Final Exam – DSC 10, Winter 2021

Q1 Sequence

3 Points

One way to use `np.arange` to produce the sequence `[2, 6, 10, 14]` is `np.arange(2, 15, 4)`. This gives three inputs to `np.arange`.

Fill in the blanks below to show a different way to produce the same sequence, this time using only one input to `np.arange`. Each blank below must be filled in with **a single number only**, and the final result, `x*np.arange(y)+z`, must produce the sequence `[2, 6, 10, 14]`.

`x =`

`y =`

`z =`

`x*np.arange(y)+z`

Q2 Restaurants

3 Points

The command `.set_index` can take as input one column, to be used as the index, or a sequence of columns to be used as a nested index (sometimes called a MultiIndex). A MultiIndex is the default behavior of the table returned by `.groupby` with multiple columns.

You are given a table called `restaurants` that contains information on a variety of local restaurants' daily number of customers and daily income. There is a row for each restaurant for each date in a given five-year time period.

The columns of `restaurants` are `name` (string), `year` (int), `month` (int), `day` (int), `num_diners` (int), and `income` (float).

Assume that in our data set, there are not two different restaurants that go by the same `name` (chain restaurants, for example).

Which of the following would be the best way to set the index for this dataset?

○ `restaurants.set_index('name')`

○ `restaurants.set_index(['year', 'month', 'day'])`

○ `restaurants.set_index(['name', 'year', 'month', 'day'])`

Q3 Merge

3 Points

If we merge a table with n rows with a table with m rows, how many rows does the resulting table have?

○ n

○ m

○ max(m, n)

○ $m * n$

○ not enough information to tell

Q4 Sampling

4 Points

You sample from a population by assigning each element of the population a number starting with 1. You include element 1 in your sample. Then you generate a random number, n, between 2 and 5, inclusive, and you take every nth element after element 1 to be in your sample. For example, if you select $n = 2$, then your sample will be elements 1, 3, 5, 7, and so on.

Q4.1 True/False 1

2 Points

True or False: Before the sample is drawn, you can calculate the probability of selecting each subset of the population.

○ True

○ False

Q4.2 True/False 2

2 Points

True or False: Each subset of the population is equally likely to be selected.

○ True

○ False

Q5 Books

9 Points

You are given a table called `books` that contains columns `author` (string), `title` (string), `num_chapters` (int), and `publication_year` (int).

Q5.1 Books 1

3 Points

What will be the output of the following code?

```
books.groupby(“publication_year”).mean().shape[1]
```

○ 1

○ 2

○ 3

○ 4

Q5.2 Books 2

3 Points

Which of the following strategies would work to compute the absolute difference in the average number of chapters per book for authors "Dean Koontz" and "Charles Dickens"?

○ group by `author`, aggregate with `.mean()`, use `get` on `num_chapters` column compute the absolute value of the difference between `iloc["Charles Dickens"]` and `iloc["Dean Koontz"]`

○ do two queries to get two separate tables (one for each of "Dean Koontz" and "Charles Dickens"), use `get` on the `num_chapters` column of each table, use the Series method `.mean()` on each, compute the absolute value of the difference in these two means

○ group by both `author` and `title`, aggregate with `.mean()`, use `get` on `num_chapters` column, use `loc` twice to find values in that column corresponding to "Dean Koontz" and "Charles Dickens", compute the absolute value of the difference in these two values

○ query using a compound condition to get all books corresponding to "Dean Koontz" or "Charles Dickens", group by `author`, aggregate with `.mean()`, compute absolute value of the difference in `index[0]` and `index[1]`

Q5.3 Books 3

3 Points

Which of the following will produce the same value as the total number of books in the table?

○ `books.groupby('Title').count().shape[0]`

○ `books.groupby('Author').count().shape[0]`

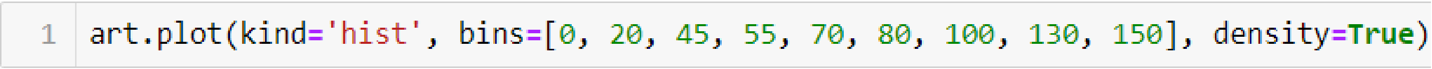
○ `books.groupby(['Author, 'Title']).count().shape[0]`

Q6 Art Galleries

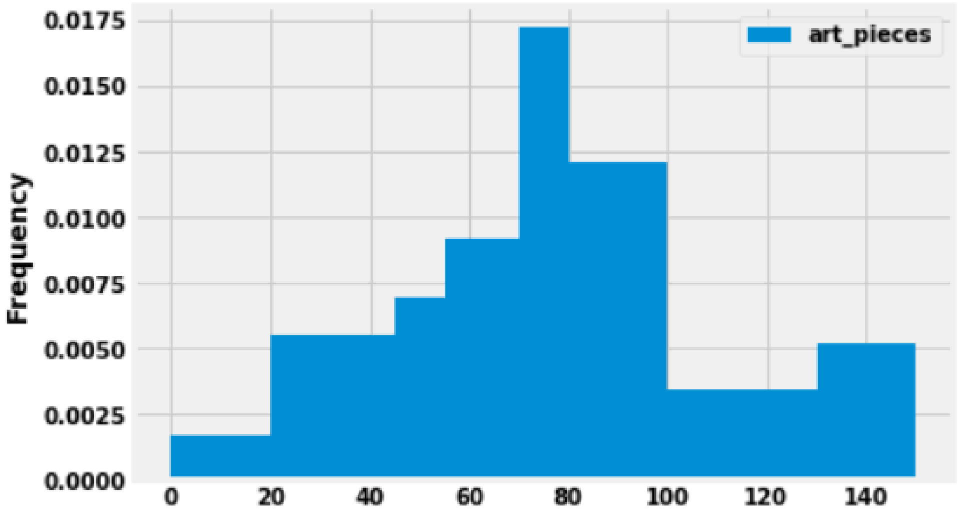
6 Points

Suppose you have a dataset of 29 art galleries that includes the number of pieces of art in each gallery.

A histogram of the number of art pieces in each gallery, as well as the code that generated it, is shown below.



```
art.plot(kind='hist', bins=[0, 20, 45, 55, 70, 80, 100, 130, 150], density=True)
```



Q6.1 Art Galleries 1

3 Points

How many galleries have at least 80 but less than 100 art pieces? Input your answer below. Make sure your answer is an **integer** and does not include any text or symbols.

Q6.2 Art Galleries 2

3 Points

If we added to our dataset two more art galleries, each containing 24 pieces of art, and plotted the histogram again for the larger dataset, what would be the height of the bin $[20, 45)$? Input your answer as a number rounded to **six decimal places**.

Q7 Sample

3 Points

Assume `df` is a DataFrame with distinct rows. Which of the following best describes `df.sample(10)`?

○ an array of length 10, where some of the entries might be the same

○ an array of length 10, where no two entries can be the same

○ a DataFrame with 10 rows, where some of the rows might be the same

○ a DataFrame with 10 rows, where no two rows can be the same

Q8 Dice

2 Points

True or False: If you roll two dice, the probability of rolling two fives is the same as the probability of rolling a six and a three.

○ True

○ False

Q9 Experiment

6 Points

Suppose you do an experiment in which you do some random process 500 times and calculate the value of some statistic, which is a count of how many times a certain phenomenon occurred out of the 500 trials. You repeat the experiment 10,000 times and draw a histogram of the 10,000 statistics.

Q9.1 Experiment 1

2 Points

Is this histogram a probability histogram or an empirical histogram?

○ probability histogram

○ empirical histogram

Q9.2 Experiment 2

2 Points

If you instead repeat the experiment 100,000 times, how will the histogram change?

○ it will become wider

○ it will become narrower

○ it will barely change at all

Q9.3 Experiment 3

2 Points

For each experiment, if you instead do the random process 5,000 times, how will the histogram change?

○ it will become wider

○ it will become narrower

○ it will barely change at all

Q10 Open-Ended: Permutation Testing

3 Points

Give an example of a dataset and a question you would want to answer about that dataset which you could answer by performing a permutation test (also known as an A/B test).

Creative responses that are different than ones we've already seen in this class will earn the most credit.

Q11 Sample

3 Points

Suppose you draw a sample of size 100 from a population with mean 50 and standard deviation 15. What is the probability that your sample has a mean between 50 and 53? Input the probability below, as a number between 0 and 1, rounded to **two decimal places**.

Q12 Vaccine

3 Points

You need to estimate the proportion of American adults who want to be vaccinated against Covid-19. You plan to survey a random sample of American adults, and use the proportion of adults in your sample who want to be vaccinated as your estimate for the true proportion in the population. Your estimate must be within 0.04 of the true proportion, 95% of the time. Using the fact that the standard deviation of any dataset of 0's and 1's is no more than 0.5, calculate the minimum number of people you would need to survey. Input your answer below, as an **integer**.

Q13 Exam Scores

3 Points

- Hector earned a 77 on an exam where the mean was 70 and the standard deviation was 5.
- Clara earned an 80 on an exam where the mean was 75 and the standard deviation was 3.
- Vivek earned an 83 on an exam where the mean was a 75 and the standard deviation was 6.

Rank these three students in **ascending** order of their exam performance *relative to their classmates.*

○ Hector, Clara, Vivek

○ Vivek, Hector, Clara

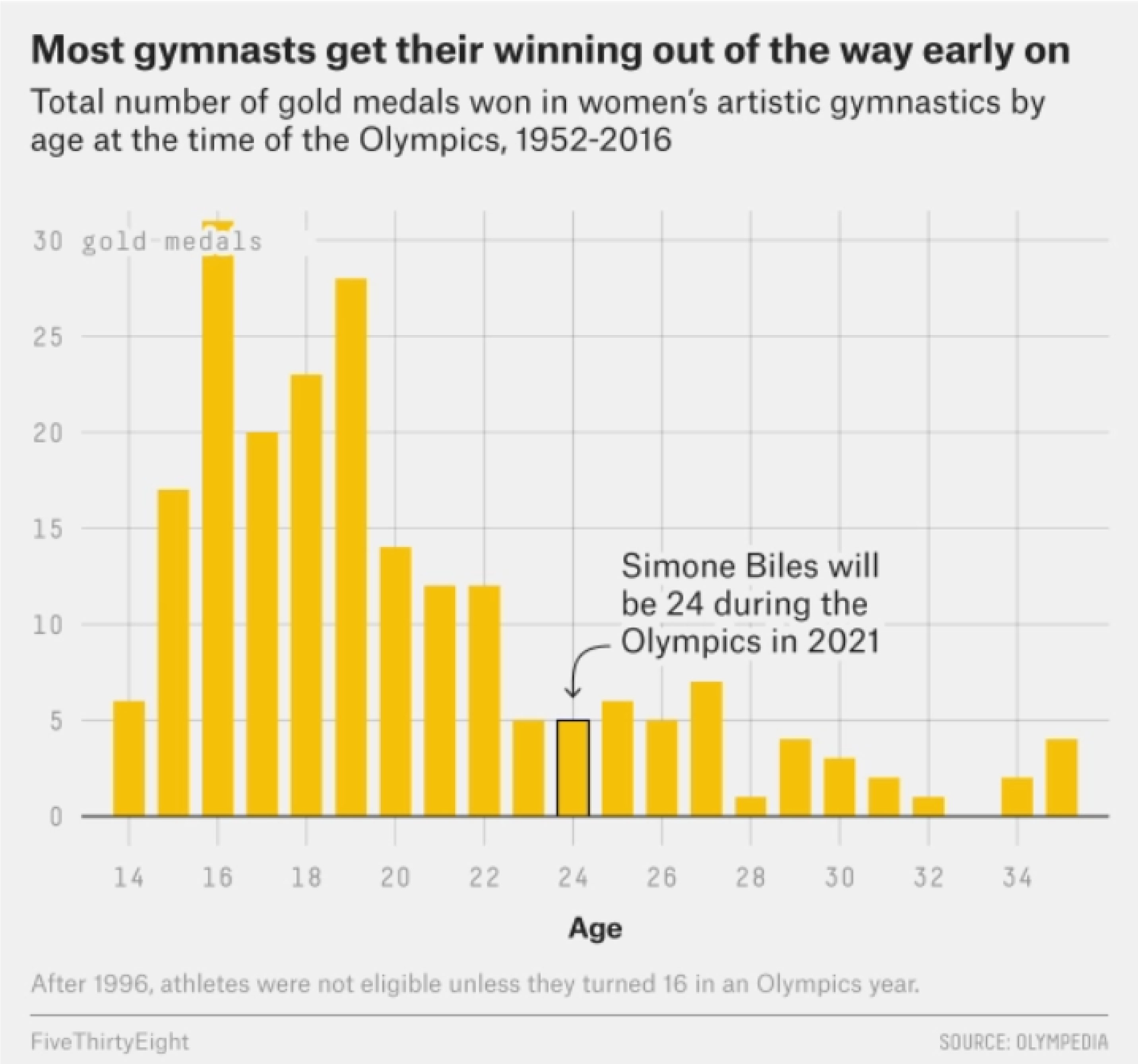
○ Clara, Hector, Vivek

○ Vivek, Clara, Hector

Q14 Gymnastics

6 Points

The data visualization below shows all Olympic gold medals for women’s gymnastics, broken down by the age of the gymnast.



Most gymnasts get their winning out of the way early on

Total number of gold medals won in women’s artistic gymnastics by age at the time of the Olympics, 1952-2016

Simone Biles will be 24 during the Olympics in 2021

After 1996, athletes were not eligible unless they turned 16 in an Olympics year.

FiveThirtyEight

SOURCE: OLYMPEDIA

Q14.1 Gymnastics 1

3 Points

Based on this data, rank the following three quantities in **ascending** order: the median age at which gold medals are earned, the mean age at which gold medals are earned, the standard deviation of the age at which gold medals are earned.

○ mean, median, SD

○ median, mean, SD

○ SD, mean, median

○ SD, median, mean

Q14.2 Gymnastics 2

3 Points

Which is larger for this dataset?

○ the difference between the 50th percentile of ages and the 25th percentile of ages

○ the difference between the 75th percentile of ages and the 50th percentile of ages

○ both are the same

Q15 Board Game

9 Points

In a board game, whenever it is your turn, you roll a six-sided die and move that number of spaces. You get 10 turns, and you win the game if you've moved 50 spaces in those 10 turns. Suppose you create a simulation, based on 10,000 trials, to show the distribution of the number of spaces moved in 10 turns. Let's call this distribution $Dist_{10}$. You also wonder how the game would be different if you were allowed 15 turns instead of 10, so you create another simulation, based on 10,000 trials, to show the distribution of the number of spaces moved in 15 turns, which we'll call $Dist_{15}$.

Q15.1 Board Game 1

3 Points

What can we say about the shapes of $Dist_{10}$ and $Dist_{15}$?

○ both will be roughly normally distributed

○ only one will be roughly normally distributed

○ neither will be roughly normally distributed

Q15.2 Board Game 2

3 Points

What can we say about the centers of $Dist_{10}$ and $Dist_{15}$?

○ both will have approximately the same mean

○ the mean of $Dist_{10}$ will be smaller than the mean of $Dist_{15}$

○ the mean of $Dist_{15}$ will be smaller than the mean of $Dist_{10}$

Q15.3 Board Game 3

3 Points

What can we say about the spread of $Dist_{10}$ and $Dist_{15}$?

○ both will have approximately the same standard deviation

○ the standard deviation of $Dist_{10}$ will be smaller than the standard deviation of $Dist_{15}$

○ the standard deviation of $Dist_{15}$ will be smaller than the standard deviation of $Dist_{10}$

Q16 True/False

2 Points

True/False: The slope of the regression line, when both variables are measured in standard units, is never more than 1.

○ True

○ False

Q17 True/False

2 Points

True/False: The slope of the regression line, when both variables are measured in original units, is never more than 1.

○ True

○ False

Q18 True/False

2 Points

True/False: Suppose that from a sample, you compute a 95% bootstrapped confidence interval for a population parameter to be the interval $[L, R]$. Then the average of L and R is the mean of the original sample.

○ True

○ False

Q19 True/False

2 Points

True/False: Suppose that from a sample, you compute a 95% normal confidence interval for a population parameter to be the interval $[L, R]$. Then the average of L and R is the mean of the original sample.

○ True

○ False

Q20 Open-Ended: Pizza

5 Points

You order 25 large pizzas from your local pizzeria. The pizzeria claims that these pizzas are 16 inches in diameter, but you're not so sure. You measure each pizza's diameter and collect a dataset of 25 actual pizza diameters. You want to run a hypothesis test to determine whether the pizzeria's claim is accurate.

Q20.1 State Hypotheses

2 Points

What would your Null Hypothesis be?

What would your Alternative Hypothesis be?

Q20.2 Test Statistic

1 Point

What test statistic would you use?

Q20.3 Explanation

2 Points

Explain how you would do your hypothesis test and how you would draw a conclusion from your results.

Q21 Restaurant

3 Points

A restaurant keeps track of each table's number of people (average 3; standard deviation 1) and the amount of the bill (average $60, standard deviation $12). If the number of people and amount of the bill are linearly associated with correlation 0.8, what is the predicted bill for a table of 5 people? Input your answer below, **to the nearest cent**. Make sure your answer is just a number and does not include the $ symbol or any text.

Q22 Bootstrap

6 Points

From a population with mean 500 and standard deviation 50, you collect a sample of size 100. The sample has mean 400 and standard deviation 40. You bootstrap this sample 10,000 times, collecting 10,000 resample means.

Q22.1 Bootstrap 1

3 Points

Which of the following is the most accurate description of the mean of the distribution of the 10,000 bootstrapped means?

○ The mean will be exactly equal to 400.

○ The mean will be exactly equal to 500.

○ The mean will be approximately equal to 400.

○ The mean will be approximately equal to 500.

Q22.2 Bootstrap 2

3 Points

Which of the following is closest to the standard deviation of the distribution of the 10,000 bootstrapped means?

○ 400

○ 40

○ 4

○ 0.4

Q23 Percentile

4 Points

Recall the mathematical definition of percentile and how we calculate it.

> Let p be a number between 0 and 100. The pth percentile of a collection is the smallest value in the collection that is *at least as large* as p% of all the values.

By this definition, any percentile between 0 and 100 can be computed for any collection of values and is always an element of the collection. Suppose there are n elements in the collection. To find the pth percentile:

1. Sort the collection in increasing order.
2. Find p% of n: $\frac{p}{100} * n$. Call that h. If h is an integer, define $k = h$. Otherwise, let k be the smallest integer greater than h.
3. Take the kth element of the sorted collection.

You have a dataset of 7 values, which are [3, 6, 7, 9, 10, 15, 18]. Using the mathematical definition of percentile above, find the smallest and largest integer values of p so that the pth percentile of this dataset corresponds to the value 10. Input your answers below, as **integers between 0 and 100**.

Q23.1 Smallest Percentile

2 Points

Smallest =

Q23.2 Largest Percentile

2 Points

Largest =

Q24 Books

3 Points

Are nonfiction books longer than fiction books?

Choose the best data science tool to help you answer this question.

○ hypothesis testing

○ permutation (A/B) testing

○ Central Limit Theorem

○ regression

Q25 Friends

3 Points

Do people have more friends as they get older?

Choose the best data science tool to help you answer this question.

○ hypothesis testing

○ permutation (A/B) testing

○ Central Limit Theorem

○ regression

Q26 Ice Cream

3 Points

Does an ice cream shop sell more chocolate or vanilla ice cream cones?

Choose the best data science tool to help you answer this question.

○ hypothesis testing

○ permutation (A/B) testing

○ Central Limit Theorem

○ regression

Q27 Extra Credit! Open-Ended: Important to You

3 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.