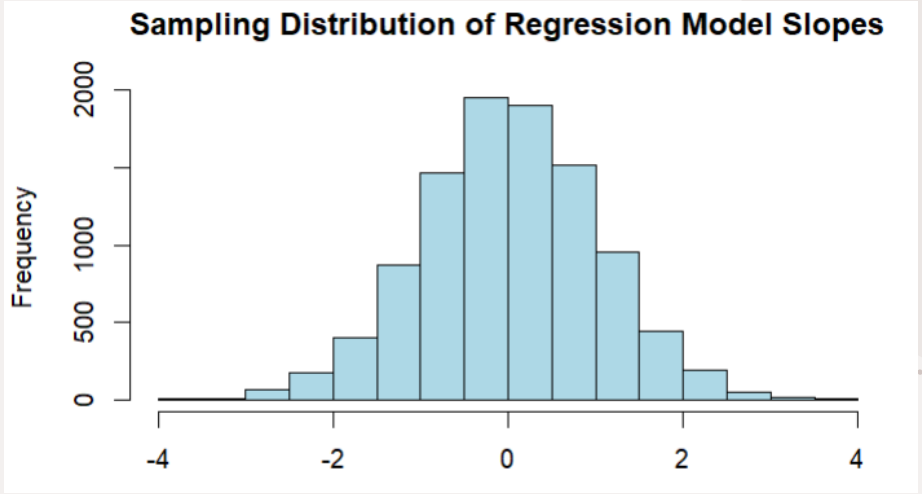
- **Sampling Distribution of Regression Model Slopes**
  - We can compute this sampling distribution using the same method as before. Collect all possible samples of 100 participants, calculate the linear model in every sample, and plot the values of b_1 across all the models in a sampling distribution.



Sampling Distribution of Regression Model Slopes

Sampling Distribution of Regression Model Slopes

- Mean = the population-level slope, β_1
- Standard error = $\sqrt{\frac{MSE}{(SS_X)(1 - {R_p}^2)}}$
  - MSE = variance unexplained by the model
  - SS_X = SS on the predictor variable being tested
  - Tolerance = $(1 - {R_p}^2)$
- Shape = *t*-distribution



Testing Hypotheses about Populations Using Sample Data

- Q: Using the sampling distribution as a tool, how can we test a hypothesis about a whether a relationship exists between age and approval ratings in the population using data from a single sample?

- There are two methods:
  - Null hypothesis significance testing
  - Confidence intervals

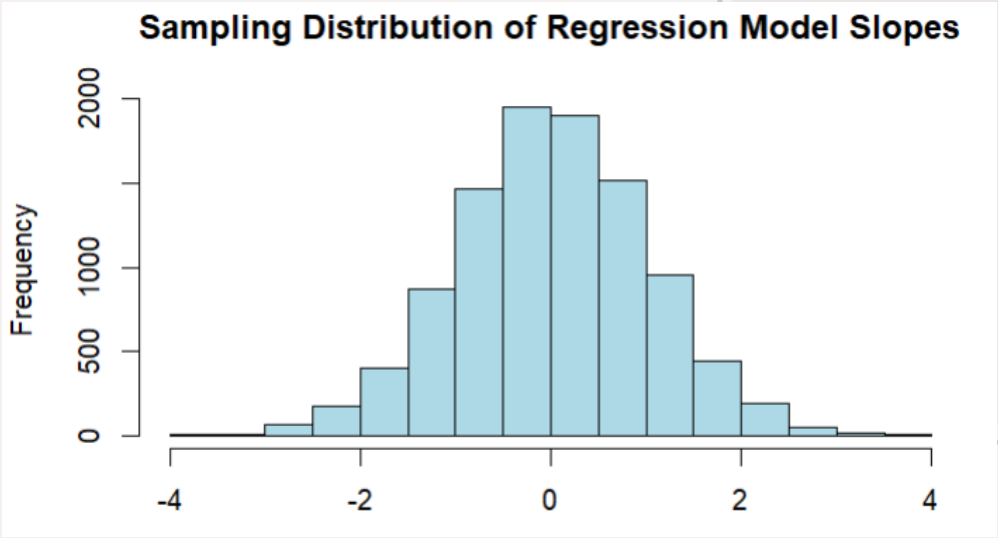
Null Hypothesis Significance Testing

- First, state the **null hypothesis** which makes a specific prediction about the population-level slope of the model
  - H0: $\beta_1 = 0$

- Second, construct a **sampling distribution of regression slopes** representing all the regression slopes that could result when sampling 100 people from the population *described by the null hypothesis*.

- Third, calculate a **test statistic** to examine where your estimate of the regression slope lands on the sampling distribution.

- Calculate the **probability** (aka, *p*-value) of getting a test statistic as or more extreme than your test statistic on the sampling distribution <u>if the null hypothesis is true</u>.



Analysis in R

Let's perform the analysis in R and see how the results map onto this process.

Fit the model

```
model <- lm(approval ~ age, data = sample)
```

summary( ) output

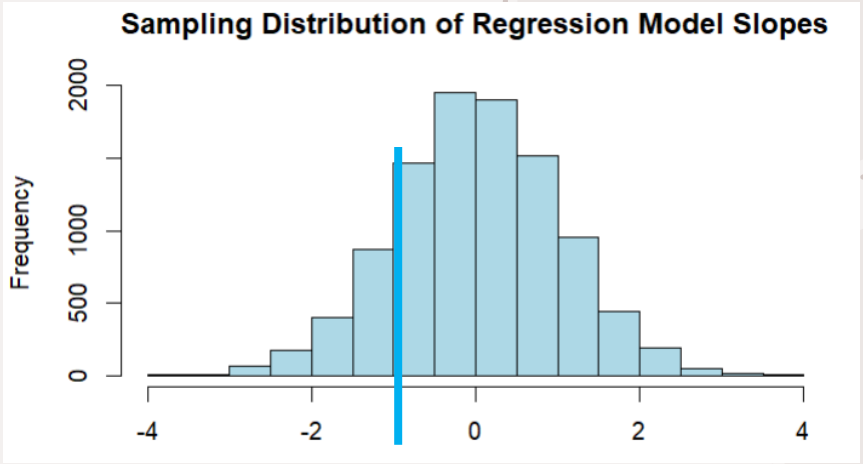
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.76157    0.62341  18.866   <2e-16 ***
age         -0.01638    0.01357  -1.207     0.23
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.898 on 98 degrees of freedom
Multiple R-squared:  0.01465,   Adjusted R-squared:  0.004591
F-statistic: 1.457 on 1 and 98 DF,  p-value: 0.2304
```

p-Value

```
> pt(-1.207, df = 98, lower.tail = TRUE)
[1] 0.1151687
```

```
> 2*pt(-1.207, df = 98, lower.tail = TRUE)
[1] 0.2303375
```



$t = -1.207$

$$t = \frac{\text{parameter estimate} - \text{mean of the sampling distribution}}{\text{standard error}}$$

$$t = \frac{-0.01638 - 0}{0.01357}$$

$$t = -1.207$$

What is a *p*-Value?

- Let's say we concluded that age was a significant predictor of approval ratings because $p < .05$. Which of the following statements are true?

1. The probability that the null hypothesis is true is .05.
2. The probability of mistakenly rejecting the null hypothesis is less than .05.
3. The difference between the two groups is large.
4. The probability of obtaining a difference as extreme or more extreme than ours is .05.
5. The probability of rejecting the null hypothesis is .95.
6. The probability of obtaining a difference as extreme or more extreme than ours, if the null hypothesis is true, is less than .05.
7. The probability that the alternative hypothesis is true is .95.

Confidence Intervals

- We can calculate a confidence interval around our point estimate of the model's slope using the same general formula as before:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.76157    0.62341  18.866   <2e-16 ***
age         -0.01638    0.01357  -1.207     0.23
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.898 on 98 degrees of freedom
Multiple R-squared:  0.01465,   Adjusted R-squared:  0.004591
F-statistic: 1.457 on 1 and 98 DF,  p-value: 0.2304
```

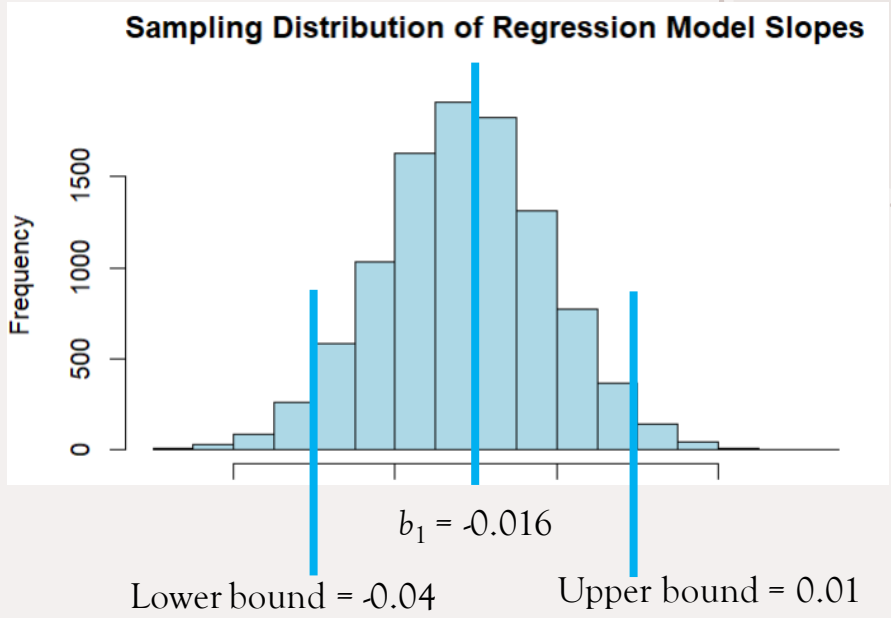
```
> confint(model)
                 2.5 %      97.5 %
(Intercept) 10.52443551 12.99871199
age         -0.04331134  0.01055267
```

- 95%CI = point estimate ± t_{CV}*SE

  The *t*-critical value for *df* = 98 and α = .05 are

  - t_{CV} = ±1.9845

- 95%CI = -0.01638 ± (1.9845*0.01357)
- 95%CI = -0.01638 ± (0.0269)
- 95%CI = [-0.04, 0.01]



p-Values vs Confidence Intervals

- *p*-Values and confidence intervals can both be used to judge the significance of a predictor

- Decision criteria using *p*-Values:
  - If $p < .05$ → Significant
  - If $p > .05$ → Non-significant

- Decision criteria using confidence intervals:
  - If CI does not contain zero → Significant
  - If CI contains zero → Non-significant

The Argument for Confidence Intervals

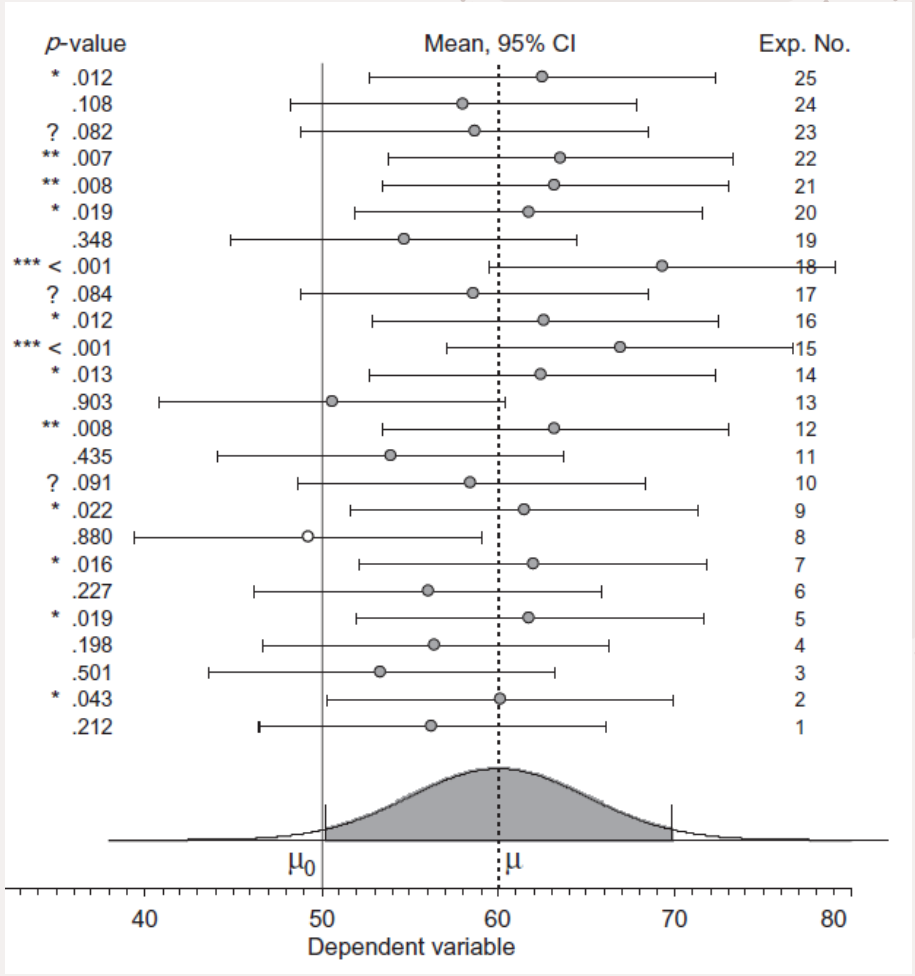
- *p*-Values have been critiqued for promoting dichotomous decision making
  - Results are either "significant" (important and worth publishing)
  - Or "not significant" (not important and not worth publishing)

- Confidence intervals may encourage more nuanced assessments of one's results
  - The range of values within which the true population parameter likely lies within
  - Have the added benefit of also communicating degree of precision in estimating a population parameter
    - Narrower confidence intervals estimate population parameter with greater precision
    - Wider confidence intervals estimate population parameter with worse precision

The Argument for Confidence Intervals

> "The most striking feature of the column of *p* values in Figure 1 is their astonishing variability: from greater than .50 to less than .001. It seems that practically any *p* value might be given by a replication of our experiment. Repeat your experiment and you are likely to get an entirely different *p*. How, then, can the *p* value from a single experiment be a reliable basis for interpretation? That is an extremely good question!"

- Confidence intervals may also be better at communicating what to expect on replication attempts
  - Cumming (2012) found in a simulation study that, across 25 samples, the *p*-values corresponding to the means of each sample varied widely
    - Across replication attempts, the *p*-value ranged from < .001 to .903
  - Whereas the 95%CI contained the true population mean 83% of the time

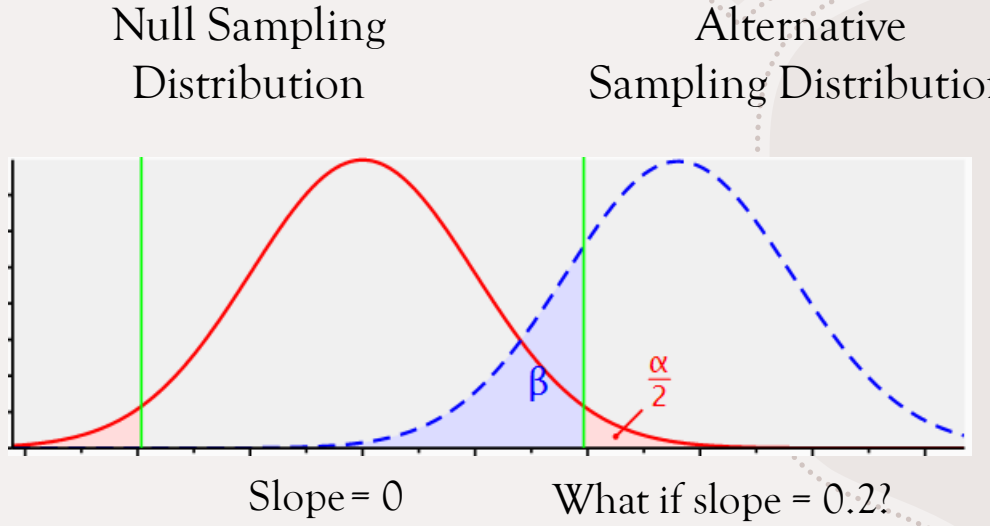


Power Analysis

- Sampling distributions also play an important role in conducting power analyses

- Compare the sampling distribution that occurs if the null hypothesis is true to a sampling distribution that would occur a *specific* alternative hypothesis is true (aka, if there *is* an effect or relationship).

  - α = .05, chances of making a Type I error
  - β = chances of making a Type II error
  - 1 – β = power



- We decide whether our results are significant based on where our test statistic lands **on the null sampling distribution**

Factors Affecting Power

- Power refers to the ability to find a significant effect (or relationship) if there is one
  - Results are more likely to be significant when:

The test statistic is large:

Test Statistic: $t = \frac{\text{parameter estimate}}{\text{standard error}}$

The confidence interval is narrow:

95%CI = parameter estimate ± t_{CV}*SE

Factors Affecting Power to Detect a Significant Regression Slope

Power increases when:

- The parameter estimate is larger (i.e., when the slope of the model is larger)
- The variance left unaccounted for by the model decreases ($MSE = \frac{SSE(A)}{n - PA}$)
  - Improve model accuracy by adding good predictors
  - Increase sample size (n)
- The multicollinearity ($R_p{}^2$) between a model's predictors decreases

The test statistic is large:

Test Statistic: $t = \frac{\text{parameter estimate}}{\text{standard error}}$

SE for a regression slope: $\text{SE} = \sqrt{\frac{MSE}{(SS_X)(1 - R_p{}^2)}}$

The confidence interval is narrow:

95%CI = parameter estimate ± t_{CV}*SE