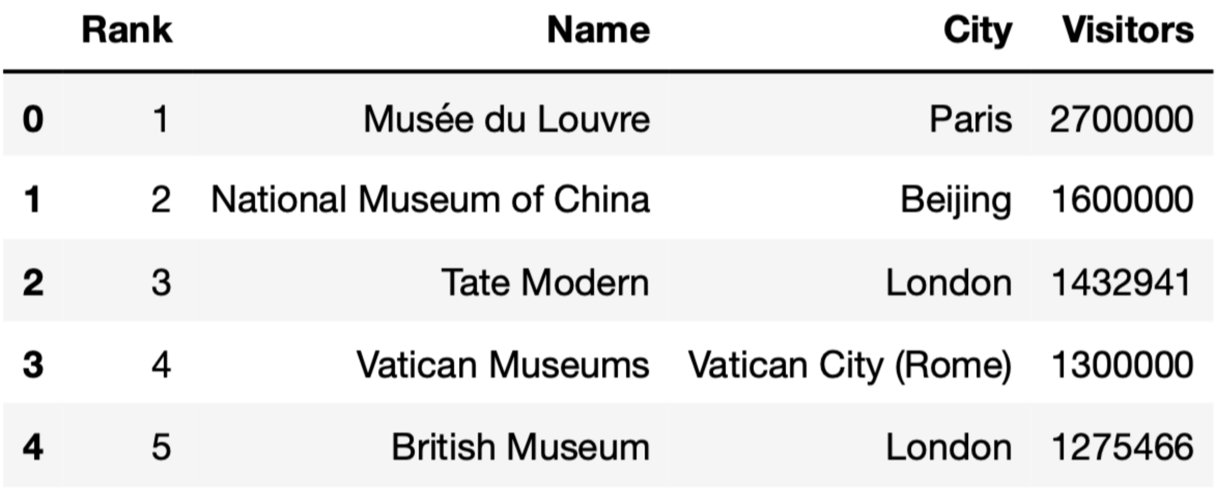
Final Exam SOLUTIONS – DSC 10, Fall 2021

Q1 A Work of Art 🎨

8 Points

Welcome to the Final Exam! In honor of the brand-new Comic-Con Museum that just opened at Balboa Park here in San Diego, this exam will contain questions about various museums and zoos around the world.

In this question, we'll work with the DataFrame `art_museums`, which contains the name, city, number of visitors in 2019, and rank (based on number of visitors) for the 100 most visited art museums in 2019. The first few rows of `art_museums` are shown below.



	Rank	Name	City	Visitors
0	1	Musée du Louvre	Paris	2700000
1	2	National Museum of China	Beijing	1600000
2	3	Tate Modern	London	1432941
3	4	Vatican Museums	Vatican City (Rome)	1300000
4	5	British Museum	London	1275466

Q1.1

5 Points

Which of the following blocks of code correctly assigns `random_art_museums` to an array of the names of 10 art museums, randomly selected without replacement from those in `art_museums`? Select all that apply.

Option 1:

```
def get_10(df):
    return np.array(df.sample(10).get('Name'))

random_art_museums = get_10(art_museums)
```

Option 2:

```
def get_10(art_museums):
    return np.array(art_museums.sample(10).get('Name'))

random_art_museums = get_10(art_museums)
```

Option 3:

```
def get_10(art_museums):
    random_art_museums = np.array(art_museums.sample(10).get('Name'))

random_art_museums = get_10(art_museums)
```

Option 4:

```
def get_10():
    return np.array(art_museums.sample(10).get('Name'))

random_art_museums = get_10()
```

Option 5:

```
random_art_museums = np.array([])

def get_10():
    random_art_museums = np.array(art_museums.sample(10).get('Name'))
    return random_art_museums

get_10()
```

☑ Option 1

☑ Option 2

□ Option 3

☑ Option 4

□ Option 5

□ None of the above

Q1.2

3 Points

For your convenience, we show the first few rows of `art_museums` again below.

	Rank	Name	City	Visitors
0	1	Musée du Louvre	Paris	2700000
1	2	National Museum of China	Beijing	1600000
2	3	Tate Modern	London	1432941
3	4	Vatican Museums	Vatican City (Rome)	1300000
4	5	British Museum	London	1275466

London has the most art museums in the top 100 of any city in the world. The most visited art museum in London is `'Tate Modern'`.

Which of the following blocks of code correctly assigns `best_in_london` to `'Tate Modern'`? Select all that apply.

Option 1:

```
def most_visited(df, col, value):
    return df[df.get(col)==value].sort_values(by='Visitors', ascending=False)

def most_common(df, col):
    return df.groupby(col).count().sort_values(by='Rank', ascending=False).in

best_in_london = most_visited(art_museums, 'City', most_common(art_museums, '
```

Option 2:

```
def most_visited(df, col, value):
    print(df[df.get(col)==value].sort_values(by='Visitors', ascending=False).

def most_common(df, col):
    print(df.groupby(col).count().sort_values(by='Rank', ascending=False).ind

best_in_london = most_visited(art_museums, 'City', most_common(art_museums, '
```

Option 3:

```python
def most_common(df, col):
    return df.groupby(col).count().sort_values(by='Rank', ascending=False).in

def most_visited(df, col, value):
    print(df[df.get(col)==value].sort_values(by='Visitors', ascending=False).

best_in_london = most_visited(art_museums, 'City', most_common(art_museums, '
```

☑ Option 1

□ Option 2

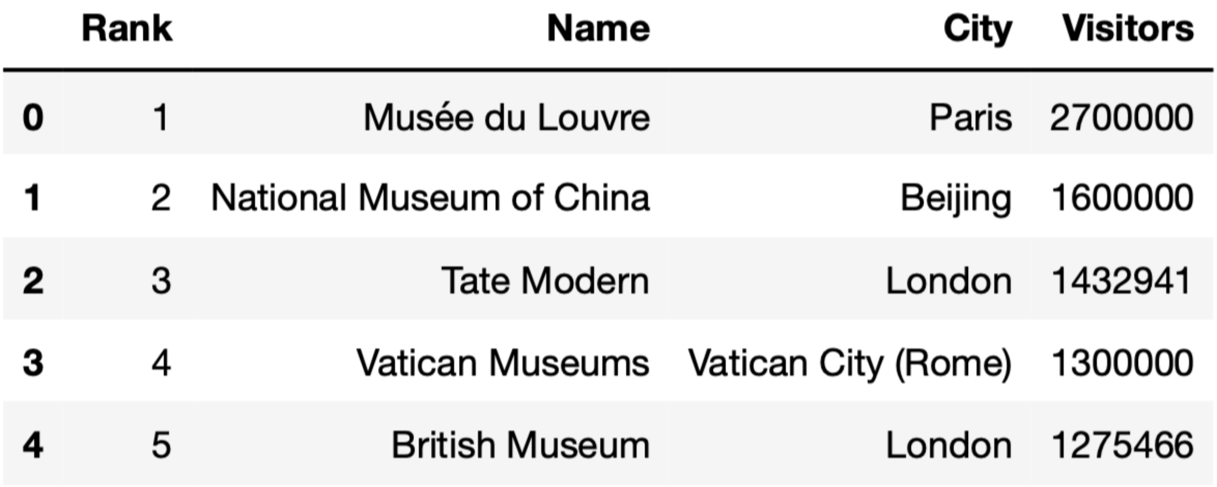
□ Option 3

□ None of the above

Q2 Around the World 🌎

10 Points

In this question, we'll keep working with the `art_museums` DataFrame.



	Rank	Name	City	Visitors
0	1	Musée du Louvre	Paris	2700000
1	2	National Museum of China	Beijing	1600000
2	3	Tate Modern	London	1432941
3	4	Vatican Museums	Vatican City (Rome)	1300000
4	5	British Museum	London	1275466

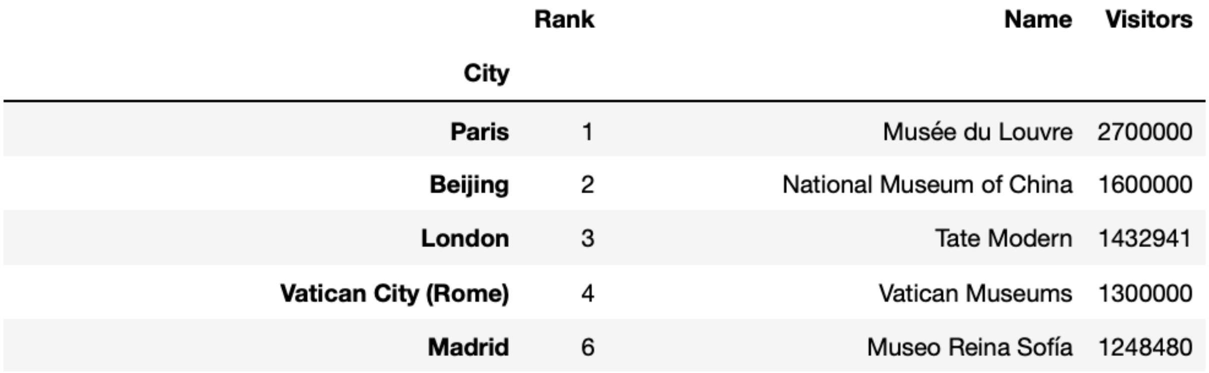
Q2.1

6 Points

`'Tate Modern'` is the most popular art museum in London. But what's the most popular art museum in each city?

It turns out that there's no way to answer this easily using the tools that you know about so far. To help, we've created a new Series method, `.last()`. If `s` is a Series, `s.last()` returns the last element of `s` (i.e. the element at the very end of `s`). `.last()` works with `.groupby`, too (just like `.mean()` and `.count()`).

Fill in the blanks so that the code below correctly assigns `best_per_city` to a DataFrame with one row per city, that describes the name, number of visitors, and rank of the most visited art museum in each city. `best_per_city` should be sorted in decreasing order of number of visitors. The first few rows of `best_per_city` are shown below.



	Rank	Name	Visitors
City			
Paris	1	Musée du Louvre	2700000
Beijing	2	National Museum of China	1600000
London	3	Tate Modern	1432941
Vatican City (Rome)	4	Vatican Museums	1300000
Madrid	6	Museo Reina Sofía	1248480

```
best_per_city = __(a)__.groupby(__(b)__).last().__(c)__
```

What goes in blank (a)?

art_museums.sort_values('Visitors', ascending=True)

What goes in blank (b)?

'City'

What goes in blank (c)?

sort_values('Visitors', ascending=False)

Q2.2

4 Points

Assume you've defined `best_per_city` correctly.

Which of the following options evaluates to the number of visitors to the most visited art museum in Amsterdam? Select all that apply.

☑ `best_per_city.get('Visitors').loc['Amsterdam']`

☑ `best_per_city[best_per_city.index == 'Amsterdam'].get('Visitors').iloc[0]`

☑ `best_per_city[best_per_city.index == 'Amsterdam'].get('Visitors').iloc[-1]`

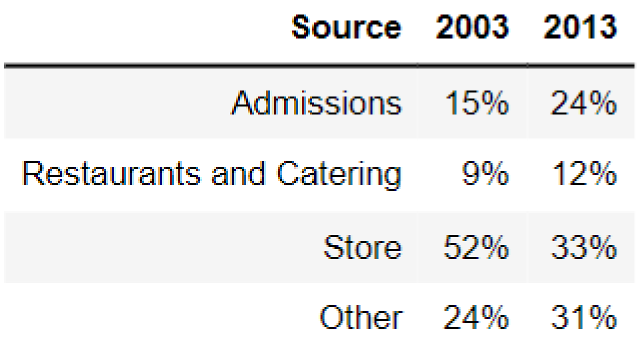
☑ `best_per_city[best_per_city.index == 'Amsterdam'].get('Visitors').loc['Ams`

□ None of the above

Q3 Money Talks 💸

6 Points

The table below shows the average amount of revenue from different sources for art museums in 2003 and 2013.



Source	2003	2013
Admissions	15%	24%
Restaurants and Catering	9%	12%
Store	52%	33%
Other	24%	31%

Q3.1

2 Points

What is the total variation distance between the distribution of revenue sources in 2003 and the distribution of revenue sources in 2013? Give your answer as a proportion (i.e. a decimal between 0 and 1), **not** a percentage. Round your answer to three decimal places.

=0.19+-0

Q3.2

2 Points

Which type of visualization would be best suited for comparing the two distributions in the table?

○ Scatter plot

○ Line plot

○ Overlaid histogram

● Overlaid bar chart

Q3.3

2 Points

Notably, there was an economic recession in 2008-2009. Which of the following can we conclude was an effect of the recession?

○ The increase in revenue from admissions, as more people were visiting museums.

○ The decline in revenue from museum stores, as people had less money to spend.

○ The decline in total revenue, as fewer people were visiting museums

● None of the above

Q4 Dinosaur Bones 🦕🦴

5 Points

Q4.1

3 Points

The Museum of Natural History has a large collection of dinosaur bones, and they know the approximate year each bone is from. They want to use this sample of dinosaur bones to estimate **the total number of years that dinosaurs lived on Earth**. We'll make the assumption that the sample is a uniform random sample from the population of all dinosaur bones. Which statistic below will give the best estimate of the population parameter?

○ sample sum

◉ $\text{sample max} - \text{sample min}$

○ $2 \cdot (\text{sample mean} - \text{sample min})$

○ $2 \cdot (\text{sample max} - \text{sample mean})$

○ $2 \cdot \text{sample mean}$

○ $2 \cdot \text{sample median}$

Q4.2

2 Points

The curator at the Museum of Natural History, who happens to have taken a data science course in college, points out that the estimate of the parameter obtained from this sample could certainly have come out differently, if the museum had started with a different sample of bones. The curator suggests trying to understand the distribution of the sample statistic. Which of the following would be an appropriate way to create that distribution?

○ bootstrapping the original sample

○ using the Central Limit Theorem

○ both bootstrapping and the Central Limit Theorem

◉ neither bootstrapping nor the Central Limit Theorem

Q5 Number of Visitors 🧑‍🤝‍🧑

5 Points

Q5.1

2 Points

Now, the Museum of Natural History wants to know how many visitors they have in a year. However, their computer systems are rather archaic and so they aren't able to keep track of the number of tickets sold for an entire year. Instead, they randomly select five days in the year, and keep track of the number of visitors on those days. Let's call these numbers v_1, v_2, v_3, v_4, and v_5.

Which of the following is the best estimate the number of visitors for the entire year?

○ $v_1 + v_2 + v_3 + v_4 + v_5$

○ $\frac{5}{365} \cdot (v_1 + v_2 + v_3 + v_4 + v_5)$

◉ $\frac{365}{5} \cdot (v_1 + v_2 + v_3 + v_4 + v_5)$

○ $365 \cdot v_3$

Q5.2

3 Points

Now we're interested in predicting the admission cost of a museum based on its number of visitors. Suppose:

- admission cost and number of visitors are linearly associated with a correlation coefficient of 0.25,
- the number of visitors at the Museum of Natural History is six standard deviations below average,
- the average cost of museum admission is 15 dollars, and
- the standard deviation of admission cost is 3 dollars.

What would the regression line predict for the admission cost (in dollars) at the Museum of Natural History? Give your answer as a number without any units, rounded to three decimal places.

=10.5+-0

Q6 Club Penguin 🐧

23 Points

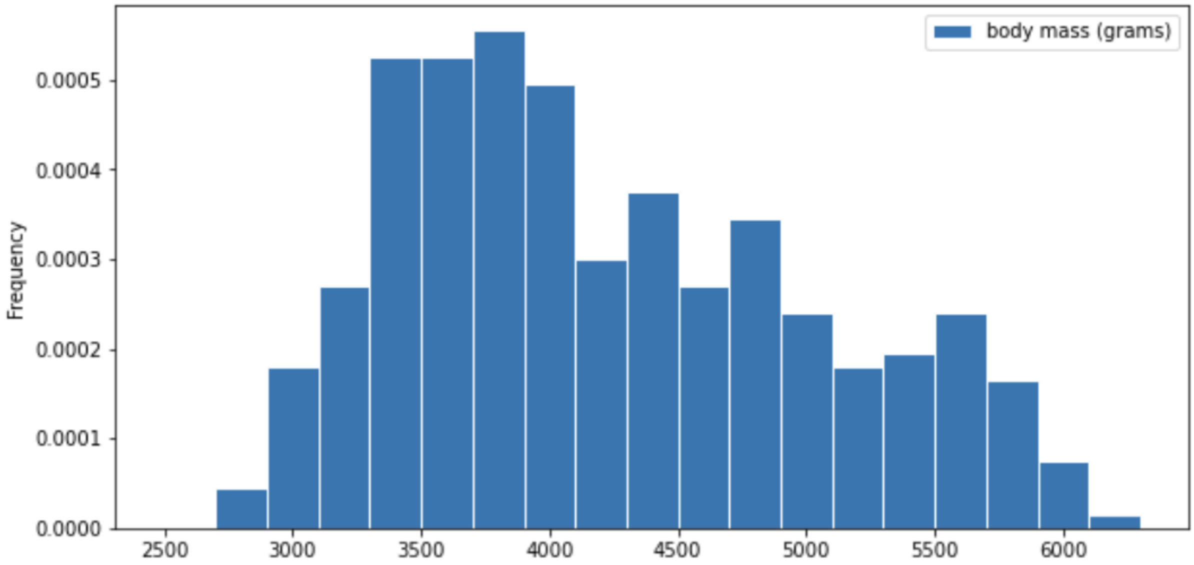
Researchers from the San Diego Zoo, located within Balboa Park, collected physical measurements of several species of penguins in a region of Antarctica.

One piece of information they tracked for each of 330 penguins was its mass in grams. The average penguin mass is 4200 grams, and the standard deviation is 840 grams.

Q6.1

2 Points

Consider the histogram of mass below.



Select the true statement below.

○ The median mass of penguins is larger than the average mass of penguins

○ The median mass of penguins is roughly equal to the average mass of penguins (within 50 grams)

● The median mass of penguins is less than the average mass of penguins

○ It is impossible to determine the relationship between the median and average mass of penguins just by looking at the above histogram

Q6.2

2 Points

Which of the following is a valid conclusion that we can draw solely from the histogram above?

○ The number of penguins with a mass of exactly 3500 grams is greater than the number of penguins with a mass of exactly 5500 grams.

● The number of penguins with a mass of at most 3500 grams is greater than the number of penguins with a mass of at least 5500 grams.

○ There is an odd number of penguins in the dataset.

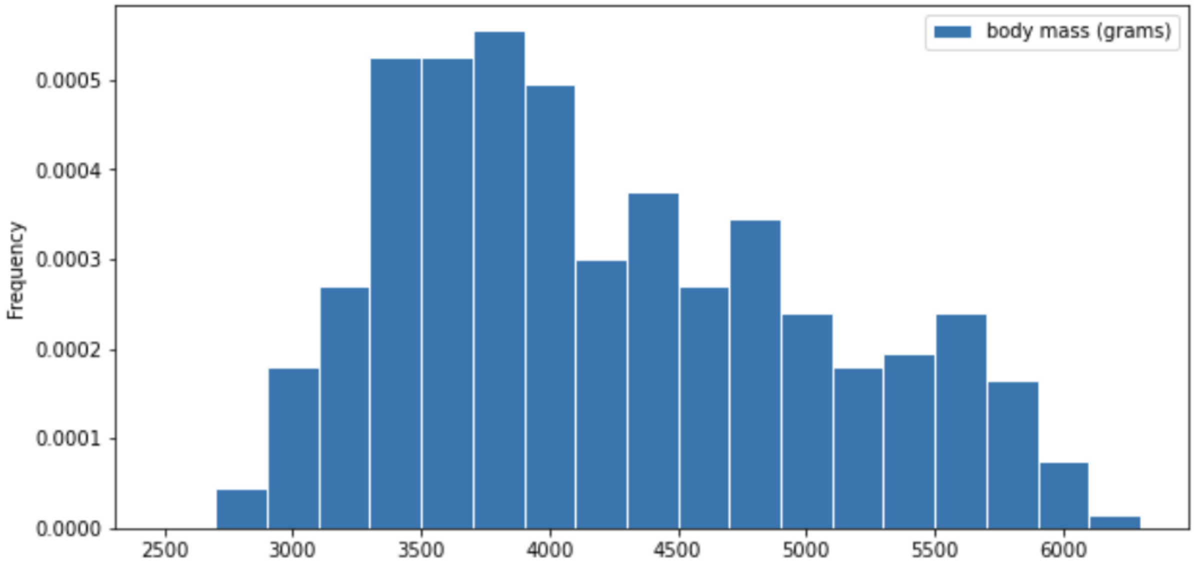
○ The number penguins with a mass of exactly 4000 grams is greater than zero.

○ None of the above.

Q6.3

3 Points

For your convenience, we show the histogram of mass again below.



Recall, there are 330 penguins in our dataset. Their average mass is 4200 grams, and the standard deviation of mass is 840 grams.

Per Chebyshev's inequality, **at least** what percentage of penguins have a mass between 3276 grams and 5124 grams? Input your answer as a **percentage** between 0 and 100, without the % symbol. Round to three decimal places.

=17.355+-0.1

Q6.4

2 Points

Per Chebyshev's inequality, **at least** what percentage of penguins have a mass between 1680 grams and 5880 grams?

○ 50%

○ 55.5%

○ 65.25%

○ 68%

● 75%

○ 88.8%

○ 95%

Q6.5

2 Points

The distribution of mass in grams is not roughly normal. Is the distribution of mass in standard units roughly normal?

○ Yes

● No

○ Impossible to tell

Q6.6

4 Points

Suppose all 330 penguin body masses (in grams) that the researchers collected are stored in an array called `masses`. We'd like to estimate the probability that two different randomly selected penguins from our dataset have body masses within 50 grams of one another (including a difference of exactly 50 grams). Fill in the missing pieces of the simulation below so that the function `estimate_prob_within_50g` returns an estimate for this probability.

```
def estimate_prob_within_50g():
    num_reps = 10000
    within_50g_count = 0
    for i in np.arange(num_reps):
        two_penguins = np.random.choice(__(a)__)
        if __(b)__:
            within_50g_count = within_50g_count + 1
    return within_50g_count / num_reps
```

What goes in blank (a)?

masses, 2, replace=False

What goes in blank (b)?

abs(two_penguins[0] - two_penguins[1])
<=50

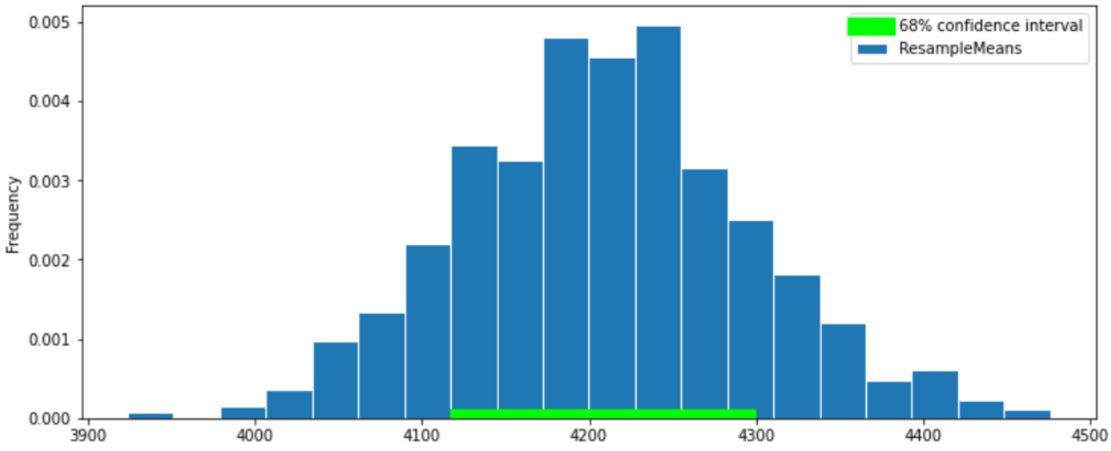
Q6.7

2 Points

Recall, there are 330 penguins in our dataset. Their average mass is 4200 grams, and the standard deviation of mass is 840 grams. Assume that the 330 penguins in our dataset are a random sample from the population of all penguins in Antarctica. Our sample gives us one estimate of the population mean.

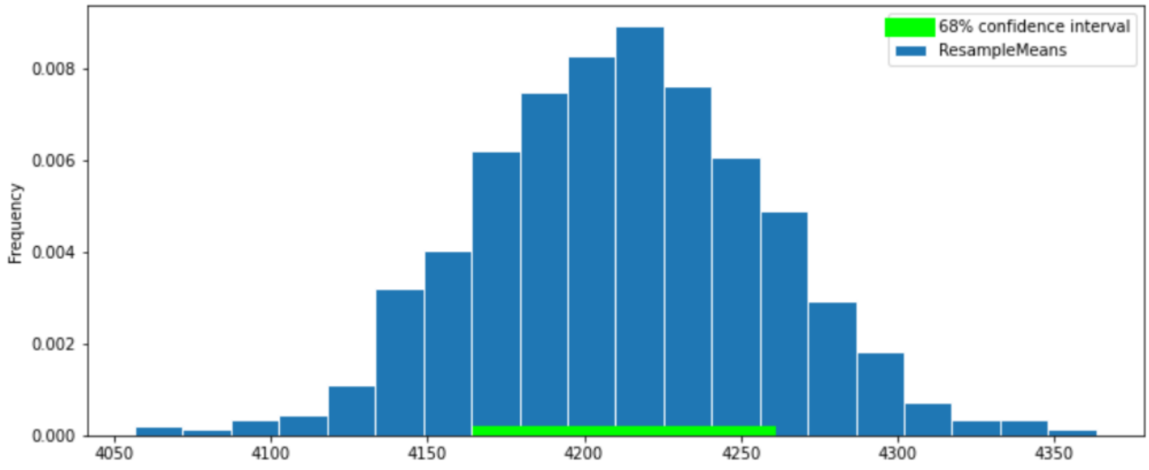
To better estimate the population mean, we bootstrapped our sample and plotted a histogram of the resample means, then took the middle 68 percent of those values to get a confidence interval. Which option below shows the histogram of the resample means and the confidence interval we found?

○



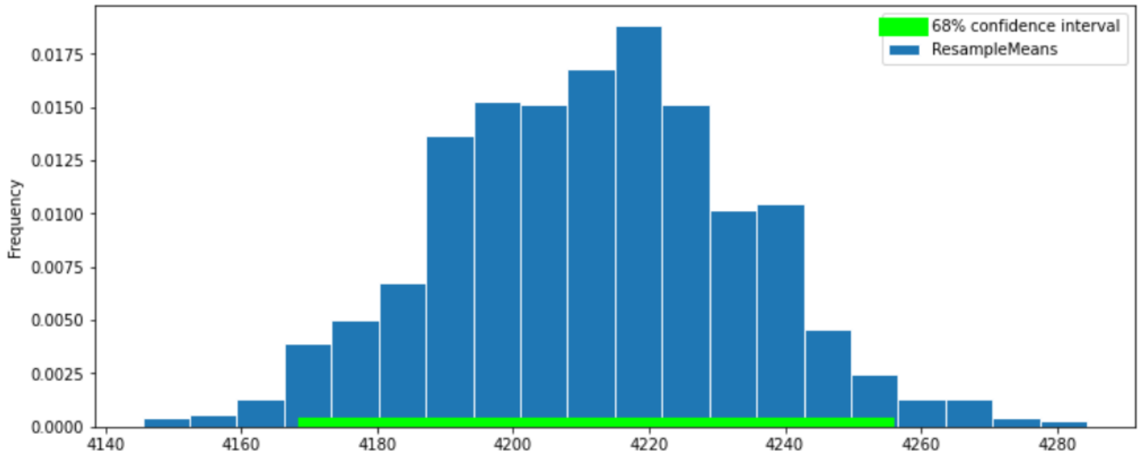
68% confidence interval
ResampleMeans
Frequency
0.005
0.004
0.003
0.002
0.001
0.000
3900
4000
4100
4200
4300
4400
4500

●



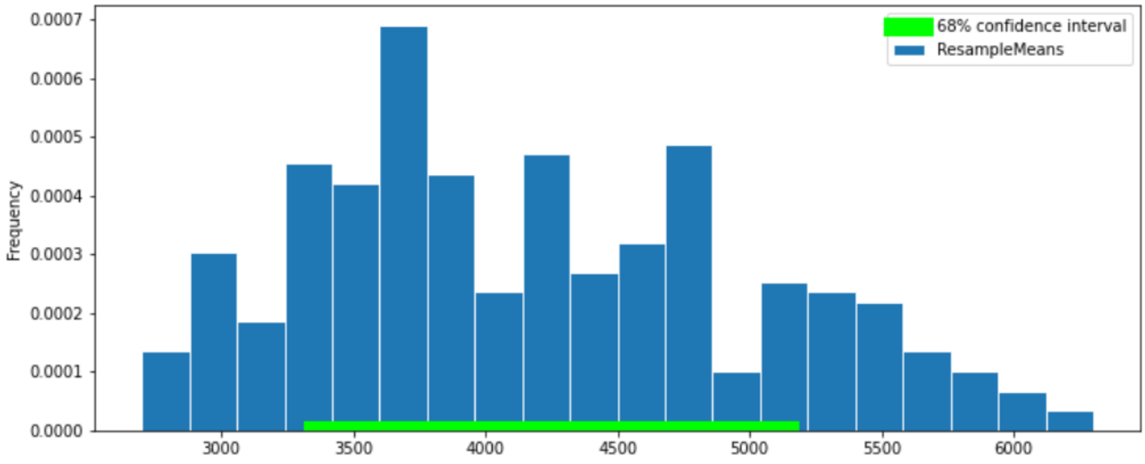
68% confidence interval
ResampleMeans
Frequency
0.008
0.006
0.004
0.002
0.000
4050
4100
4150
4200
4250
4300
4350

○



68% confidence interval
ResampleMeans
Frequency
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
4140
4160
4180
4200
4220
4240
4260
4280

○



68% confidence interval
ResampleMeans
Frequency
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
3000
3500
4000
4500
5000
5500
6000

Q6.8

2 Points

Suppose `boot_means` is an array of the resampled means. Fill in the blanks below so that `[left, right]` is a 68% confidence interval for the true mean mass of penguins.

```
left = np.percentile(boot_means, __(a)__)
right = np.percentile(boot_means, __(b)__)
[left, right]
```

What goes in blank (a)?

16

What goes in blank (b)?

84

Q6.9

4 Points

Which of the following is a correct interpretation of this confidence interval? Select all that apply.

□ There is an approximately 68% chance that mean weight of all penguins in Antarctica falls within the bounds of this confidence interval.

□ Approximately 68% of penguin weights in our sample fall within the bounds of this confidence interval.

□ Approximately 68% of penguin weights in the population fall within the bounds of this interval.

☑ If we created many confidence intervals using the same method, approximately 68% of them would contain the mean weight of all penguins in Antarctica.

□ None of the above

Q7 Bills 💵

12 Points

Now let's study the relationship between a penguin's bill length (in millimeters) and mass (in grams). Suppose we're given that

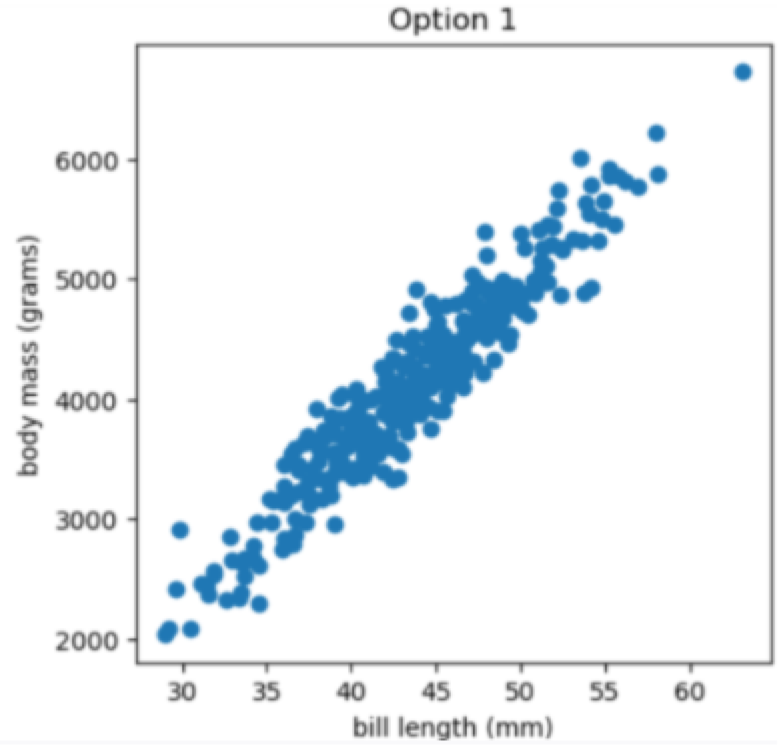
- bill length and body mass have a correlation coefficient of 0.55
- the average bill length is 44 mm and the standard deviation of bill lengths is 6 mm
- as before, the average body mass is 4200 grams and the standard deviation of body mass is 840 grams

Q7.1

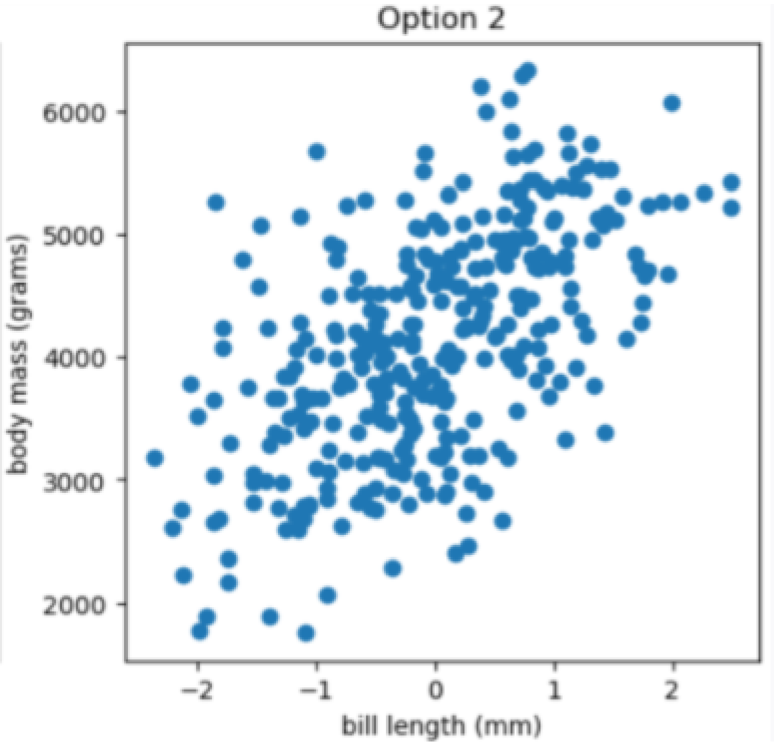
2 Points

Which of the four scatter plots below describe the relationship between bill length and body mass, based on the information provided in the question?

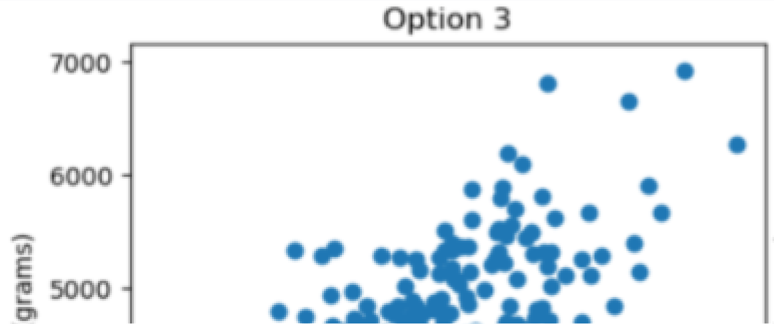
○

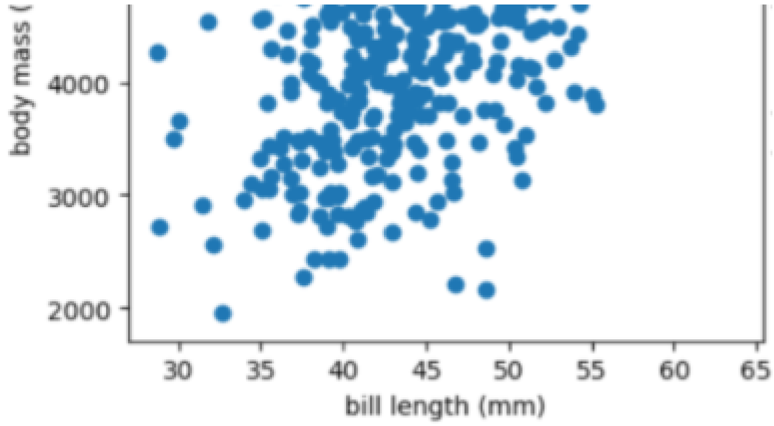

Option 1

○

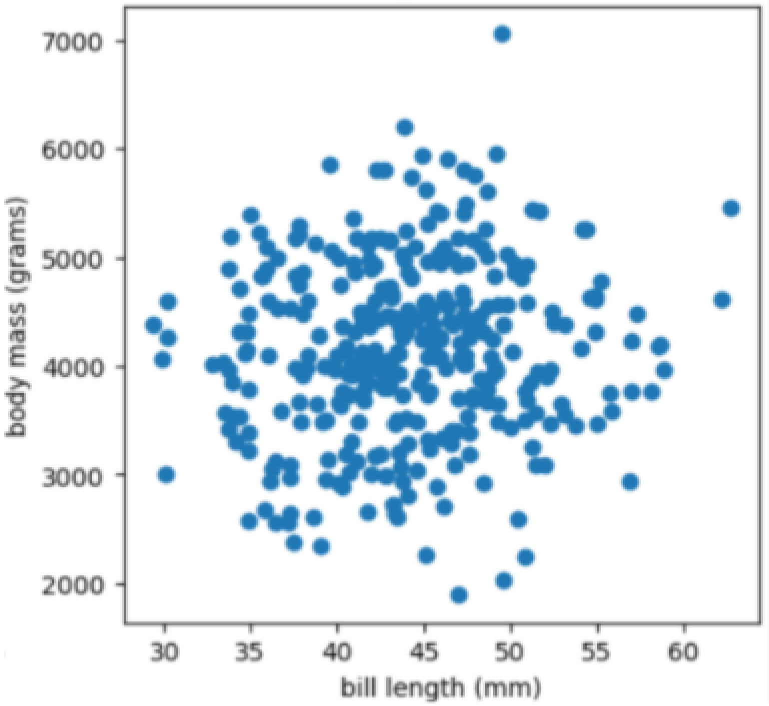

Option 2

◉


Option 3



○ Option 4



Q7.2

4 Points

Suppose we want to find the regression line that uses bill length, x, to predict body mass, y. The line is of the form $y = mx + b$. What are m and b?

What is m? Give your answer as a number without any units, rounded to three decimal places.

=77+-0

What is b? Give your answer as a number without units, rounded to three decimal places.

=812+-0

Q7.3

2 Points

What is the predicted body mass (in grams) of a penguin whose bill length is 44 mm? Give your answer as a number without any units, rounded to three decimal places.

=4200+-0

Q7.4

2 Points

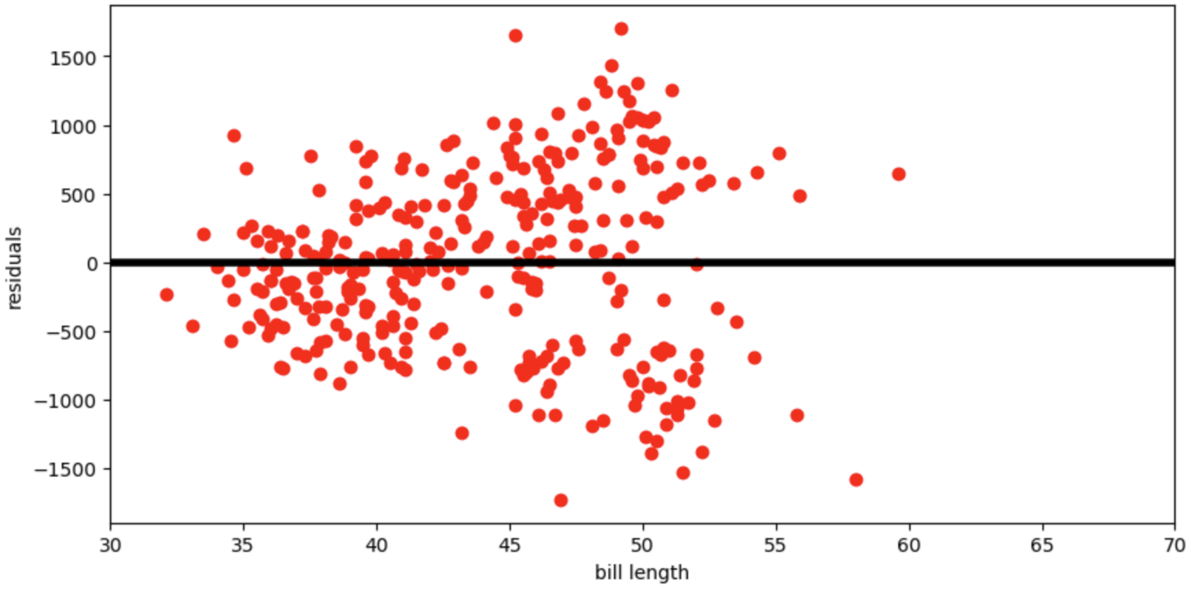
A particular penguin had a predicted body mass of 6800 grams. What is that penguin's bill length (in mm)? Give your answer as a number without any units, rounded to three decimal places.

=77.766+-0.1

Q7.5

2 Points

Below is the residual plot for our regression line.



Which of the following is a valid conclusion that we can draw solely from the residual plot above?

○ For this dataset, there is another line with a lower root mean squared error

○ The root mean squared error of the regression line is 0

● The accuracy of the regression line's predictions depends on bill length

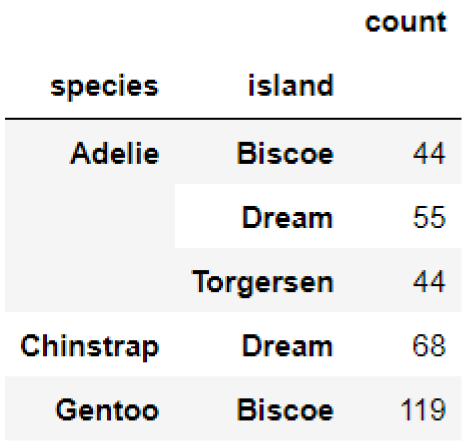
○ The relationship between bill length and body mass is likely non-linear

○ None of the above

Q8 Dream Island 💭🏝️

8 Points

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.



		count
species	**island**	
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.

Q8.1

2 Points

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.

=0.373+-0.001

Q8.2

3 Points

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.

=0.447+-0.001

Q8.3

3 Points

If we know that Chester is not from Dream island, what is the probability that Chester is not an Adelie penguin? Give your answer as a number between 0 and 1, rounded to three decimal places.

=0.575+-0.001

Q9 Measuring Up 📏

13 Points

We're now interested in investigating the differences between the masses of Adelie penguins and Chinstrap penguins. Specifically, our null hypothesis is that their masses are drawn from the same population distribution, and any observed differences are due to chance only.

Below, we have a snippet of working code for this hypothesis test, for a specific test statistic. Assume that `adelie_chinstrap` is a DataFrame of only Adelie and Chinstrap penguins, with just two columns – `'species'` and `'mass'`.

```
stats = np.array([])
num_reps = 500
for i in np.arange(num_reps):
    # --- line (a) starts ---
    shuffled = np.random.permutation(adelie_chinstrap.get('species'))
    # --- line (a) ends ---

    # --- line (b) starts ---
    with_shuffled = adelie_chinstrap.assign(species=shuffled)
    # --- line (b) ends ---

    grouped = with_shuffled.groupby('species').mean()

    # --- line (c) starts ---
    stat = grouped.get('mass').iloc[0] - grouped.get('mass').iloc[1]
    # --- line (c) ends ---

    stats = np.append(stats, stat)
```

Q9.1

2 Points

Which of the following statements best describe the procedure above?

○ This is a standard hypothesis test, and our test statistic is the total variation distance between the distribution of Adelie masses and Chinstrap masses

○ This is a standard hypothesis test, and our test statistic is the difference between the expected proportion of Adelie penguins and the proportion of Adelie penguins in our resample

○ This is a permutation test, and our test statistic is the total variation distance between the distribution of Adelie masses and Chinstrap masses

● This is a permutation test, and our test statistic is the difference in the mean Adelie mass and mean Chinstrap mass

Q9.2

2 Points

Currently, `line (c)` (marked with a comment) uses `.iloc`. Which of the following options compute the exact same statistic as `line (c)` currently does?

Option 1:

```
stat = grouped.get('mass').loc['Adelie'] - grouped.get('mass').loc['Chinstrap
```

Option 2:

```
stat = grouped.get('mass').loc['Chinstrap'] - grouped.get('mass').loc['Adelie
```

● Option 1 only

○ Option 2 only

○ Both options

○ Neither option

Q9.3

1 Point

Is it possible to re-write `line (c)` in a way that uses `.iloc[0]` twice, without any other uses of `.loc` or `.iloc`?

● Yes, it's possible

○ No, it's not possible

Q9.4

2 Points

For your convenience, we copy the code for the hypothesis test below.

```
stats = np.array([])
num_reps = 500
for i in np.arange(num_reps):
    # --- line (a) starts ---
    shuffled = np.random.permutation(adelie_chinstrap.get('species'))
    # --- line (a) ends ---

    # --- line (b) starts ---
    with_shuffled = adelie_chinstrap.assign(species=shuffled)
    # --- line (b) ends ---

    grouped = with_shuffled.groupby('species').mean()

    # --- line (c) starts ---
    stat = grouped.get('mass').iloc[0] - grouped.get('mass').iloc[1]
    # --- line (c) ends ---

    stats = np.append(stats, stat)
```

What would happen if we removed `line (a)`, and replaced `line (b)` with

```
with_shuffled = adelie_chinstrap.sample(adelie_chinstrap.shape[0], replace=Fa
```

Select the best answer.

○ This would still run a valid hypothesis test

● This would not run a valid hypothesis test, as all values in the `stats` array would be exactly the same

○ This would not run a valid hypothesis test, even though there would be several different values in the `stats` array

○ This would not run a valid hypothesis test, as it would incorporate information about Gentoo penguins

Q9.5

2 Points

For your convenience, we copy the code for the hypothesis test below.

```
stats = np.array([])
num_reps = 500
for i in np.arange(num_reps):
    # --- line (a) starts ---
    shuffled = np.random.permutation(adelie_chinstrap.get('species'))
    # --- line (a) ends ---

    # --- line (b) starts ---
    with_shuffled = adelie_chinstrap.assign(species=shuffled)
    # --- line (b) ends ---

    grouped = with_shuffled.groupby('species').mean()

    # --- line (c) starts ---
    stat = grouped.get('mass').iloc[0] - grouped.get('mass').iloc[1]
    # --- line (c) ends ---

    stats = np.append(stats, stat)
```

What would happen if we removed `line (a)`, and replaced `line (b)` with

```
with_shuffled = adelie_chinstrap.sample(adelie_chinstrap.shape[0], replace=Tr
```

Select the best answer.

○ This would still run a valid hypothesis test

○ This would not run a valid hypothesis test, as all values in the `stats` array would be exactly the same

● This would not run a valid hypothesis test, even though there would be several different values in the `stats` array

○ This would not run a valid hypothesis test, as it would incorporate information about Gentoo penguins

Q9.6

2 Points

For your convenience, we copy the code for the hypothesis test below.

```
stats = np.array([])
num_reps = 500
for i in np.arange(num_reps):
    # --- line (a) starts ---
    shuffled = np.random.permutation(adelie_chinstrap.get('species'))
    # --- line (a) ends ---

    # --- line (b) starts ---
    with_shuffled = adelie_chinstrap.assign(species=shuffled)
    # --- line (b) ends ---

    grouped = with_shuffled.groupby('species').mean()

    # --- line (c) starts ---
    stat = grouped.get('mass').iloc[0] - grouped.get('mass').iloc[1]
    # --- line (c) ends ---

    stats = np.append(stats, stat)
```

What would happen if we replaced `line (a)` with

```
with_shuffled = adelie_chinstrap.assign(
    species=np.random.permutation(adelie_chinstrap.get('species')
)
```

and replaced `line (b)` with

```
with_shuffled = with_shuffled.assign(
    mass=np.random.permutation(adelie_chinstrap.get('mass')
)
```

Select the best answer.

● This would still run a valid hypothesis test

○ This would not run a valid hypothesis test, as all values in the `stats` array would be exactly the same

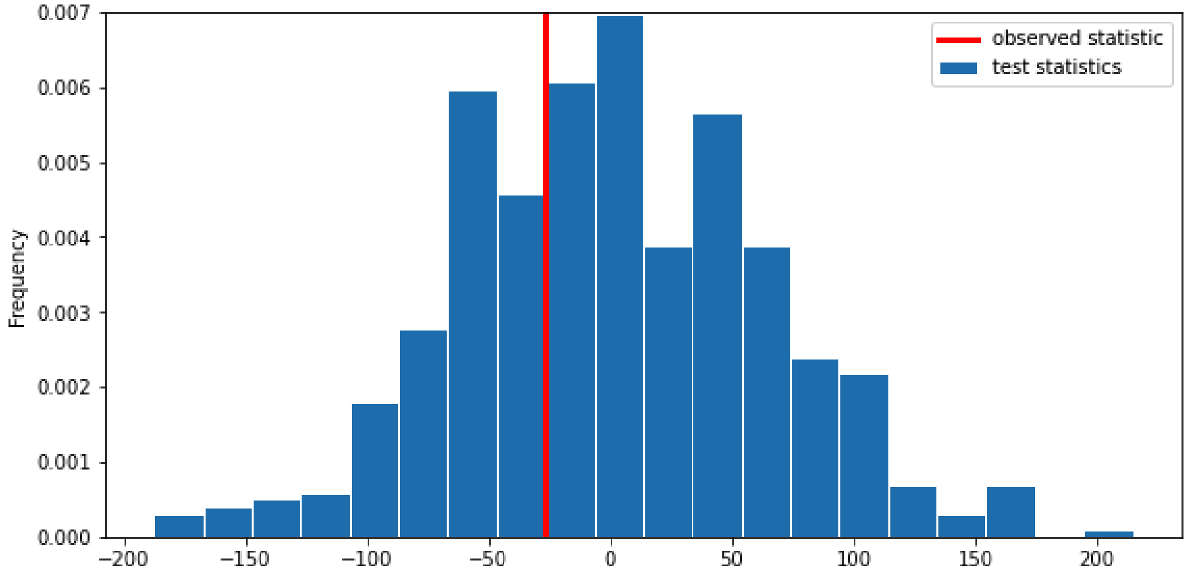
○ This would not run a valid hypothesis test, even though there would be several different values in the `stats` array

○ This would not run a valid hypothesis test, as it would incorporate information about Gentoo penguins

Q9.7

2 Points

Suppose we run the code for the hypothesis test and see the following empirical distribution for the test statistic. In red is the observed statistic.



Suppose our alternative hypothesis is that Chinstrap penguins weigh more on average than Adelie penguins. Which of the following is closest to the p-value for our hypothesis test?

○ 0

○ $\frac{1}{4}$

● $\frac{1}{3}$

○ $\frac{2}{3}$

○ $\frac{3}{4}$

○ 1

Q10 Model Railroad Museum 🚂

20 Points

At the San Diego Model Railroad Museum, there are different admission prices for children, adults, and seniors. Over a period of time, as tickets are sold, employees keep track of how many of each type of ticket are sold. These ticket counts (in the order child, adult, senior) are stored as follows.

```
admissions_data = np.array([550, 1550, 400])
```

Q10.1

2 Points

Complete the code below so that it creates an array `admissions_proportions` with the proportions of tickets sold to each group (in the order child, adult, senior).

```
def as_proportion(data):
    return __(a)__

admissions_proportions = as_proportion(admissions_data)
```

What goes in blank (a)?

data/data.sum()

Q10.2

2 Points

The museum employees have a model in mind for the proportions in which they sell tickets to children, adults, and seniors. This model is stored as follows.

```
model = np.array([0.25, 0.6, 0.15])
```

We want to conduct a hypothesis test to determine whether the admissions data we have is consistent with this model. Which of the following is the null hypothesis for this test?

○ Child, adult, and senior tickets might plausibly be purchased in proportions 0.25, 0.6, and 0.15.

● Child, adult, and senior tickets are purchased in proportions 0.25, 0.6, and 0.15.

○ Child, adult, and senior tickets might plausibly be purchased in proportions other than 0.25, 0.6, and 0.15.

○ Child, adult, and senior tickets, are purchased in proportions other than 0.25, 0.6, and 0.15.

Q10.3

4 Points

Which of the following test statistics could we use to test our hypotheses? Select all that could work.

□ sum of differences in proportions

☑ sum of squared differences in proportions

□ mean of differences in proportions

☑ mean of squared differences in proportions

□ none of the above

Q10.4

2 Points

Below, we'll perform the hypothesis test with a different test statistic, the mean of the absolute differences in proportions.

Recall that the ticket counts we observed for children, adults, and seniors are stored in the array `admissions_data  = np.array([550, 1550, 400])`, and that our model is `model = np.array([0.25, 0.6, 0.15])`.

For our hypothesis test to determine whether the admissions data is consistent with our model, what is the observed value of the test statistic? Input your answer as a decimal between 0 and 1. Round to three decimal places.

=0.02+-0

Moving forward, suppose that the value you calculated above is assigned to the variable `observed_stat`.

Q10.5

6 Points

Now, we want to simulate the test statistic 10,000 times under the assumptions of the null hypothesis. Fill in the blanks below to complete this simulation and calculate the p-value for our hypothesis test. Assume that the variables `admissions_data`, `admissions_proportions`, `model`, and `observed_stat` are already defined as specified earlier in the question.

```
simulated_stats = np.array([])
for i in np.arange(10000):
    simulated_proportions = as_proportions(np.random.multinomial(__(a)__, __(
    simulated_stat = __(c)__
    simulated_stats = np.append(simulated_stats, simulated_stat)

p_value = __(d)__
```

What goes in blank (a)?

admissions_data.sum()

What goes in blank (b)?

model

What goes in blank (c)?

np.abs(simulated_proportions - model).mean()

What goes in blank (d)?

np.count_nonzero(simulated_stats >= observed_stat) / 10000

Q10.6

2 Points

True or False: the p-value represents the probability that the null hypothesis is true.

○ True

● False

Q10.7

2 Points

The new statistic that we used for this hypothesis test, the mean of the absolute differences in proportions, is in fact closely related to the total variation distance. Given two arrays of length three, `array_1` and `array_2`, suppose we compute the mean of the absolute differences in proportions between `array_1` and `array_2` and store the result as `madp`. What value would we have to multiply `madp` by to obtain the total variation distance `array_1` and `array_2`? Input your answer below, rounding to three decimal places.

=1.5+-0

Q11 Extra Credit

0 Points

Hopefully, after taking this course, you have a sense of how broadly applicable data science can be. Tell us about a cause or an issue that's important to you, and how you might use data science to support the cause or address the issue you care about. Be specific and reference particular data science tools you have learned in this class.