EXAM 2 REVIEW

Exam 2 covers content covered in Lessons 15-22, HW 7-9 and Quizzes 6-7.

Learning Objectives in Scope of Exam 2
Practice these learning objectives using active recall (click here for a description).

1. Joint Random Variables (Lesson 15; including previous concepts on independence from Lesson 10)
   (a) Use joint PMFs to calculate joint probabilities, conditional probabilities and marginal probabilities
   (b) Define and calculate probabilities for multinomial random variables.
   (c) Use Python to simulate random samples from a multinomial distribution.
   (d) State the mathematical definition of what it means for 2 random variables to be independent.
   (e) Determine whether 2 discrete RV are independent using the mathematical definition
   (f) Convert random variables to standard units
   (g) Calculate and interpret covariance and correlation
   (h) Distinguish between association, correlation and causation
   (i) Calculate expectation and variance for sums of random variables
   (j) Define what it means for RV to be IID and determine when RV meet this criteria
2. Sampling and The Central Limit Theorem (Lessons 16-17)
   (a) Explain the differences between a target population, a sampling frame and a sample
   (b) Define 3 common types of sampling bias
   (c) Define a simple random sample
   (d) Explain how to gather an IID sample and explain what the 10% rule is in this context
   (e) Explain the difference between parameters and statistics. Which one is random?
   (f) Define what an estimator means in statistics. Define what it means for an estimator to be unbiased.
   (g) Explain the difference between a population distribution, a sample distribution and the sampling distribution of a statistic
   (h) Simuate the empirical sampling distribution of a statistic
   (i) Apply and interpret the Central Limit Theorem. What is it used for? What conditions must hold for it to apply?
   (j) Define the **standard error** of a statistic.
3. Hypothesis Tests (Lessons 18-20)
   (a) Explain when to use hypothesis tests
   (b) Implement hypothesis tests comparing a sample to a model and interpret results
   (c) Implement a hypothesis test to determine if a sample is consistent with a specific multinomial distribution
   (d) Define null and alternative hypotheses and explain guidelines for how to define them.
   (e) Define the test statistic and explain what it is used for. Provide guidelines for types of test statistics to use based on your null and alternative hypotheses (including TVD).
   (f) Define the significance level, how to choose it and explain what it is used for.
   (g) Define the p-value and interpret it both mathematically as well as visually.
   (h) Describe the types of conclusions you can make in hypothesis testing
   (i) Implement A/B hypothesis tests and interpret conclusions
   (j) Explain how and when to use permutations to simulate under the null
   (k) Explain when you can make causal conclusions from an A/B test.
   (l) Explain the difference between an **empirical p-value** and a **theoretical p-value**.

(m) State the mathematical definition of statistical power. State 3 factors that influence the power of a hypothesis test.

(n) Explain the difference between Type I and Type II errors and how to minimize them.

(o) Define p-hacking and provide guidelines to avoid this.

4. Confidence Intervals (Lessons 21-22)

   (a) Define what a confidence interval is and how to interpret it.

   (b) Calculate confidence intervals using the bootstrapping percentile method.

      i. When we conduct a bootstrap resample, what size resample should we draw from our sample? Why?

      ii. Why do we need to resample from our sample with replacement?

      iii. When we conduct a bootstrap resample, what is the underlying assumption/reasoning for resampling from our sample? Why does it work?

      iv. Explain key assumptions and limitations of the bootstrapping percentile method.

   (c) Use the CLT to create and interpret Confidence Intervals for the population mean

   (d) Use CLT to create and interpret confidence intervals for population proportions

   (e) Determine sample sizes based on desired confidence interval widths

   (f) Explain when and how you can use confidence interaqls to conduct hypothesis tests.

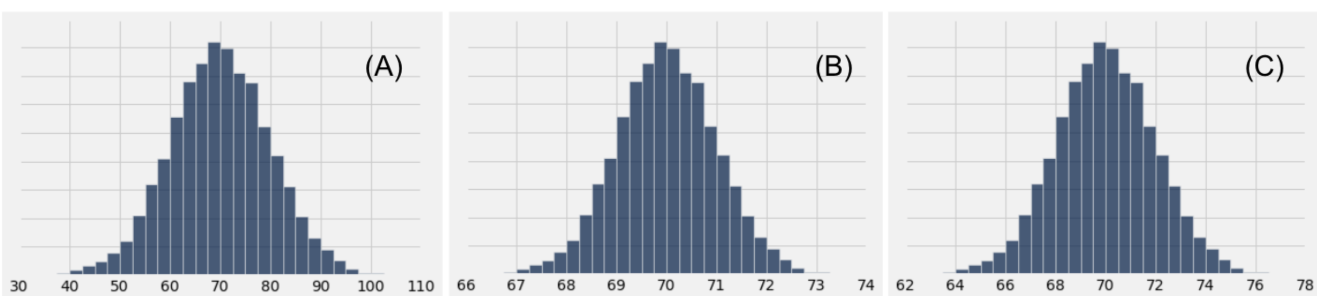
Review Questions

Here are a selection of practice questions for you to use to quiz yourself while studying in addition to HW, quiz and examples we completed in class. These practice questions are listed in random order to give you practice distinguishing what concept/strategy to apply.

Try to answer these questions like you are taking an exam (using only your crib sheet and calculator as resources). The answers to the questions from the slides are provided in the next slide and/or in the video on that topic in the supporting materials for that lesson. The answers to questions not from the slides are posted in a separate Exam 2 Review answers file posted in the modules on Canvas.

1. Lesson 17 Slide 27

   A population distribution has an expected value of 70 and a standard deviation of 10. One of the histograms below is the distribution of the averages of 10,000 random samples of size 100 drawn from the population. Which one? Explain your reasoning.



2. Let X be a discrete random variable that can only take on the values $0, 1, 2, 3$. Let Y be a discrete random variable that can only take on the values $0, 1, 2$.

   The joint probability table for X and Y is given below.

Y \ X	0	1	2	3
0	.16	?	.07	.04
1	.12	.14	.12	0
2	.07	.12	0	0

   a). What is the value of the missing probability?

   b). Are X and Y independent or dependent? Justify your answer mathematically.

   c). What is $P(X = 1|Y = 0)$?

3. Suppose that the rents paid by a random group of 900 people are i.i.d (independent and identically distributed) each with expectation \$1500 dollars and standard deviation \$810.

   There is about 95% chance that the average rent of the 900 people is in the range \$1500 plus or minus $\$x$. Find x.

4.

The dataframe `WELCOME_TBL` contains the results of this semester's welcome survey. The first two rows are shown below. Each row corresponds to a student. In the column **Extraversion**, each student scored themselves on a scale of 1 (not extraverted) to 10 (extremely extraverted).

Year	Extraversion	Number of Textees	Hours of Sleep	Handedness	First Pant Leg	Sleep Position
Second	8	5	6	Right-handed	Right	Left
Second	7	8	7.5	Right-handed	Right	Left

(...1000 rows omitted)

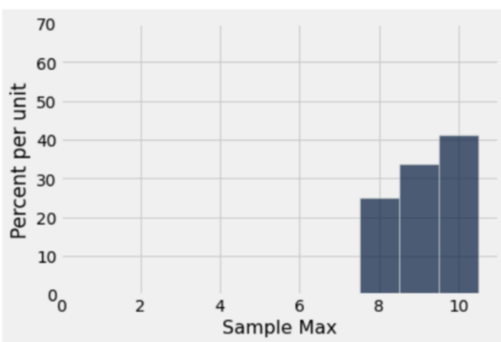
(a) **(4 pt)** Complete the code below to define a function `FUN_NAME` that takes a sample size as its argument. The function should sample that many times at random **without** replacement from all the students and return the maximum extraversion score of the sampled students.

```
def FUN_NAME(...):
    ...
    ...
```

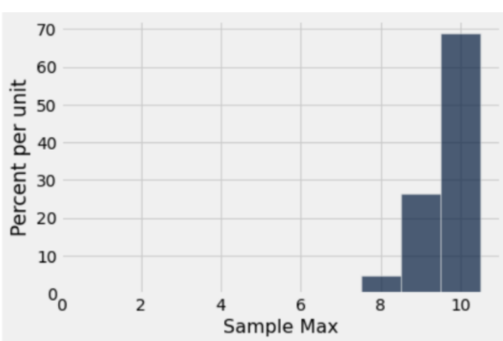
(b) Complete the code below so that the last line evaluates to an array of 10,000 simulated values of the maximum extraversion score in a random sample of size 25 drawn without replacement from all the students. Your code should use the function `FUN_NAME` that you defined above.

```
repetitions = ...
SIM_VALS = ...

for ... in ...:
    ...

SIM_VALS
```

(c) A student mistypes the sample size in the previous question to be 55 instead of 25. One of the histograms below shows the distribution of the maximum values simulated by this student. The other shows the distribution of the maximum values that you simulated using a sample size of 25. Which is which?



A:



B:

○ A is sample of 25, B is sample of 55

○ A is sample of 55, B is sample of 25

5.

The pygmy hippo is a small, reclusive (and cute) hippopotamid type that is native to the forests and swamps of West Africa. Two teams of zoologists set out to estimate the proportion that are male by sampling at random from the population. The first team samples 100 hippos and finds the proportion of males in their sample to be A. The second team samples 40 hippos and finds the proportion of males in their sample to be B. The full population has all 2,500 wild pygmy hippos; the proportion P of males in the population is 50% (but unknown to the zoologists).

i. Which of the following are more likely than not? Select **all** that apply.

☐ A is smaller than B.

☐ A is larger than B.

☐ P is closer to A than B.

☐ P is closer to B than A.

☐ None of these.

ii. Which of the following is largest?

○ The chance that A is above 55%

○ The chance that B is above 55%

○ The chance that A is above 60%

○ The chance that B is above 60%

iii. Which Python expression evaluates to the probability that B is not 0 and not 1, but instead a proportion between 0 and 1?

○ 0

○ 1

○ 0.5 ** 40

○ 1 - (0.5 ** 40)

○ 0.5 ** 40 + 0.5 ** 40

○ 1 - (0.5 ** 40 + 0.5 ** 40)

6. You are assigned to conduct a randomized control trial as part of an effort to encourage high school students from under-resourced communities to apply for college.

   Group A receives special coaching for the ACT. Group B receives no intervention.

   You are interested in determining whether there is a difference in mean ACT scores between Groups A and B that is significant at the $\alpha = 5\%$ level.

   (a) State the null hypothesis and the alternative hypothesis.

   (b) What is a valid test statistic you could use?

   Before conducting the study, you use a power analysis to determine the number of participants to include in your study.

   (c) In a power analysis, the necessary sample size depends on what 3 other variables? Explain what these variables mean in a sentence.

   (d) What are the two types of hypothesis test errors? (Give examples in the context of this problem).

   (e) You conduct your power analysis and find the minimum necessary sample size is 90 participants per group. You end up recruiting 200 participants. A simple random sample of 90 participants received special coaching for the ACT. The remaining participants received no intervention. You find that the 95% confidence interval for the difference in mean ACT scores between those who used the intervention and those who didn't is given by: $[-1, 3]$. What is the conclusion of your hypothesis test? Explain.

7. You are interested in investigating the liters of water consumed every day by CU students. In particular, you want to study the proportion of students drinking less than 3 liters of water per day. You contact 150 random students from the directory and obtain the amounts of water each one of them drinks, storing them in an np.array `water`. The entries of the array are the number of liters of water drunk by each student.

   (a) What is the parameter and what is the statistic in this scenario?

   (b) Write a line of code to calculate the proportion of students in your sample who drank less than 3 liters of water per day.

   (c) Write a line of code to perform a single bootstrap resample of the data stored in the `water` array.

   (d) Write code to conduct 10,000 bootstrap resamples of your data, calculating the proportion of students in each resample that drink less than 3 liters of water per day. Then write code to calculate an 85% confidence interval for the proportion of students in the population who drink less than 3 liters of water per day.

   (e) Your friend tells you that the 85% confidence interval you found above has an 85% probability of containing the true population proportion. Is this the correct interpretation of the meaning of the confidence interval? If not, rewrite a correct interpretation.

8. Ciara is interested in the heights of professional women tennis players. She's collected a random sample of 100 heights of professional women tennis players. She wants to use this sample to estimate the true interquartile range (IQR) of all heights of professional women tennis players.

   (a) In order to construct a 99% confidence interval for the IQR, what should our upper and lower percentile endpoints be?

   (b) Define a function `ci_iqr` that constructs a 99% confidence interval for the IQR as follows. The function takes the following arguments:

   `tbl`: A dataframe consisting of a random sample from the population. The height data is in a column titled `height`.

   `reps`: The number of bootstrap repetitions.

9. Suppose you're only given the following information about two joint random variables X and Y:

$$\mu_X = 2, \ \ \mu_Y = 5, \ \ \sigma_X^2 = 4, \ \ \sigma_Y^2 = 9 \text{ and } \rho(X, Y) = \frac{3}{8}$$

   .

   For each of the quantities below, calculate if you have enough information, showing all steps. If not, explain what additional info you'd need.

   i). $E[X + Y]$

   ii). $Cov(X, Y)$

   iii). $Var[X + Y]$

   iii). $E[XY]$

10. Lesson 17 Slide 26

Determine if you have enough information to answer each of the following. If so, answer them. If not, describe what additional information you'd need.

Suppose salaries at a very large corporation have a mean of $162,000 and a standard deviation of $32,000.

a). If a single employee is randomly selected, what is the probability that their salary exceeds $175,000?

b). If 100 employees are randomly sampled, what is the probability that their average salary exceeds $175,000?

```
from scipy import stats
```

```
stats.norm.cdf(175000, 162000, 32000/10)
```

```
0.9999757250261433
```

11. Lesson 17 Slide 39

True or False:

A). No matter what population you are drawing from, the sample distribution is roughly normal (for large enough n).

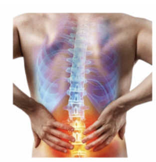
B). No matter what population you are drawing from, the sampling distribution of the sample mean is roughly normal (for large enough n)

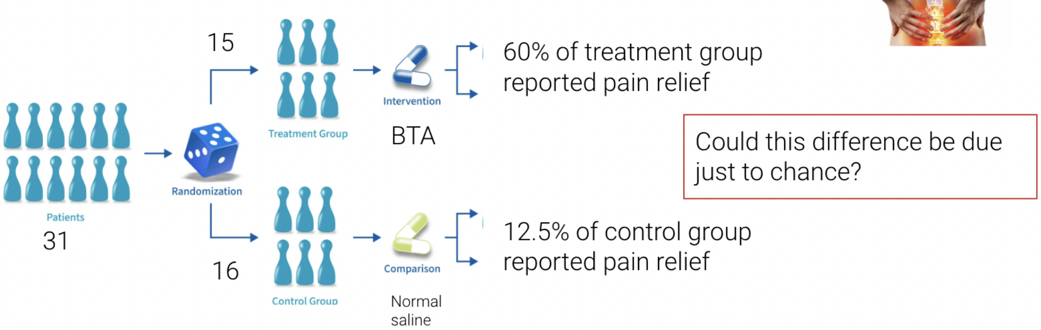
C). No matter what population you are drawing from, the sampling distribution of the sample median is roughly normal (for large enough n)

D). If you are drawing from a Bernoulli distribution, the sampling distribution of the sample proportion is roughly normal (for large enough n)

12. Lesson 19 Slide 21

A randomized controlled trial (RCT) examined the effect of using Botulinum Toxin A (BTA) as a treatment for low-back pain.





The trials were run double-blind so that neither doctors nor patients knew which group they were in.

Set-up and conduct a hypothesis test to determine if the difference in the outcomes in this trial is significantly statistically significant.

13. Lesson 20 Slide 20

    Manufacturers of Super Soda run a taste test. 91 out of 200 tasters prefer Super Soda over its rival. Question: Do fewer people prefer Super Soda than its rival, or is this just chance? Set-up and conduct a hypothesis test to answer this question. Include the following in your test:

    Null hypothesis:

    Alternative hypothesis:

    Test statistic:

    p-value: Start at the observed statistic and look which way?

    Write code to simulate the distribution of the test statistic under the null.

    Write code to calculate the empirical p-value.

    What is the conclusion of your test?

14. Lesson 22 Slide 29

    I am going to use a 68% confidence interval to estimate a population proportion.

    I want the total width of my interval to be no more than 2.5%.

    What is the minimum size I should use for my sample?

15. Lesson 22 Slide 30

    A researcher is estimating a population proportion based on a random sample of size 10,000.

    Fill in ?? with a decimal:

    With chance at least 95%, the estimate will be correct to within ??

16. Lesson 22 Slide 35

    Consider the following hypothesis test:

    Null: The average age of mothers in the population is 25 years; the random sample average is different due to chance.

    Alternative: The average age of the mothers in the population is **not** 25 years.

    Suppose you use the 5% cutoff for the p-value.

    In the previous example we found the 95% confidence interval for the average age of mothers in the population is (26.9, 27.6).

    Based on this information, what should you conclude for your hypothesis test and why?

    A). Reject the null B). Fail to reject the null

CONT'D ON NEXT PAGE

17). Simulation and Hypothesis Testing

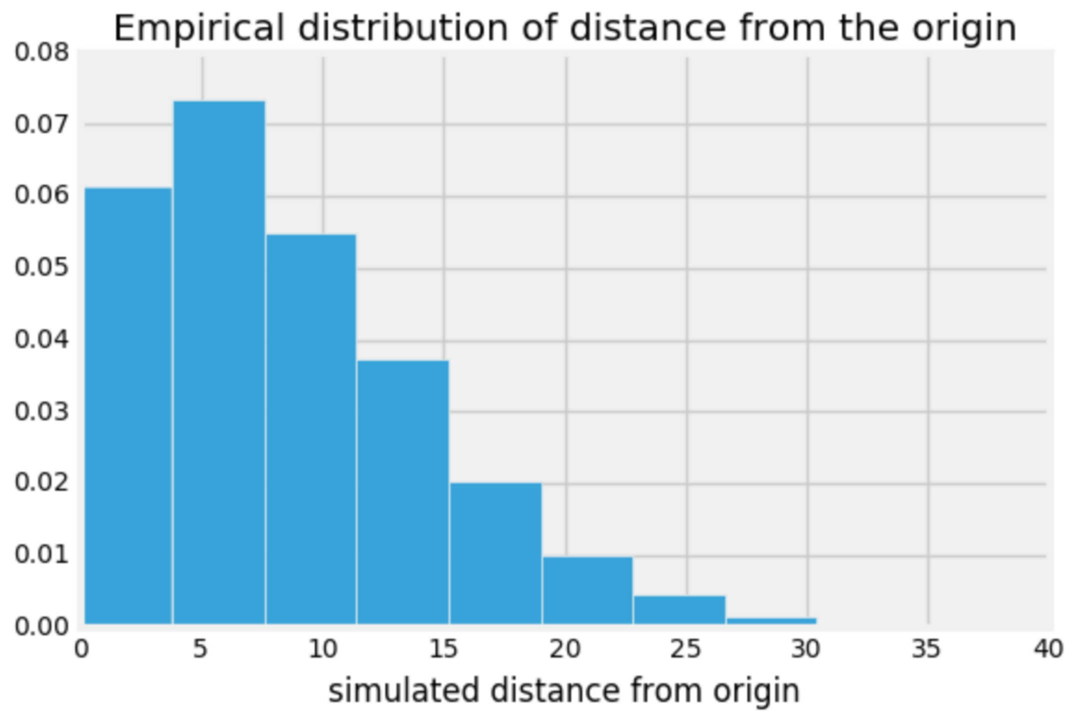
Achilles the turtle sits on the number line. Achilles loves long random walks that last a total of 100 time steps. At each time step, Achilles moves based on the following scheme: He flips a coin and moves one step to the right if the coin comes up heads or one step to the left if the coin comes up tails.

a). Assuming that Achilles' coin is fair, write a function called `one_walk` that simulates one random walk of 100 time steps and returns how far from the origin Achilles ends up at the end of his walk. You may assume that Achilles always starts from the origin.

```
def one_walk():
```

b). Assuming that Achilles' coin is fair, we would like to simulate what would happen if Achilles took 10000 different random walks. Complete the simulation below and keep track of how far Achilles ends up from the origin in each of his walks in an array called `distances`. The histogram shown below is an example of a histogram plotted from `distances`.

```
distances =

for i in np.arange(10000):
    new_distance = ___________________________________________________

    distances = ______________________________________________________
```



c). Achilles goes for a walk and claims that at the end of his walk, he ended up 30 steps away from the origin. You notice this is strange, so you want to run a hypothesis test to test whether or not Achilles used a fair coin. Fill in the blanks below for the null and alternative hypotheses as well as a good test statistic for this experiment.

Hint: When considering your alternative hypothesis, note that we do not really care about whether the coin is biased towards heads or towards tails.

Null Hypothesis:

Alternative Hypothesis:

Test Statistic:

d). Write the code to calculate the p-value given the test statistic listed above and using a 5% p-value cut-off. Then, describe the different conclusions that you would arrive at depending on the p-value.

Hint: We simulated an array in part(b) of test statistics under the null hypothesis. Try to use the `distances` *array.*

```
p_value = ____________________________________________________________
```

18). True/False

Respond with true or false to the following questions. If your answer is false, explain why.

1. In the U.S. in 2000, there were 2.4 million deaths from all causes, compared to 1.9 million in 1970, which represents a 25% increase. The data shows that the public's health got worse over the period 1970-2000.

2. A company is interested in knowing whether women are paid less than men in their organization. They share *all* their salary data with you. An A/B test is the best way to examine the hypothesis that all employees in the company are paid equally.

3. Consider a randomized control trial where participants are randomly split into treatment and control groups. There will be no systematic differences between the treatment and control groups if the process is followed correctly.

4. A researcher considers the following scheme for splitting a people into control and treatment groups. People are arranged in a line and for each person, a fair, six-sided die is rolled. If the die comes up to be a 1 or a 2, the person is allocated to the treatment group. If the die comes up to be a 3, 4, 5 or 6 then the person is allocated to the control group. This is a randomized control experiment.

5. You are conducting a hypothesis test to check whether a coin is fair. After you calculate your observed test statistic, you see that its p-value is below the 5% cutoff. At this point, you can claim with certainty that the null hypothesis can not be true.

6. You roll a fair die a large number of times. While you are doing that, you observe the frequencies with which each face appears and you make the following statement: As I increase the number of times I roll the die, the probability histogram of the observed frequencies converges to the empirical histogram.

19). **Multiple Choice**

. Gary is playing with a coin and he wants to test whether his coin is fair. His experiment is to toss the coin 100 times. He chooses the following null hypothesis:

Null Hypothesis: The coin is fair and any deviation observed is due to chance.

For each of the alternative hypotheses listed below, determine whether or not the test statistic is valid.

a. **Alternative Hypothesis:** The coin is biased towards heads.
Test Statistic: # of heads

b. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: # of heads

c. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: |# of heads - expected # of heads|

d. **Alternative Hypothesis:** The coin is biased towards heads.
Test Statistic: |# of heads - expected # of heads|

e. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: ½ - proportion of heads

20). Fun with Functions

1. Write a function called `compute_pvalue` that, given an empirical distribution in the form of an array and the observed value of your test statistic, calculates the p-value for that test statistic. You may assume that large values of your test statistic provide evidence against the null hypothesis.

```
def compute_p_value(empirical_dist, observed_ts):
```

2. Now write a function called `is_significant` that takes in an empirical distribution, the observed test statistic and a p-value cutoff, returns `True` if the p-value of the observed test statistic is statistically significant based on the cutoff provided and `False` otherwise.

Hint: Use the function you defined in Question 1!

```
def is_significant(empirical_dist, observed_ts, cutoff):



    return
```

21). More Hypothesis Testing

Chloe is a big fan of Trader Joes' frozen mac n cheese, but she noticed that the cheese used in it varies from box to box. A Trader Joe's employee provides her with some data about the 4 different cheeses used and the probability of them being used in each box:

Cheese	**Probability**
Velveeta	0.05
Gruyère	0.55
Sharp Cheddar	0.25
Monterey Jack	0.15

Chloe is suspicious about this distribution. After all, Velveeta is much cheaper to use than Gruyère, and she has also never bought a box that uses Gruyère. Chloe decides to buy many boxes throughout the next month and tracks the type of cheese used in each box. She uses this to conduct a hypothesis test.

1. Write the correct null hypothesis for this experiment
   - Null Hypothesis:
   - Alternative Hypothesis:

```
observed_proportions =  np.array([0.2, 0.3, 0.45, 0.05])
employee_proportions = np.array([0.05, 0.55, 0.25, 0.15])
```

The array `observed_proportions` contains the proportions of cheese that Chloe observed in 20 boxes of Mac n Cheese.

2. Chloe wants to use the mean as a test statistic, but Katherine suggests that she uses the TVD (total variation distance) instead. Which test statistic should Chloe use in this case? Briefly justify your answer. Then write a line of code to assign the observed value of the test statistic to `observed_stat`.

```
observed_stat = ______________________________________
```

3. Define the function `one_simulated_test_stat` to simulate a random sample according to the null hypothesis and return the test statistic for that sample.

```
def one_simulated_test_stat():
      sample_prop = _________________________________
      return ________________________________________
```

4. Chloe simulates the test statistic 10,000 times and stores the results in an array called simulated_stats. The observed value of the test statistic is stored in `observed_stat`. Complete the code below so that it evaluates to the p-value of the test:

```
_____________(simulated_stats _________ observed_statistic) /
_____________
```

5. Given that the computed p-value is 0.0825, which of the following are true? Select all that may apply.
   a. Using an 8% p-value cutoff, the null hypothesis should be rejected
   b. Using a 10% p-value cutoff, the null hypothesis should be rejected.
   c. There is an 8.25% chance that the null hypothesis is true

d. There is an 8.25% chance that the alternative hypothesis is true

22). A/B Testing

Choose True/False for each of the statements below, and explain your answer.

a) A/B testing is used to determine whether or not we believe two samples come from the same underlying distribution.

b) To conduct a permutation test, you should sample your data with replacement with a sample size equal to the number of rows in the original table.

c) A/B testing is the same as using total variation distance as a test statistic for a hypothesis test.

23). You and your friend are both huge Warriors fans, and are watching a game together. You're enjoying the absolute dominance the Warriors are displaying as a team. You attribute the Warriors' recent success to Stephen Curry's shooting skills, but your friend argues that Klay Thompson has actually been the better shooter as of late. Since you are a data scientist who loves Stephen Curry, you decide to test your friend's claim.

You decide to compare the two players by taking a sample of games from this year and finding the field goal percentages (the percentage of shots a player makes in a game) for both Curry and Thompson. You record these in a table, and at a glance, it seems that Curry has better field goal percentages. You decide to run an A/B test to determine if this is just due to chance.

a) Choose a null hypothesis for your A/B test.

b) Choose an alternative hypothesis.

c) Choose a test statistic.

d) Describe how you would use a permutation test to simulate your test statistic and calculate a p-value.

So far in the course, you have used the bootstrap to estimate multiple different parameters of a population such as the maximum, median, and mean. You are now capable of building *empirical distributions* for these sample statistics. An empirical distribution for a sample statistic is obtained by repeatedly resampling and calculating the statistic for those resamples. However, there is special theory, namely the **Central Limit Theorem**, that tells us the empirical distribution of the *sample mean* is unique: if you draw a large random sample **with replacement** from a population, then, regardless of the distribution of the population, the probability distribution for that sample's mean is roughly normal, centered at the population mean.

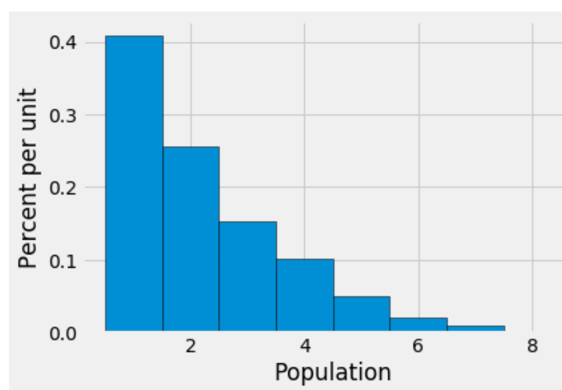
Furthermore, the *standard deviation* (spread) of the distribution of sample means is governed by a simple equation, shown below:

$$\text{SD of all possible sample means} = \frac{\text{Population SD}}{\sqrt{\text{sample size}}}$$

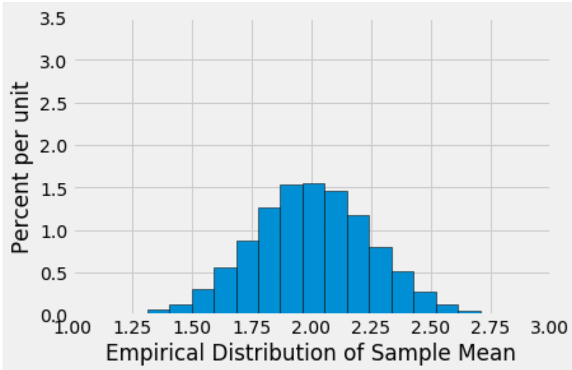
"SD of the distribution of all sample means" is the same thing as saying "sample mean SD".

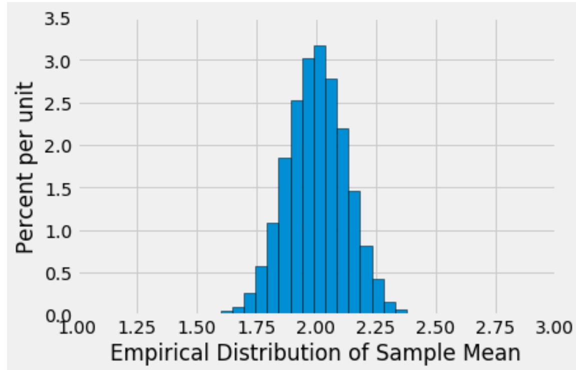
24). **Sample Means**: Assume that you have a certain population of interest whose histogram is to the right.

a) Aarushi takes many large random samples **with replacement** from the population with the goal of generating an empirical distribution of the sample mean. What shape do you expect this distribution to have? Which value will it be centered around?



b) Suppose that Aarushi creates two empirical distributions of sample means, with different sample sizes. Which distribution corresponds to a larger sample size? Why?





c) Suppose you were told that the distribution on the left has a standard deviation of 0.3 and was generated based on a sample size of 100. How big of a sample size would you need if you wanted the standard deviation of my distribution of sample means to be 0.03 instead?

25). **Confidence Intervals:** You are working with Oscar on constructing a confidence interval for the mean height of all Berkeley students. You take a random sample of 400 Berkeley students and compute the mean height of students in the sample; it is 170 cm. We also calculate the standard deviation of our sample to be 10 cm.

a) Oscar claims that the distribution of all possible sample means is normal with SD 0.5 cm. Use this information to construct an approximate 95% confidence interval for the mean height of all Berkeley students.

Hint: If you know the distribution is normal, what do you know about the proportion of values that lie within a few SDs of its mean?

b) If Oscar hadn't told you what the SD of the sample mean was, could you estimate it from the data in the sample? If yes, how?

c) Does your answer from part (b) agree with what Oscar claims in part (a)?

26). **Standard Units and Correlation**

a) When calculating the correlation coefficient, why do we convert data to standard units?

b) Write a function called `convert_su` which takes in an array of elements called data and returns an array of the values represented in standard units.

```
def convert_su(data):

    sd = ________________________________________________

    mean = ______________________________________________

    return ______________________________________________
```

27).

Lesson 18 Slide 83 . (And see See Jupyter nb Lesson 18 for solutions).

Example: Consider a Calculus class with 359 students divided into 12 recitation sections. TA's lead the sections. After the midterm, students in Section 3 notice that the average score in their section is lower than in others. They complain that it must be due to their particular TA.

Design a Hypothesis Test to test the students' concerns.

a). What is the null?

b). What is the alternative?

c). What is a valid test statistic?

d). How would you simulate (or calculate theoretically) the test statistic distribution under the null?