Review

1 Arrays and Dataframes

Problem 1

Suppose `x` and `y` are both `int`s that have been previously defined, with $x < y$. Now, define:

$$\texttt{peach} = \texttt{np.arange}(x, y, 2)$$

Say that the *spread* of `peach` is the difference between the largest and smallest values in `peach`. The spread should be a non-negative integer.

Problem 1.1 Using array methods, write an expression that evaluates to the spread of `peach`.

```
peach.max() - peach.min()
np.max(peach) - np.min(peach)
```

Problem 1.2 Without using any methods or functions, write an expression that evaluates to the spread of `peach`.

Hint: Use `[]`.

```
peach[-1] - peach[0]
peach[len(peach) - 1] - peach[0]
```

Problem 1.3 Choose the correct way to fill in the blank in this sentence:

The spread of `peach` is _____ the value of $y - x$.

Problem 2

Suppose `students` is a `DataFrame` of all students who took DSC 10 last quarter. `students` has one row per student, where:

- The index contains students' PIDs as strings starting with "A".
- The `"Overall"` column contains students' overall percentage grades as floats.
- The `"Animal"` column contains students' favorite animals as strings.

Problem 2.1 What type is `students.get("Overall")`?

Problem 2.2 What type is `students.get("PID")`?

2 Regression

Problem 3
Suppose the price of an IKEA product and the cost to have it assembled are linearly associated with a correlation of 0.8. Product prices have a mean of 140 dollars and a standard deviation of 40 dollars. Assembly costs have a mean of 80 dollars and a standard deviation of 10 dollars. We want to predict the assembly cost of a product based on its price using linear regression.

Problem 3.1 The NORDMELA 4-drawer dresser sells for 200 dollars. How much do we predict its assembly cost to be?

$$m = r \cdot \frac{SDofy}{SDofx} \tag{1}$$
$$= 0.8 \cdot \frac{10}{40} \tag{2}$$
$$= 0.2 \tag{3}$$

$$b = mean_y - m \cdot mean_x \tag{4}$$
$$= 80 - 0.2 \cdot 140 \tag{5}$$
$$= 52 \tag{6}$$

$$y = 0.2x + 52$$
$$0.2 \cdot 200 + 52 = 92$$

Problem 3.2 The IDANÄS wardrobe sells for 80 dollars more than the KLIPPAN loveseat, so we expect the IDANÄS wardrobe will have a greater assembly cost than the KLIPPAN loveseat. How much do we predict the difference in assembly costs to be?

1. Predict what the wardrobe is selling for: $0.2 \cdot 280 + 52 - (0.2 \cdot 200 + 52) = 0.2 \cdot 80 = 16$

Problem 4
Problem 4.1 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.
predicted $y_{\text{su}} = r \cdot x_{\text{su}}$
$\frac{4}{5} = r \cdot \frac{8}{3}$

Problem 4.2 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.

3 Hypothesis Testing, Permutation Testing, and CLT

Problem 5 Q6, Final Exam Winter 2025

Problem 6 Suppose we want to sample the average amount of money spent by UCSD graduate students on blind boxes. Because graduate students are not paid very much, we know that the maximum possible standard deviation for amount of money spent is \$50. What is the minimum sample size of graduate students I need to take in order to guarantee that my 95% confidence interval for the average amount of money spent has width at most 0.20?

Problem 7 The `DataFrame` `apps` contains application data for a random sample of 1,000 applicants for a particular credit card from the 1990s. The `"age"` column contains the applicants' ages, in years, to the nearest twelfth of a year.

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 random sample of 100 rows of `apps` called `hundred_apps`.

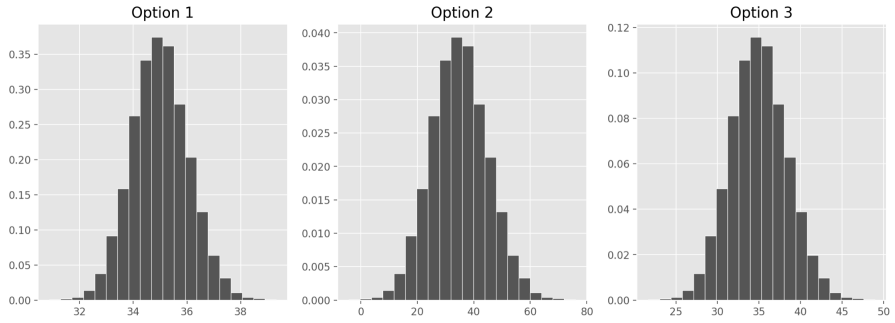
We are interested in estimating the mean age of all applicants in `apps` given only the data in `hundred_apps`. The ages in `hundred_apps` have a mean of 35 and a standard deviation of 10.

Problem 7.1 Give the endpoints of the CLT-based 95% confidence interval for the mean age of all applicants in `apps`, based on the data in `hundred_apps`.

Problem 7.2 BruinCard reinstates our access to `apps` so that we can now easily extract information about the ages of all applicants. We determine that, just like in `hundred_apps`, the ages in `apps` have a mean of 35 and a standard deviation of 10. This raises the question of how other samples of 100 rows of `apps` would have turned out, so we compute 10,000 sample means as follows.

```
sample_means = np.array([])
for i in np.arange(10000):
    sample_mean = apps.sample(100, replace=True).get("age").mean()
    sample_means = np.append(sample_means, sample_mean)
```

Which of the following three visualizations best depict the distribution of `sample_means`?



4 Probability

Problem 8 Q3, Final Exam Winter 2025

Problem 9 Each individual penguin in our dataset is of a certain species (Adelie, Chinstrap, or Gentoo) and comes from a particular island in Antarctica (Biscoe, Dream, or Torgerson). There are 330 penguins in our dataset, grouped by species and island as shown below.

species	island	count
Adelie	Biscoe	44
	Dream	55
	Torgersen	44
Chinstrap	Dream	68
Gentoo	Biscoe	119

Suppose we pick one of these 330 penguins, uniformly at random, and name it Chester.

Problem 9.1 What is the probability that Chester comes from Dream island? Give your answer as a number between 0 and 1, rounded to three decimal places.

Problem 9.2 If we know that Chester comes from Dream island, what is the probability that Chester is an Adelie penguin? Give your answer as a number between 0 and 1, rounded to three decimal places.

Problem 9.3 Suppose I pick 3 more penguins and name them Aaron, Logan, and Cole. What is the probability that none of them are Dream penguins ?