DATA ANALYSIS

Week 7: Sampling and hypothesis testing

logistics: midterm 1

- exam grades will be made available latest by Thursday morning
- review my comments in rubric
- please come see me if you have questions!

mid-semester check-in

- everyone schedules a 15-minute meeting post spring break [calendly link]
- fill out **anonymous** mid semester survey [opens on Friday]

7	T: March 4, 2025	W7: Sampling and Hypothesis Testing
7	Th: March 6, 2025	W7 continued…
7	F: March 7, 2025	**PS3 revision due**
7	F: March 7, 2025	**Week 7 Quiz due**
8	T: March 11, 2025	**Spring Break!**
8	Th: March 13, 2025	**Spring Break!**
9	T: March 18, 2025	**Spring Break!**
9	Th: March 20, 2025	**Spring Break!**
10	T: March 25, 2025	W10: Modeling Relationships
10	Th: March 27, 2025	W10 continued…
10	Su: March 30, 2025	**Week 10 Quiz due**
11	M: March 31, 2025	**PS4 due / Opt-out Deadline 2**
11	T: April 1, 2025	W11: Special Cases
11	Th: April 3, 2025	W11 continued…
12	M: April 7, 2025	**PS5 + PS4 revision due**
12	T: April 8, 2025	W12: Loose Ends / Exam 2 review
12	Th: April 10, 2025	**Exam (Midterm) 2**

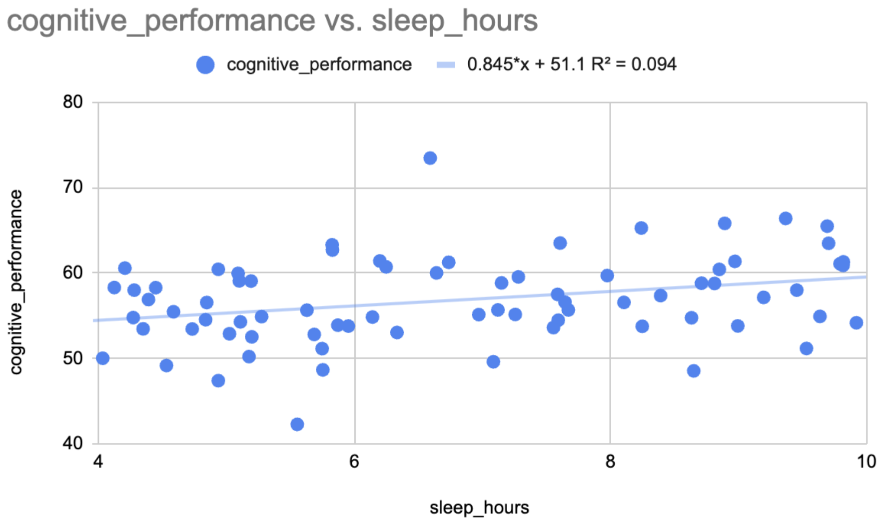
today's agenda

- sampling
- hypothesis testing

an experiment

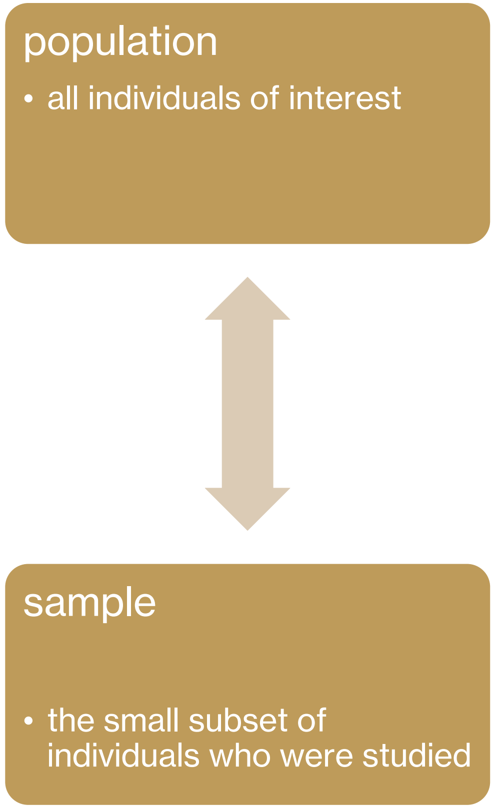
- hypothesis: sleep predicts cognitive performance
- recorded number of hours of sleep via sleep tracker
- cognitive performance via phone game
- model formulation
  - cognitive performance ~ sleep + error
- sample correlation, r = .31

sleep_hours	cognitive_performance
6.24724071	60.7189024
9.70428584	63.4792748
8.39196365	57.3561342
7.59195091	54.4745639
4.93611184	47.3953454
4.93596712	60.434202
4.34850167	53.4598411
9.19705687	57.1451403
7.60669007	63.5052718
8.24843547	53.7573958
4.12350697	58.2974841



from samples to populations

- we have a sample statistic (known: r)
- our population parameter (unknown ρ): it could be close or very far from r
- we can simulate what the population parameter would look like using our sample statistic
- basic idea:
  - samples are small subsets of the population
  - we mimic collecting MANY such samples of the same size and look at the distribution of <u>all possible sample statistics</u> we could obtain

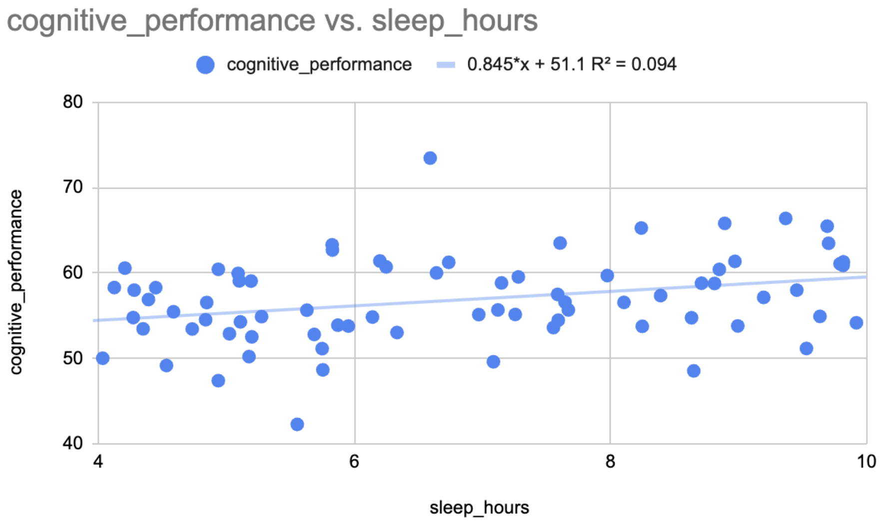


W7 Activity 1

- Jupyter notebook with data
- confirm correlation, slope, and intercept

	sleep_hours	cognitive_performance
0	6.247241	60.718902
1	9.704286	63.479275
2	8.391964	57.356134
3	7.591951	54.474564
4	4.936112	47.395345

```
Actual sample slope: 0.84
Actual sample intercept: 51.08
Correlation between sleep_hours and cognitive_performance: 0.31
```



W7 Activity 1

- what if there was no true relationship between sleep and cognitive performance in the population?
- $\rho = 0$
- how can we mimic this "no relationship" using the sample we have?
- we could keep the sleep_hours column the same but shuffle the cognitive_performance column

original

	sleep_hours	cognitive_performance
0	6.247241	60.718902
1	9.704286	63.479275
2	8.391964	57.356134
3	7.591951	54.474564
4	4.936112	47.395345

shuffled

	sleep_hours	cognitive_performance
0	6.247241	47.395345
1	9.704286	54.845377
2	8.391964	58.297484
3	7.591951	60.718902
4	4.936112	60.422949

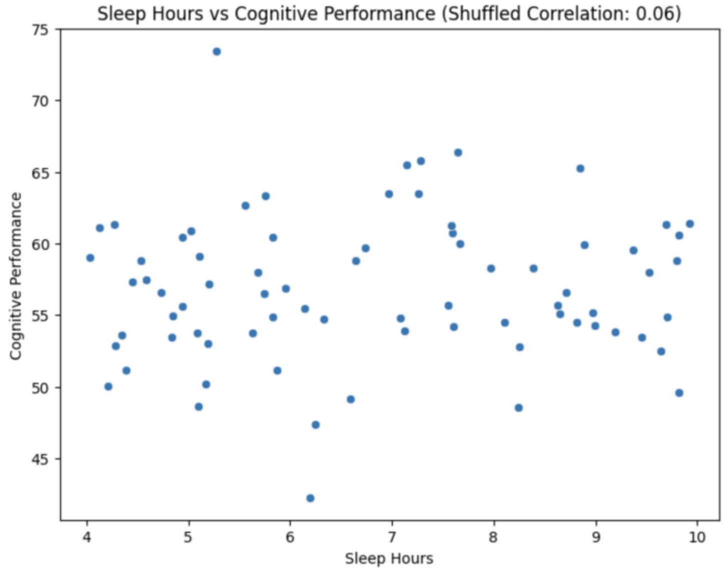
W7 Activity 1

- what if there was no true relationship between sleep and cognitive performance in the population?
- $\rho = 0$
- how can we mimic this "no relationship" using the sample we have?
- we could keep the sleep_hours column the same but shuffle the cognitive_performance column

original

	sleep_hours	cognitive_performance
0	6.247241	60.718902
1	9.704286	63.479275
2	8.391964	57.356134
3	7.591951	54.474564
4	4.936112	47.395345

shuffled



W7 Activity 1

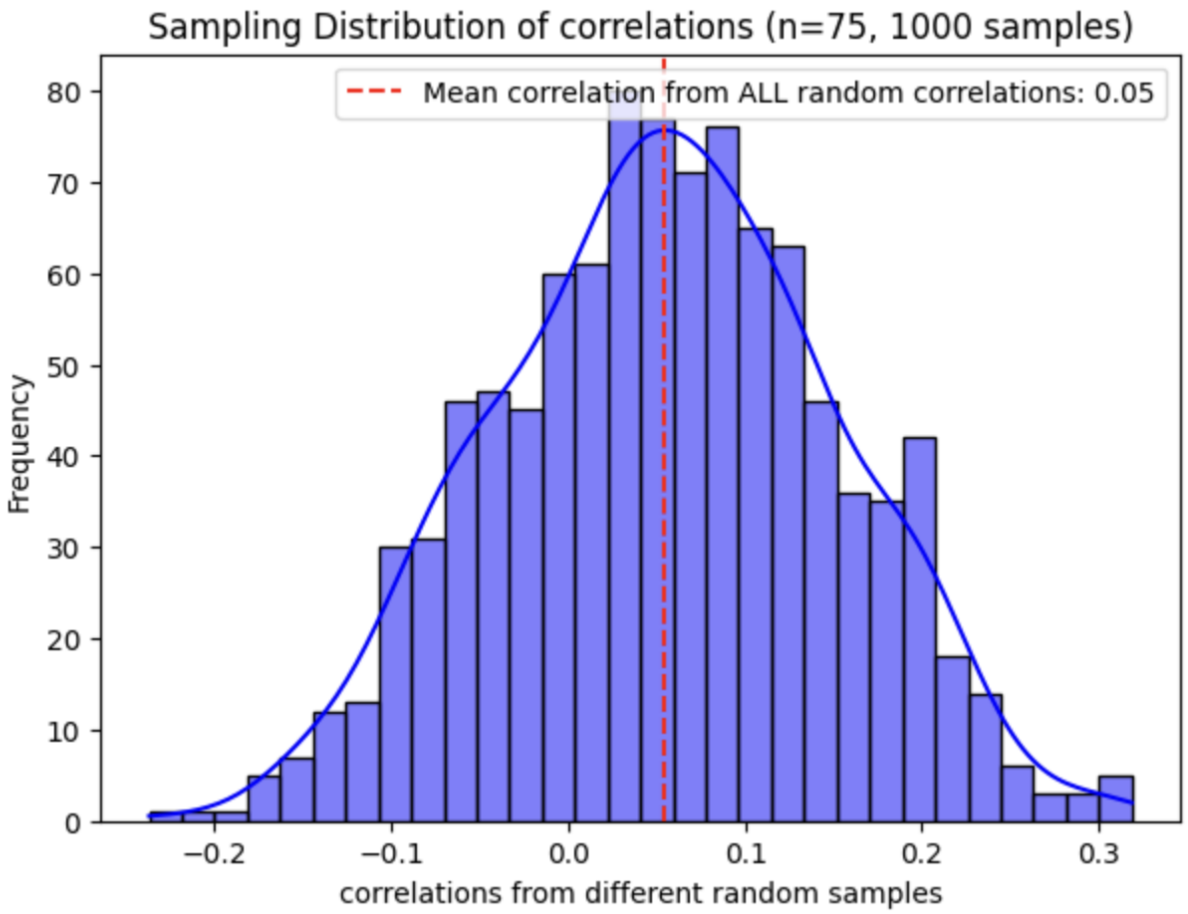
- in this shuffled dataset, what does a random sample look like?
- we repeat this 1000 times, i.e., we take 1000 random samples with replacement
    - random sample: each outcome has an equal chance of being selected
    - sampling with replacement = putting back each observation so that the probability of being selected remains constant on the second draw

```
random sample slope: -0.05
random sample intercept: 56.53
random sample correlation: -0.02

random sample slope: 0.06
random sample intercept: 55.57
random sample correlation: 0.03
```

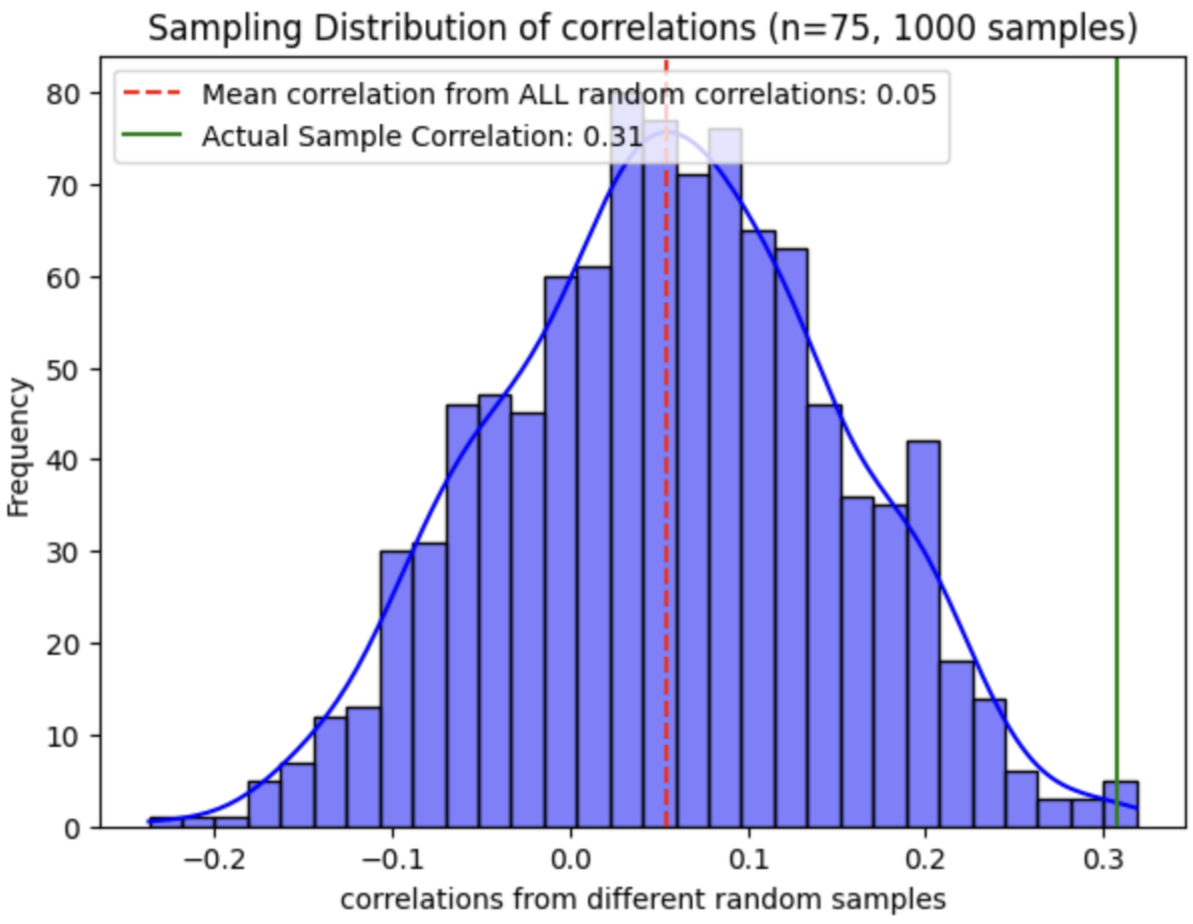
W7 Activity 1

- what does the distribution of correlations look like for MANY such random samples?
- sampling distribution: distribution of all possible values of the **sample statistic** obtained from multiple samples of a given size



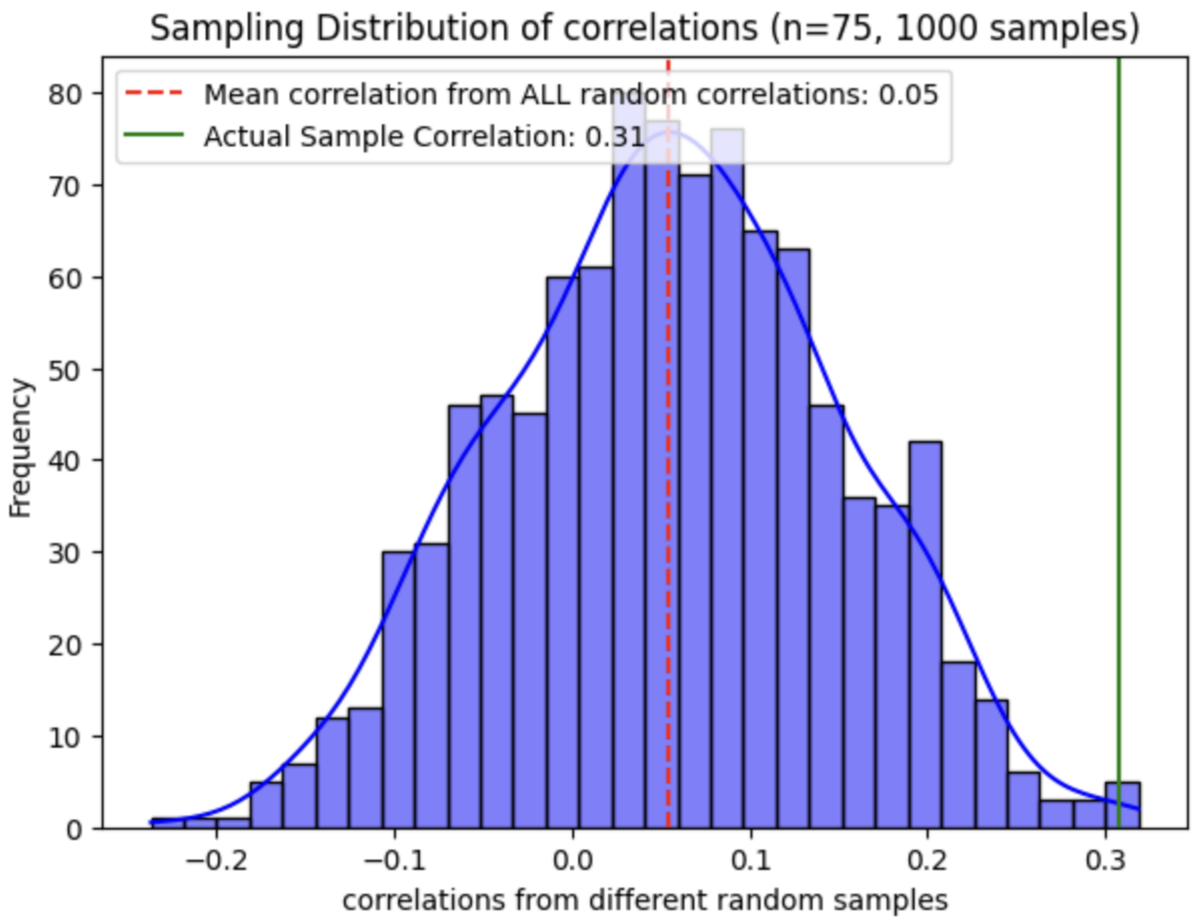
W7 Activity 1

- what does the distribution of correlations look like for MANY such random samples?
- sampling distribution: distribution of all possible values of the ***sample statistic*** obtained from multiple samples of a given size
- now compare this sampling distribution of random slopes from the shuffled dataset to the correlation in the actual sample



W7 Activity 1

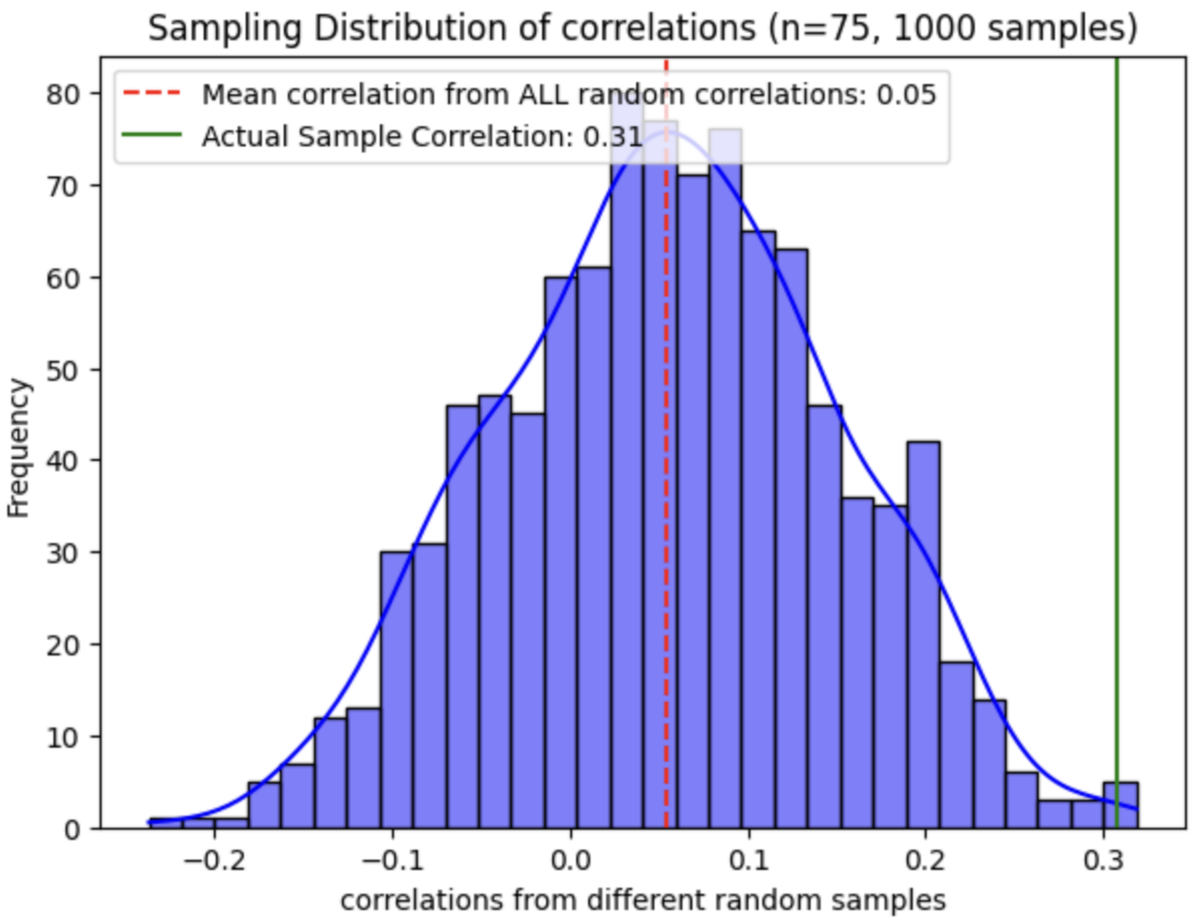
```
Mean of ALL random correlations: 0.05
Standard deviation of ALL random correlations: 0.10
```



- **<u>if there was no relationship between sleep and cognitive performance</u>**:
    - there is an expected distribution of correlations
    - what would be the average distance of a random correlation from the mean correlation?
- standard deviation of sampling distribution of correlations ≅ standard error of correlation
- $SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}}$

W7 Activity 1

```
Mean of ALL random correlations: 0.05
Standard deviation of ALL random correlations: 0.10
```

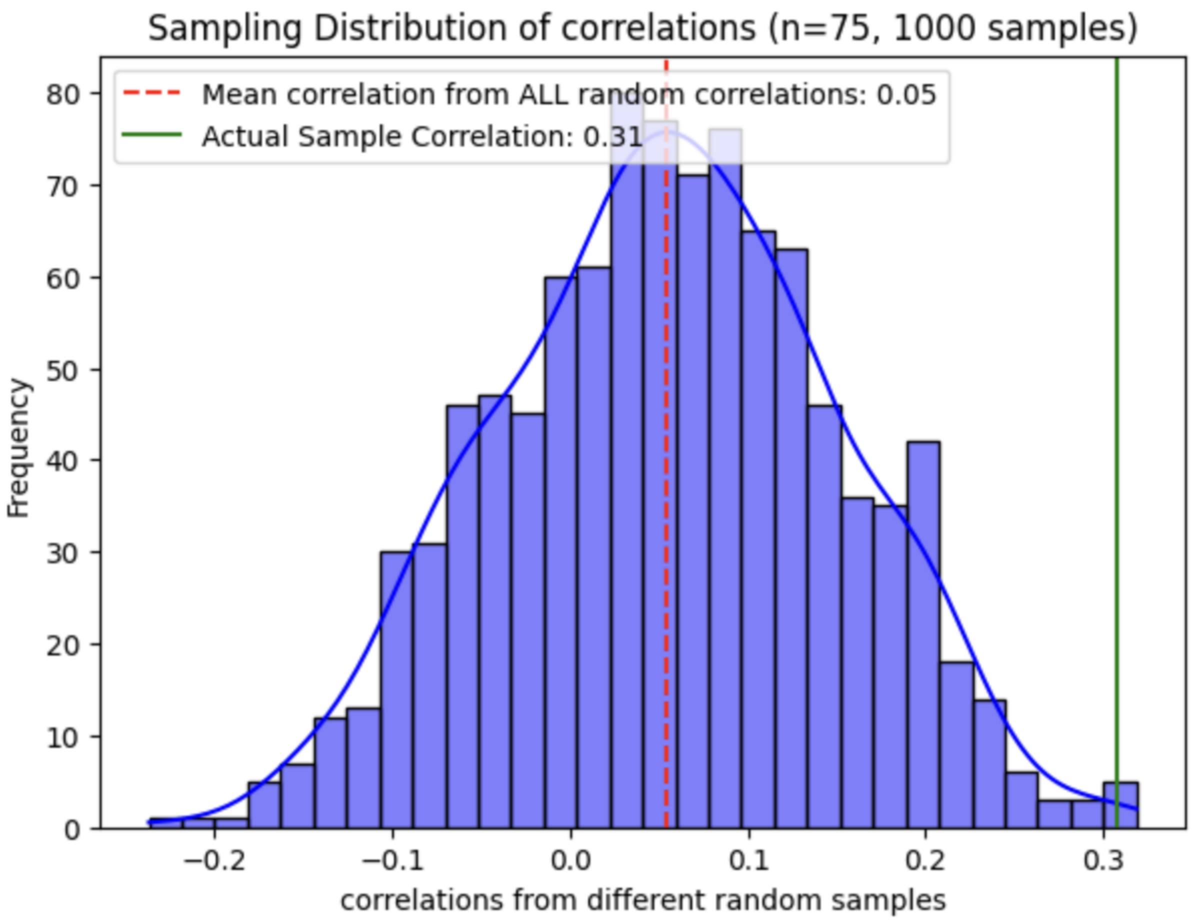


- **<u>if there was no relationship between sleep and cognitive performance</u>**: is obtaining a sample correlation of 0.31 typical?
- ratio of observed difference vs. expected difference
- $\frac{observed\ difference}{expected\ difference} = \frac{r - \rho}{SE_r}$
- like a z-score!

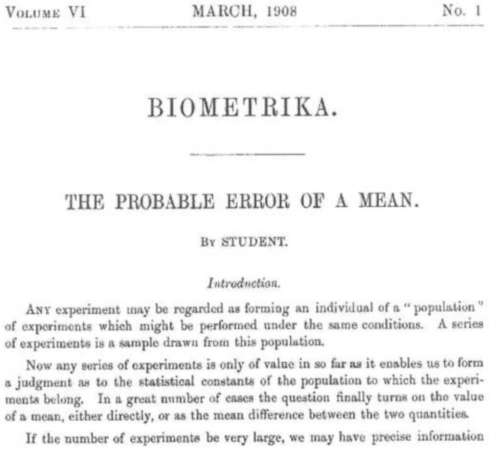
W7 Activity 1

```
Mean of ALL random correlations: 0.05
Standard deviation of ALL random correlations: 0.10
```

- $\frac{observed\ difference}{expected\ difference} = \frac{r - \rho}{SE_r}$
- what if we wanted an EXACT probability for our observed correlation?
- we would need to know the <u>exact form of the sampling distribution</u> of the sample statistic we are calculating



Student's t distribution



Volume VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

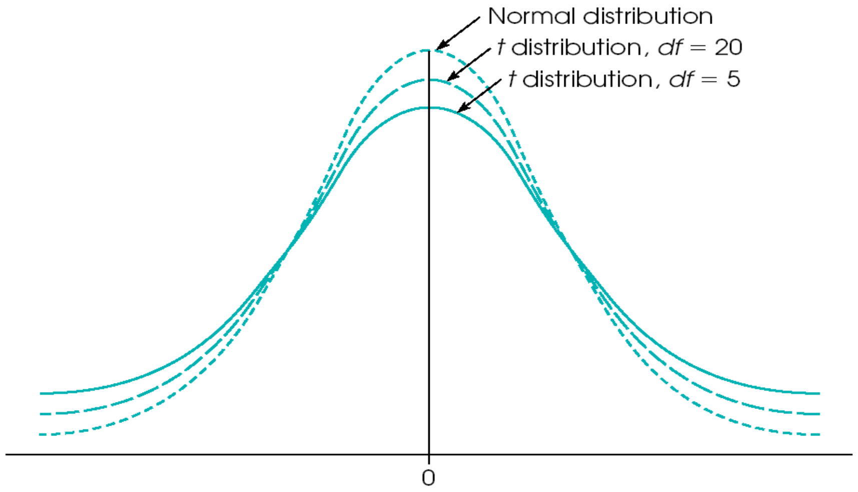
Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

$$t = \frac{sample\ statistic\ - population\ parameter}{standard\ error} = \frac{observed}{expected}$$

- *t* distribution approximates the normal distribution
- how good is this approximation?
  - depends on the sample size (n)
  - each t-curve is defined by *degrees of freedom (df)* which depend on the sample size and statistic
  - for large dfs, the t distribution approximates the normal distribution

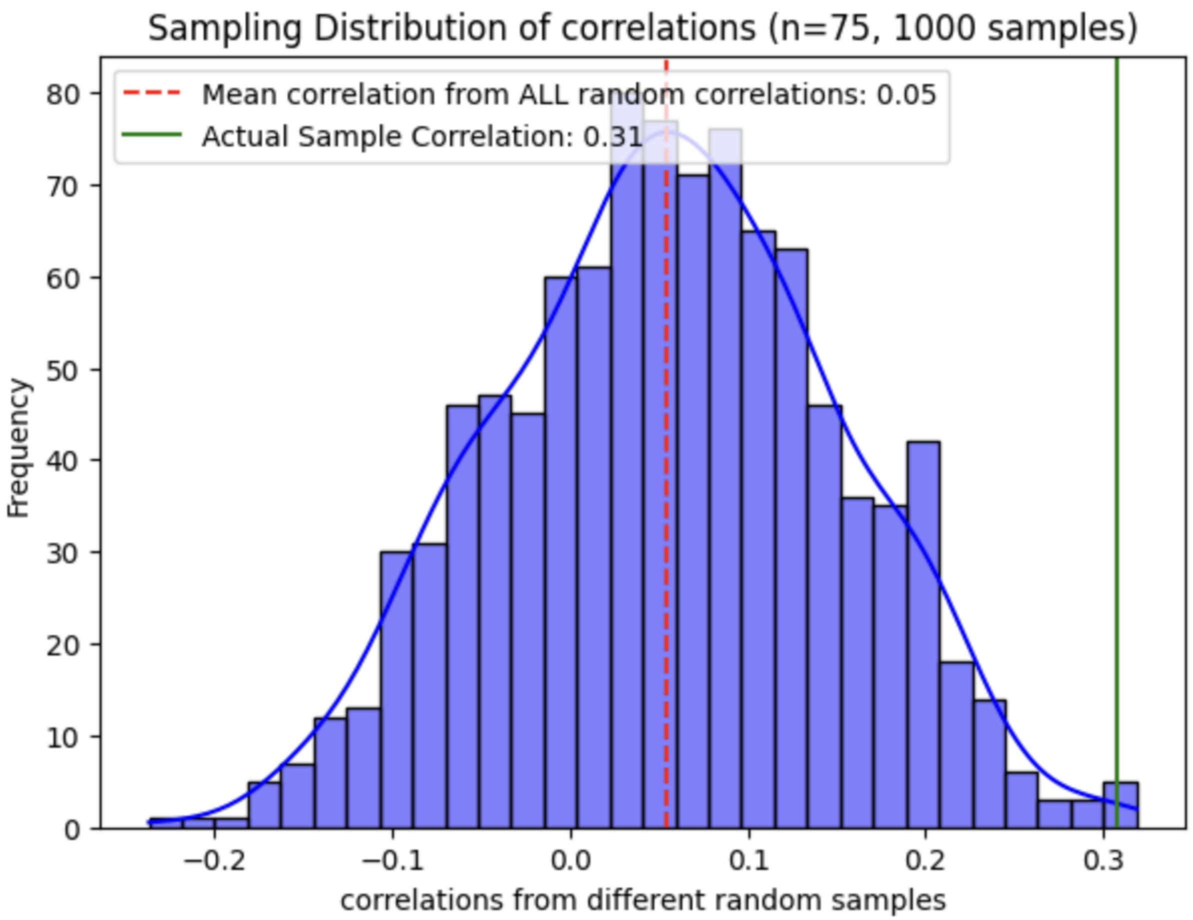


typical sampling distributions

sample statistic	sampling distribution
correlation / slopes	Student's *t* distribution
ratio of squared errors	F-distribution
means	normal (central limit theorem)

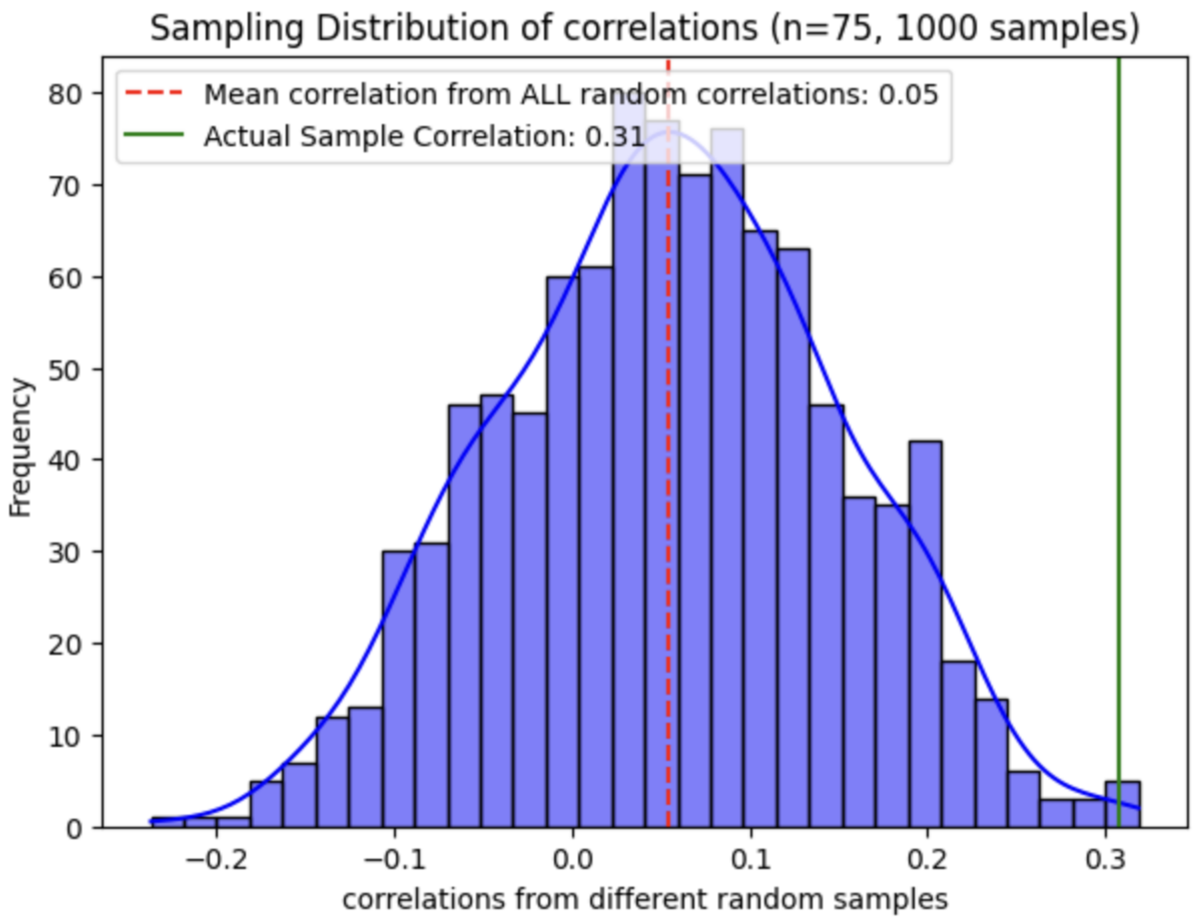
null hypothesis significance testing (NHST)

- start by assuming that our hypothesis is wrong
  - null hypothesis: there is no meaningful relationship between Y (performance) and X (sleep)
    - H_0: population parameter $\rho = 0$
  - alternative hypothesis: there is a meaningful relationship between Y (performance) and X (sleep)
    - H_1: population parameter $\rho \neq 0$
- we generate a sampling distribution under the null hypothesis using the sample statistic



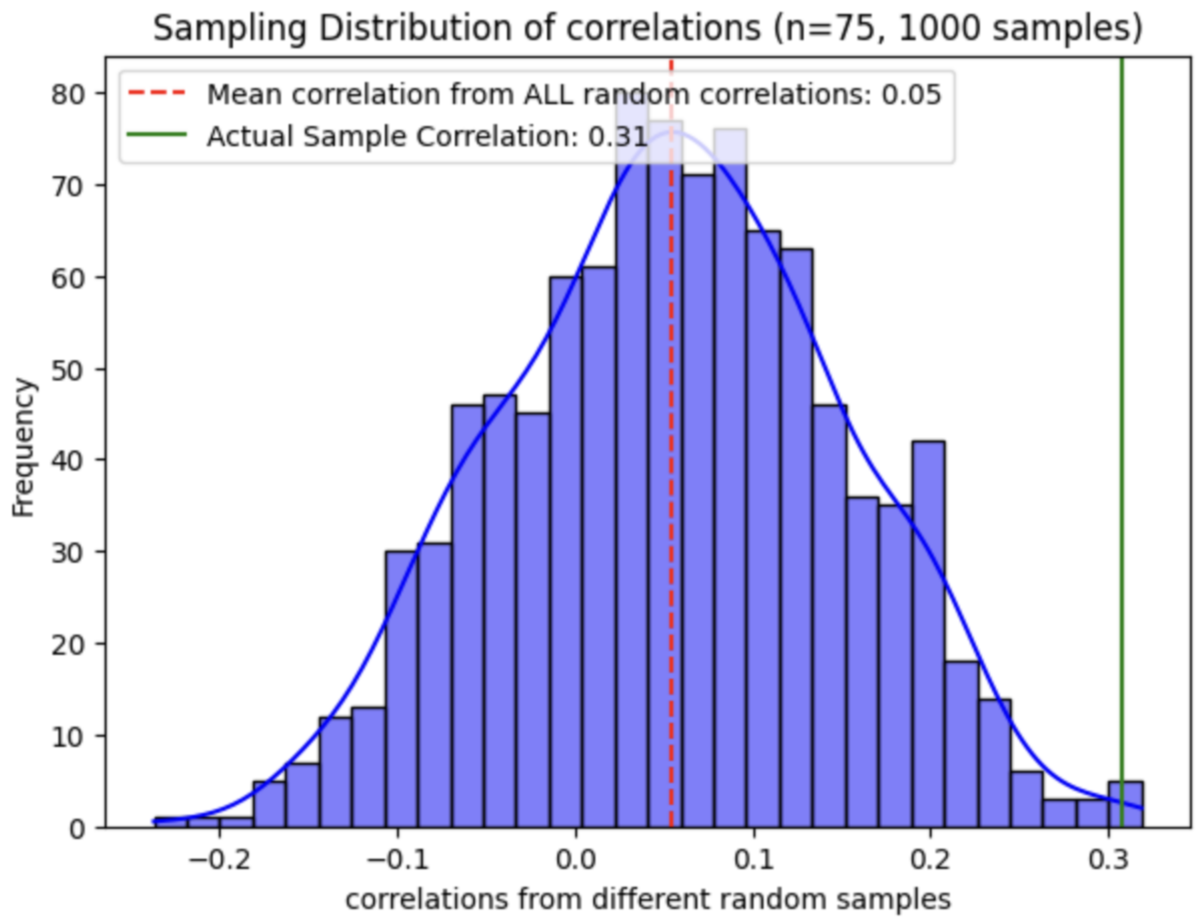
p-value: P (data | null)

- once we have a sampling distribution under the null hypothesis, we want to know how likely is the sample statistic you obtained
- p-value = probability of observing a sample statistic as or more extreme as the one observed if the null hypothesis was true
- if this probability is really low, we can infer that the null hypothesis may not be true, and subsequently infer that your actual hypothesis may be true!



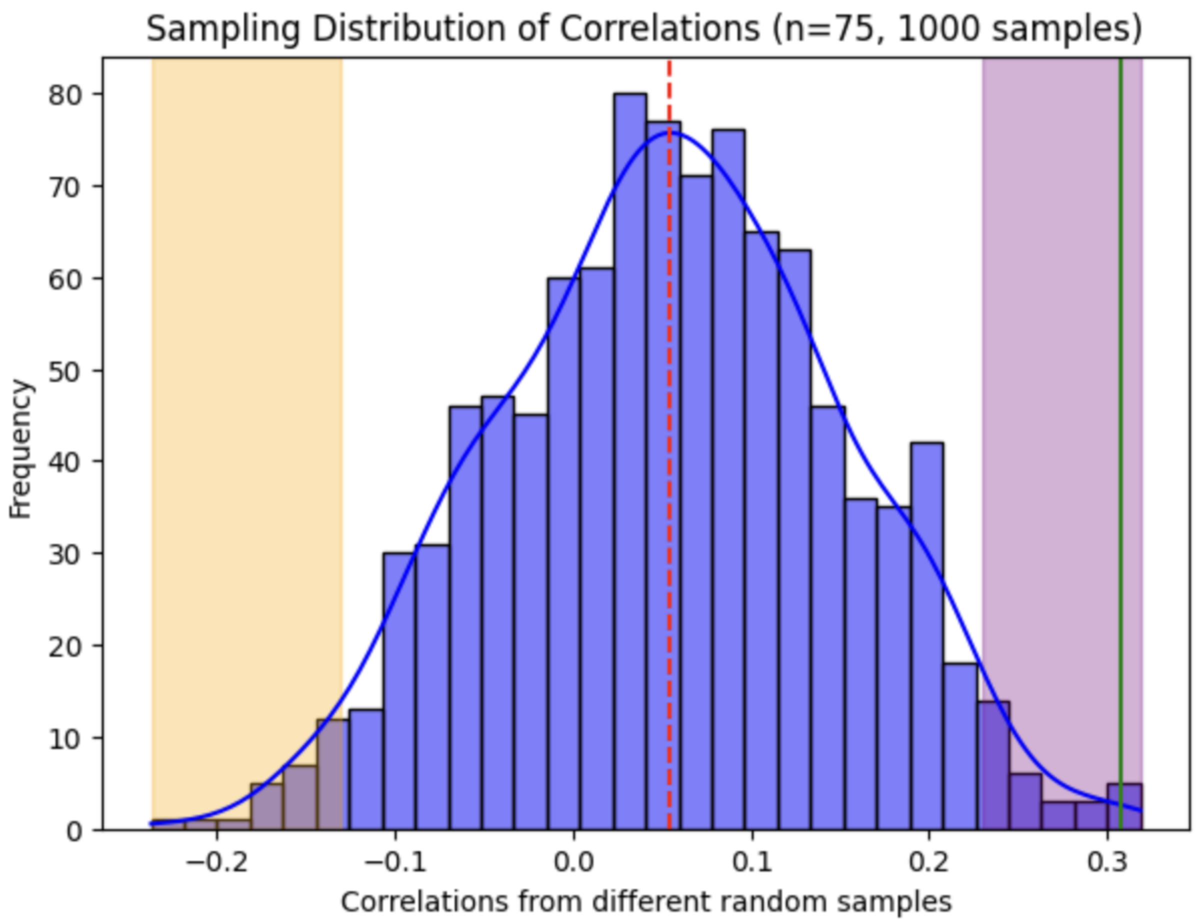
α-level and critical region

- α-level = significance-level criteria
  - typically set to 0.05, i.e., the extreme 5% of the sampling distribution



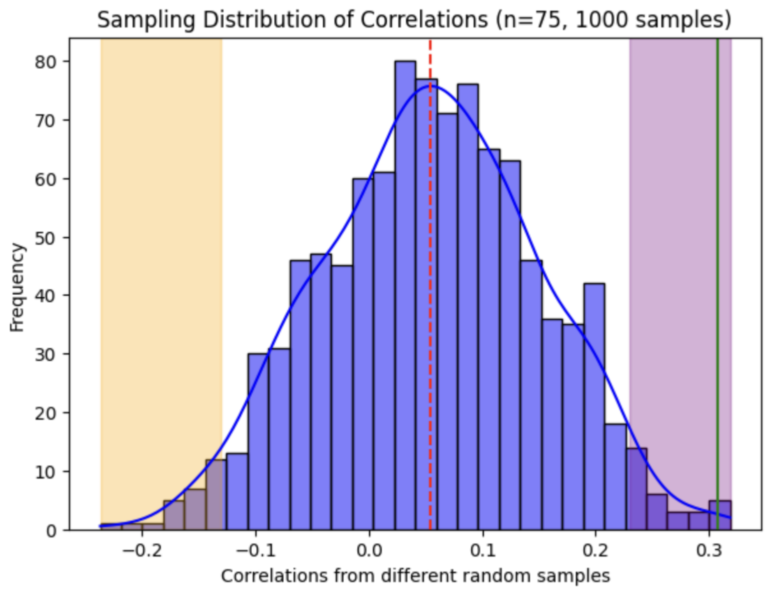
α-level and critical region

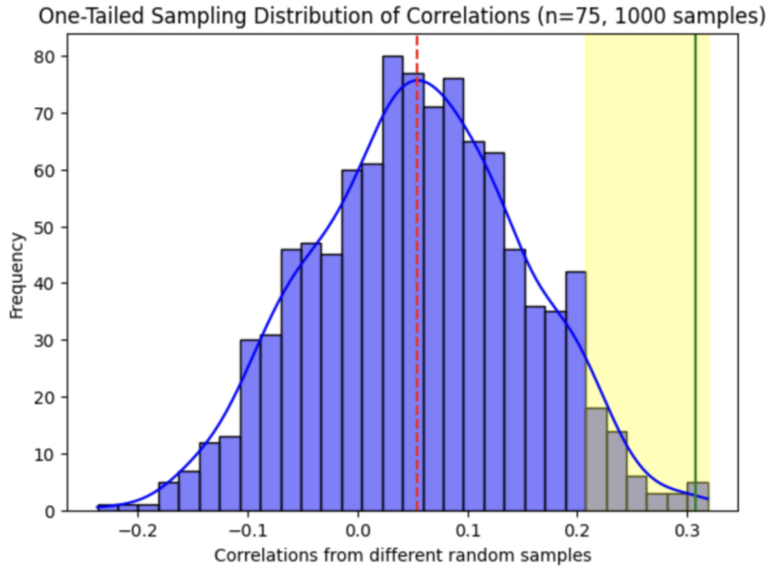
- α-level = significance-level criteria
  - typically set to 0.05, i.e., the extreme 5% of the sampling distribution
  - if the observed statistic is within this critical region, we will reject the null hypothesis in favor of the alternative hypothesis
  - commonly referred to as a statistically significant result
- we typically find a critical value based on the underlying sampling distribution (e.g., $t_{critical}$) using a calculator



one vs. two-tailed tests

- two-tailed tests make no assumptions about directionality when discussing the hypotheses
  - H_o: $\rho = 0$, $H_1: \rho \neq 0$
  - $\alpha = 0.05$ splits the null distribution into two regions (corresponding to p < .025 and p > .025)
- one-tailed (directional) tests specify a direction in the hypotheses, i.e., an increase or decrease in the population parameter
  - $H_o: \rho \leq 0$, $H_1: \rho > 0$
  - $\alpha = 0.05$ is restricted to only ONE part of the null distribution, leading to a larger area
  - more sensitive but also less conservative





W7 Activity 2

W7 Activity 2 debrief

For a two-tailed hypothesis test evaluating a Pearson correlation, what is stated by the null hypothesis?

- ○ there is a non-zero correlation for the sample
- ○ the sample correlation is zero
- ○ there is a non-zero correlation for the general population
- ○ the population correlation is zero

Which of the following accurately describes the critical region?

- ○ outcomes with a very low probability whether or not the null hypothesis is true
- ○ outcomes with a high probability whether or not the null hypothesis is true
- ○ outcomes with a high probability if the null hypothesis is true
- ○ outcomes with a very low probability if the null hypothesis is true

W7 Activity 2 debrief

Which of the following accurately describes a hypothesis test?

- an inferential technique that uses the data from a sample to draw inferences about a population
- a descriptive technique that allows researchers to describe a sample
- a descriptive technique that allows researchers to describe a population
- an inferential technique that uses information about a population to make predictions about a sample

W7 Activity 2 debrief

If a research report includes the term *significant result*, it means that the null hypothesis was accepted.

- ○ True
- ○ False

The null hypothesis is stated in terms of the population, even though the data come from a sample.

- ○ True
- ○ False

summary of NHST

NHST for **correlations**

step 1: state the hypotheses

$H_0: \rho = 0$
$H_1: \rho \neq 0$
find the form of the sampling distribution of correlations under H_0

step 2: set criteria for decision

$\alpha = .05$
find $t_{critical}$ based on one vs. two tailed test and degrees of freedom = $n - 2$

step 3: collect data

(1) compute SE_r for sampling distribution of correlations under H_0

(2) compute $t_{observed} = \frac{r - \rho}{SE_r}$

(3) find p-value for t-score

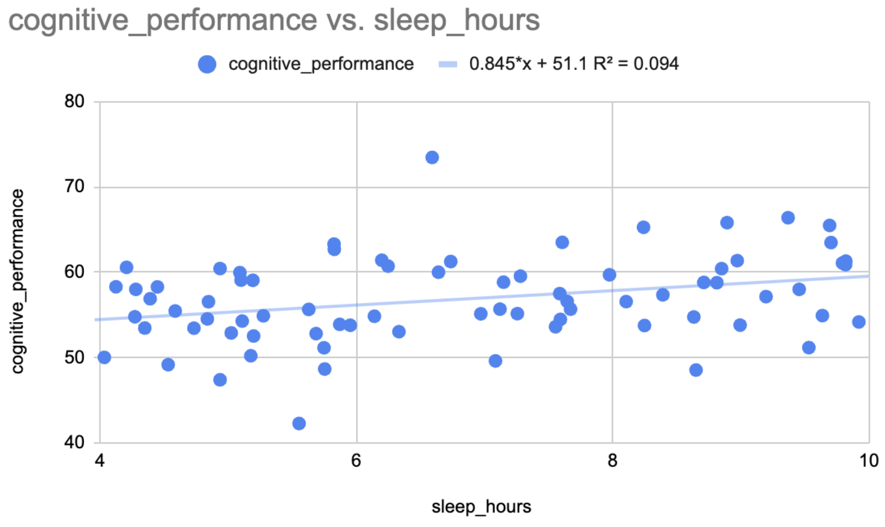
step 4: make a decision!

check whether $t_{observed}$ is beyond $t_{critical}$ and p-value < .05. if so, reject null hypothesis!

example: NHST for correlation

- hypothesis: sleep predicts cognitive performance

sleep_hours	cognitive_performance
6.24724071	60.7189024
9.70428584	63.4792748
8.39196365	57.3561342
7.59195091	54.4745639
4.93611184	47.3953454
4.93596712	60.434202
4.34850167	53.4598411
9.19705687	57.1451403
7.60669007	63.5052718
8.24843547	53.7573958
4.12350697	58.2974841

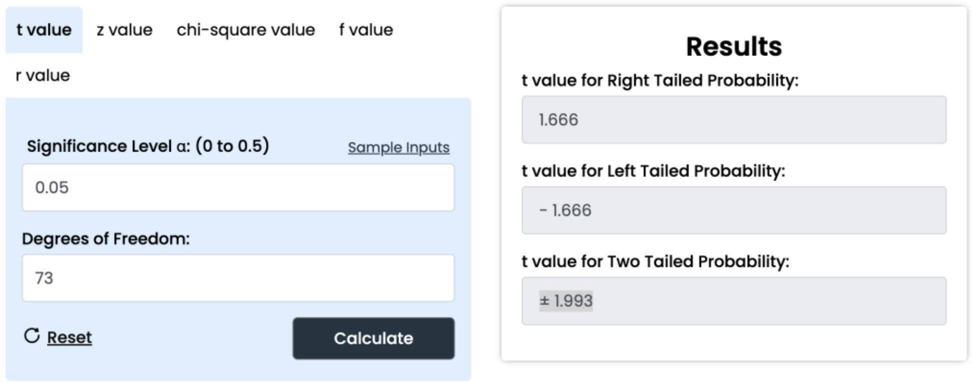


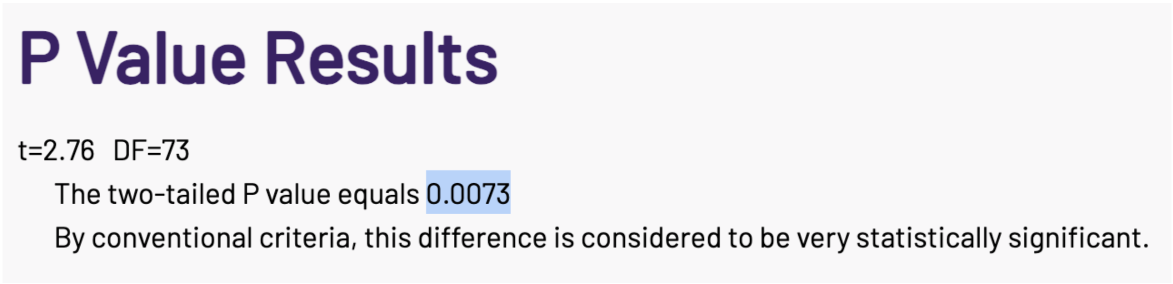
example: NHST for correlation

- **step 1: state the hypotheses**
  - H_0: $\rho = 0$
  - H_1: $\rho \neq 0$
- **step 2: set criteria for decision**
  - $t_{n-2} = t_{73} = t_{critical} = 1.99 \; at \; \alpha = .05$
- **step 3: collect data**
  - compute sample correlation $r = 0.31$
  - compute the standard error for correlation

$$SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}} = .11$$

  - compute the t-statistic: $t_{observed} = \frac{r-0}{SE_r} = \frac{.31}{.11} = 2.76$
  - compute p-value: $p_{observed} = .0073$
- **step 4: decide & report!**
  - sleep is significantly correlated with cognitive performance, $r = .31$, $t(73) = 2.76$, $p = .007$



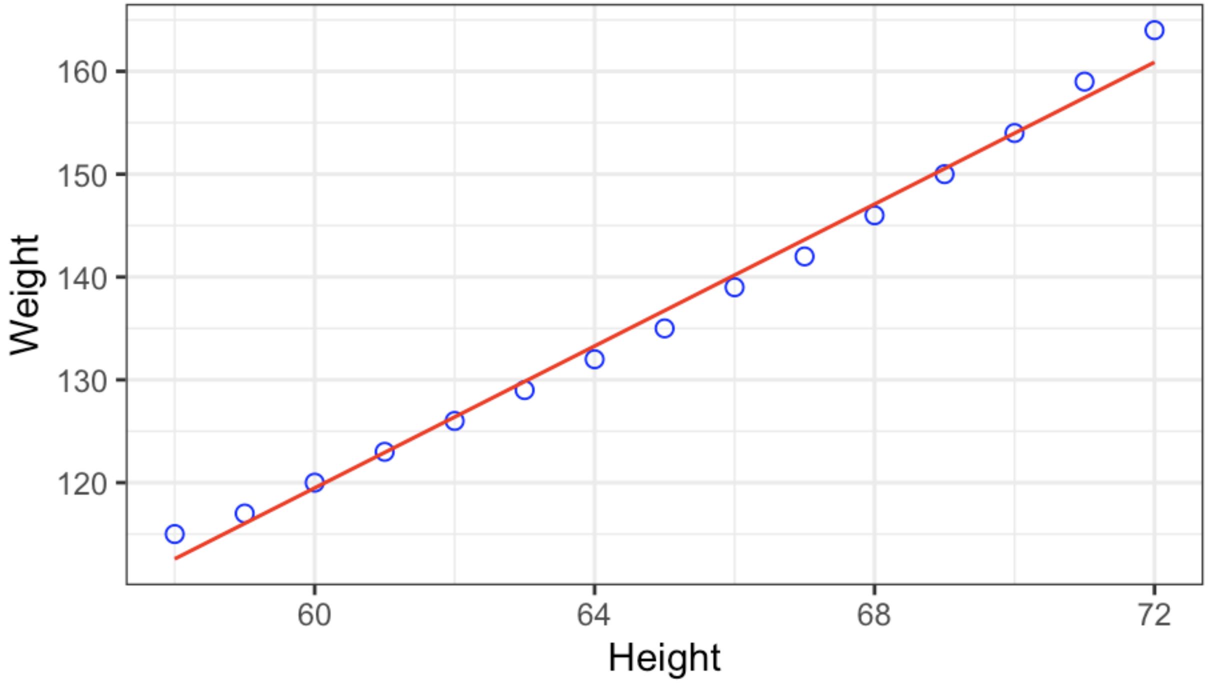


reporting statistical results

- verbal description, sample statistic = X.XX, *test statistic* (sample size) = X.XX, *p value* = .XXX
- statistics are always italicized
- p-values are reported up to third decimal
- everything else reported up to second decimal
- sleep is significantly correlated with cognitive performance,
  r = .31, *t*(73) = 2.76, *p* = .007

activity: NHST for correlation

- women data
- compute correlation and perform a hypothesis test!
- critical value calculator
- p-value calculator

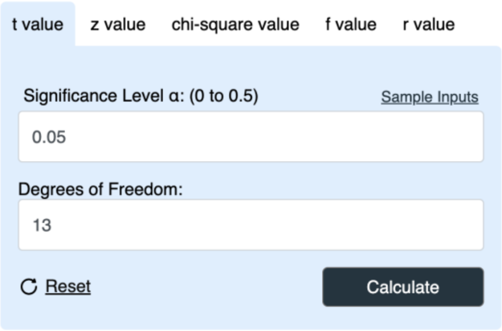


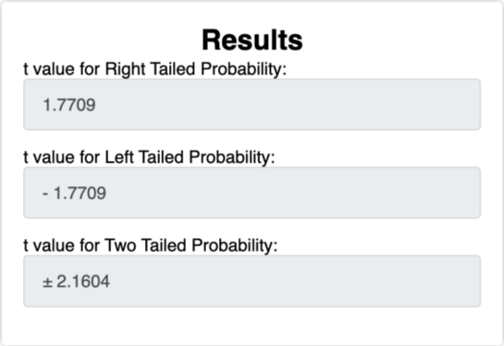
activity: NHST for correlation

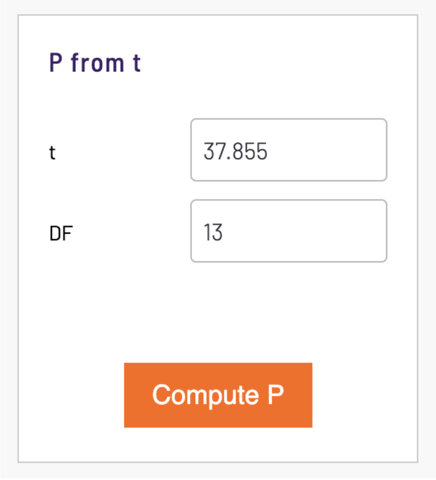
- **step 1: state the hypotheses**
  - $H_0: \rho = 0$
  - $H_1: \rho \neq 0$
- **step 2: set criteria for decision**
  - $t_{n-2} = t_{13} = t_{critical} = 2.16 \; at \; \alpha = \; .05$
- **step 3: collect data**
  - compute the correlation r = 0.995
  - compute the standard error for correlation

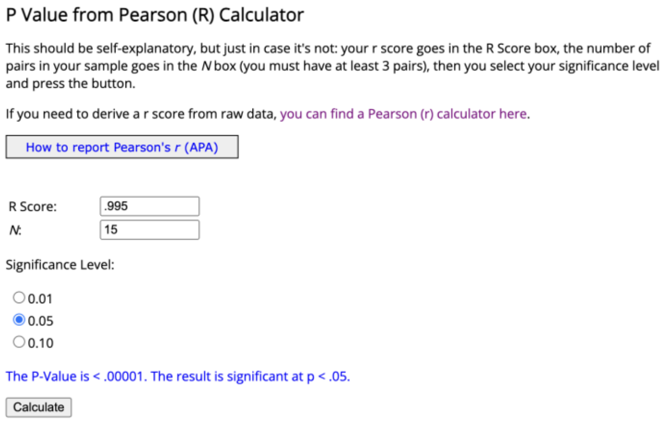
$$SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}} = \; .026$$

  - compute the t-statistic: $t_{observed} = \frac{r-0}{SE_r} = \frac{.995}{.026} = 37.855$
  - compute p-value: $p_{observed} < \; .0001$
- **step 4: decide!**
  - height significantly correlates with weight, r = .995, $t(13)$ = 37.86, p < .001









next time

- **more** hypothesis testing

Before Thursday

- Watch: Hypothesis Testing (Correlations).

After Thursday

- Fill out the Mid-Semester Survey for extra credit!
  - Due after spring break
- See Apply section.

Here are the to-do's for this week:

- Submit Week 7 Quiz
- Submit Problem Set 3 Revision
- Start working on Problem Set 4
- Submit any lingering questions here!
- Have a wonderful spring break!!
- Extra credit opportunities:
  - Submit Mid-Semester Survey
  - Submit Exra Credit Questions
  - Submit Optional Meme Submission