From the Winter 2023 Final:

Problem 16

We collect data on the play times of 100 games of *Chutes and Ladders* (sometimes known as *Snakes and Ladders*) and want to use this data to perform a hypothesis test.

Problem 16.1

Which of the following pairs of hypotheses can we test using this data?

Option 1: **Null Hypothesis:** In a random sample of Chutes and Ladders games, the average play time is 30 minutes. **Alternative Hypothesis:** In a random sample of Chutes and Ladders games, the average play time is not 30 minutes.

Option 2: **Null Hypothesis:** In a random sample of Chutes and Ladders games, the average play time is not 30 minutes. **Alternative Hypothesis:** In a random sample of Chutes and Ladders games, the average play time is 30 minutes

Option 3: **Null Hypothesis:** A game of Chutes and Ladders takes, on average, 30 minutes to play. **Alternative Hypothesis:** A game of Chutes and Ladders does not take, on average, 30 minutes to play.

Option 4: **Null Hypothesis:** A game of Chutes and Ladders does not take, on average, 30 minutes to play. **Alternative Hypothesis:** A game of Chutes and Ladders takes, on average, 30 minutes to play.

- Option 1
- Option 2
- Option 3
- Option 4

From the Winter 2023 Final:

Problem 16.2

We use our collected data to construct a 95% CLT-based confidence interval for the average play time of a game of *Chutes and Ladders*. This 95% confidence interval is [26.47, 28.47]. For the 100 games for which we collected data, what is the mean and standard deviation of the play times?

mean : 27.47, SD = 5

Problem 16.3

Does the CLT say that the distribution of play times of the 100 games is roughly normal?

- Yes
- ✓ No

Problem 16.4

Of the two hypotheses you selected in part (a), which one is better supported by the data?

- Null Hypothesis
- ✓ Alternative Hypothesis

$$\frac{\text{sample SD}}{\sqrt{\text{size}}} = \frac{\text{sample SD}}{10}$$

95% CI: sample mean ± 2 · SD of sample mean's distribution

$$\text{mean} + \frac{2 \cdot \text{sample SD}}{10} = 28.47 \rightarrow 27.47 + \frac{5\,SD}{5} = 28.47$$

$$\Rightarrow \frac{5SD}{5} = 1$$

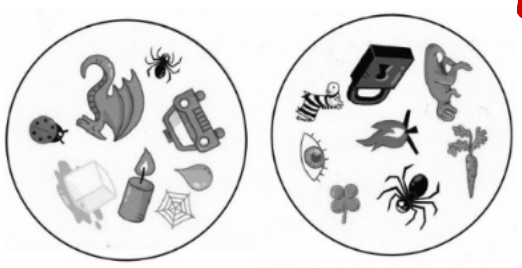
$$\Rightarrow 5SD = 5$$

$$\text{mean} - \frac{2 \cdot \text{sample SD}}{10} = 26.47$$

From the Winter 2023 Final:

Problem 12

In the game *Spot It*, players race to identify an object that appears on two different cards. Each card contains images of eight objects, and exactly one object is common to both cards.



```
def find_match(ob1, ob2):
    for x in ob1:
        if x in ob2:
            return x
```

Problem 12.1

Suppose the objects appearing on each card are stored in an array, and our task is to find the object that appears on both cards. Complete the function `find_match` that takes as input two arrays of 8 objects each, with one object in common, and returns the name of the object in both arrays.

For example, suppose we have two arrays defined as follows.

```
objects1 = np.array(["dragon", "spider", "car", "water droplet", "spiderweb", "candle", "ice c
objects2 = np.array(["zebra", "lock", "dinosaur", "eye", "fire", "shamrock", "spider", "carrot
```

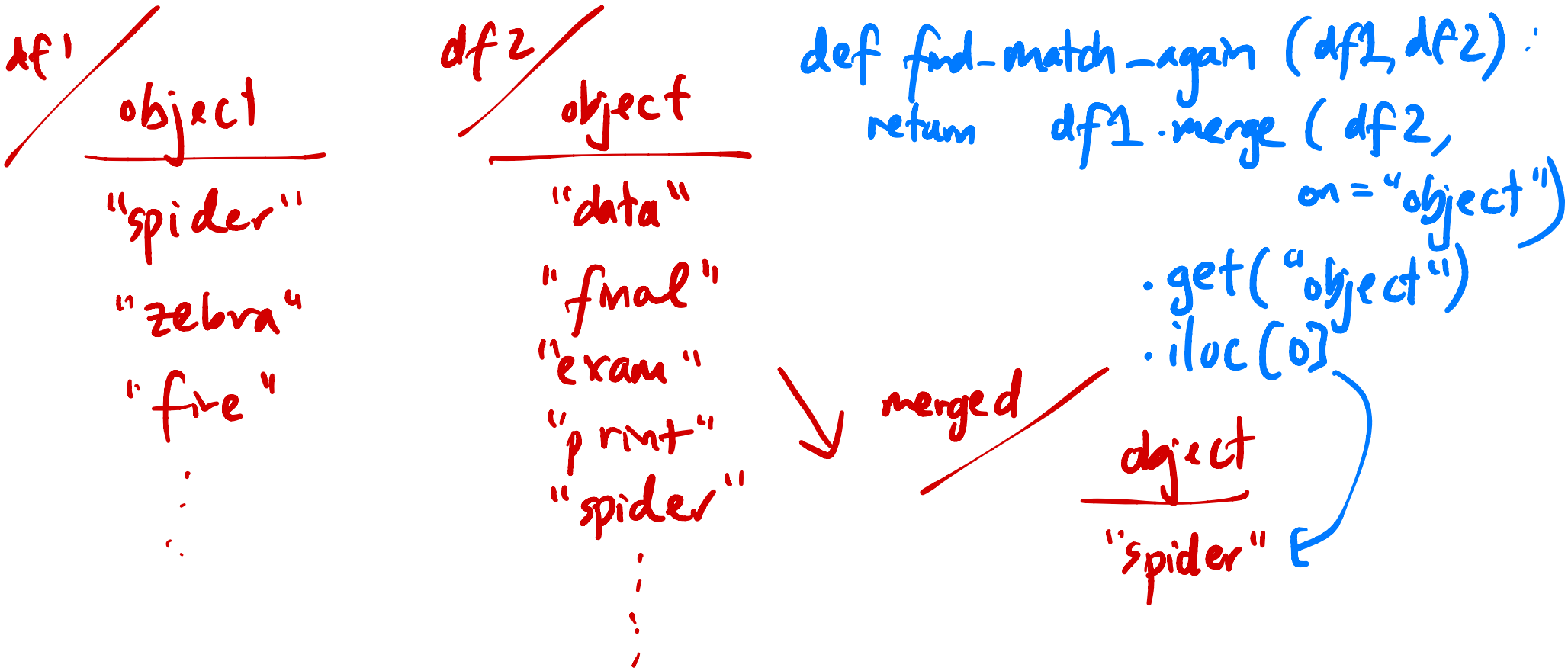
Then `find_match(objects1, objects2)` should evaluate to `"spider"`. Your function must include a for loop, and it must take **at most three lines of code** (not counting the line with `def`).

From the Winter 2023 Final:

Problem 12.2

Now suppose the objects appearing on each card are stored in a DataFrame with 8 rows and one column called `"object"`. Complete the function `find_match_again` that takes as input two such DataFrames with one object in common and returns the nameof the object in both DataFrames.

Your function may not call the previous function `find_match`, and it must take exactly **one line of code** (not counting the line with `def`).



pred $y_{(su)} = r \cdot x_{(su)}$

From the Fall 2022 Final:

income : x
age : y

Problem 6

In this question, we'll explore the relationship between the ages and incomes of credit card applicants.

$x_{(su)} = \frac{8}{3}$, pred $y_{(su)} = \frac{4}{5}$

$\frac{4}{5} = r \cdot \frac{8}{3} \Rightarrow r = \frac{4}{5} \cdot \frac{3}{8_2} = \boxed{\frac{3}{10}}$

Problem 6.1

The credit card company that owns the data in apps, BruinCard, has decided not to give us access to the entire `apps` DataFrame, but instead just a sample of `apps` called `small apps`. We'll start by using the information in `small_apps` to compute the regression line that predicts the age of an applicant given their income.

For an applicant with an income that is $\frac{8}{3}$ standard deviations above the mean income, we predict their age to be $\frac{4}{5}$ standard deviations above the mean age. What is the correlation coefficient, r, between incomes and ages in `small_apps`? Give your answer as a **fully simplified fraction**.

Click to view the solution.

Problem 6.2

Now, we want to predict the income of an applicant given their age. We will again use the information in `small_apps` to find the regression line. The regression line predicts that an applicant whose age is $\frac{4}{5}$ standard deviations above the mean age has an income that is s standard deviations above the mean income. What is the value of s? Give your answer as a **fully simplified fraction**.

Click to view the solution.

6.2: age: x
income: y

$x_{(su)} = \frac{4}{5}$ pred $y_{(su)} = ???$

pred $y_{(su)} = r \cdot x_{(su)}$

$= \frac{3}{10_5} \cdot \frac{4}{5}^2 = \boxed{\frac{6}{25}}$

From the Fall 2022 Final:

Problem 6.3

BruinCard has now taken away our access to both `apps` and `small_apps`, and has instead given us access to an even smaller sample of `apps` called `mini_apps`. In `mini_apps`, we know the following information: - All incomes and ages are positive numbers. - There is a positive linear association between incomes and ages.

We use the data in `mini_apps` to find the regression line that will allow us to predict the income of an applicant given their age. Just to test the limits of this regression line, we use it to predict the income of an applicant who is **-2 years old**, even though it doesn't make sense for a person to have a negative age.

Let I be the regression line's prediction of this applicant's income. Which of the following inequalities are guaranteed to be satisfied? Select all that apply.

- $I < 0$
- $I <$ mean income
- $|I -$ mean income$| \leq |$mean age $+ 2|$
- $\dfrac{|I - \text{mean income}|}{\text{standard deviation of incomes}} \leq \dfrac{|\text{mean age} + 2|}{\text{standard deviation of ages}}$
- None of the above.

Click to view the solution.

x: age
y: income

all $x_s, y_s > 0$

$r > 0$

$$\left|\frac{-2 - \text{mean age}}{\text{SD of age}}\right| = \frac{|\text{mean age} + 2|}{\text{SD of age}}$$

$$|-2_{(su)}| = \left|\frac{-2 - \text{mean age}}{\text{SD of age}}\right| = \frac{|\text{mean age} + 2|}{\text{SD of age}}$$

$$|I_{(su)}| \leq |-2_{(su)}| \Rightarrow |\text{pred } y_{(su)}| \leq |x_{(su)}|$$

$$\text{pred } y_{(su)} = r \cdot x_{(su)}$$

From the Fall 2022 Final:

Problem 6.4

Yet again, BruinCard, the company that gave us access to `apps`, `small_apps`, and `mini_apps`, has revoked our access to those three DataFrames and instead has given us `micro_apps`, an even smaller sample of `apps`.

Using `micro_apps`, we are again interested in finding the regression line that will allow us to predict the income of an applicant given their age. We are given the following information:

- The correlation coefficient, r, between ages and incomes is $-\frac{1}{3}$ (note the negative sign).
- The mean income is $\frac{7}{2}$ (remember, incomes are measured in tens of thousands of dollars).
- The mean age is 33.
- The regression line predicts that a 24 year old applicant has an income of $\frac{31}{2}$.

Suppose the standard deviation of incomes in `micro_apps` is an integer multiple of the standard deviation of ages in `micro_apps`. That is,

$$\text{standard deviation of income} = k \cdot \text{standard deviation of age}.$$

What is the value of k? Give your answer as an **integer**.

Click to view the solution.

From the Fall 2022 Final:

Problem 7

Below, we define a new `DataFrame` called `seven_apps` and display it fully.

```
seven_apps = apps.sample(7).sort_values(by="dependents", ascending=False)
seven_apps
```

	status	age	income	homeowner	dependents
505	approved	52.16667	2.6600	yes	3
474	approved	39.00000	3.5000	yes	2
934	approved	22.25000	2.8000	yes	1
828	approved	21.41667	1.5896	no	0
970	approved	21.83333	2.0272	no	0
18	denied	35.58333	4.0000	no	0
784	approved	32.83333	2.5000	no	0

Consider the process of **resampling 7 rows from `seven_apps` with replacement**, and computing the maximum number of dependents in the resample.

dependents: 3, 2, 1, 0, 0, 0, 0

From the Fall 2022 Final:

$P(\text{no 3 in one resample}) = \left(\frac{6}{7}\right)^7$

Problem 7.1

If we take one resample, what is the probability that the maximum number of dependents in the resample is **less than** 3? Leave your answer **unsimplified**.

Click to view the solution.

$P(\text{no 3 in any of the 50 resamples}) = P(\text{no 3 in one resample})^{50} = \left[\left(\frac{6}{7}\right)^7\right]^{50}$

Problem 7.2

If we take 50 resamples, what is the probability that the maximum number of dependents is **never** 3, in any resample? Leave your answer **unsimplified**.

Click to view the solution.

Problem 7.3

If we take 50 resamples, what is the probability that the maximum number of dependents is 3 in **every** resample? Leave your answer **unsimplified**.

Click to view the solution.

$P(\text{at least one 3 in all 50 resamples}) = P(\text{at least one 3 in a single resample})^{50}$

$= \left[1 - \left(\frac{6}{7}\right)^7\right]^{50}$