Final Exam - DSC 10, Summer 2022

Instructions:

- You have 180 minutes to complete this exam.
- This exam consists of 11 questions, worth a total of 120 points.
- Write your name in the top right of each page in the space provided.
- Please write neatly in the provided answer boxes. We will not grade work that appears elsewhere.
- Completely fill in bubbles and square boxes.
  - ◯ A bubble means that you should only **select one choice**.
  - ☐ A square box means you should **select all that apply**.
- You may refer to the DSC 10 reference sheet and two hand-written page of notes. No other resources or technology (including calculators) are permitted.

Full Name:

PID:

Name of student to your **left**:

Name of student to your **right**:

(Write "N/A" if a wall/aisle is to your left/right.)

By signing below, you are agreeing that you will behave honestly and fairly during and after this exam.

Signature:

Do not open your exam until instructed to do so.

This page is intentionally left blank. You can use it as scratch paper.

Name: _____

Question 1 (4 points)

Choose the best visualization to answer each of the following questions.

a) What kinds of dogs cost the most?

◯ Bar plot ◯ Scatter plot ◯ Histogram ◯ None of the above

b) Do toy dogs have smaller `"size"` values than other kinds of dogs?

◯ Bar plot ◯ Scatter plot ◯ Histogram ◯ None of the above

c) Is there a linear relationship between weight and height?

◯ Bar plot ◯ Scatter plot ◯ Histogram ◯ None of the above

d) Do dogs that are highly suitable for children weigh more?

◯ Bar plot ◯ Scatter plot ◯ Histogram ◯ None of the above

Question 2 (4 points)

The following code computes the breed of the cheapest toy dog.

```
df[__________].__________.__________
      (a)         (b)         (c)
```

a) Fill in blank (a).

b) Fill in blank (b).

c) Which of the following can fill in blank (c)? **Select all that apply.**

☐ `loc[0]` ☐ `iloc[0]` ☐ `index[0]` ☐ `min()`

Question 3 (8 points)

The following code computes an array containing the unique kinds of dogs that are heavier than 20 kg **or** taller than 40 cm on average.

```
foo = df._________._________
           (a)       (b)

np.array(foo[_________]._________)
               (c)         (d)
```

a) Fill in blank (a).

b) Fill in blank (b).

c) Fill in blank (c).

d) Which of the following should fill in blank (d)?

◯ `.index` ◯ `.unique()` ◯ `.get("kind")` ◯ `.get(["kind"])`

e) Let's say that right after blank (b), we added `.reset_index()`. Now, which of the following should fill in blank (d)?

◯ `.index` ◯ `.unique()` ◯ `.get("kind")` ◯ `.get(["kind"])`

Name: _____

Question 4 (7 points)

The `sums` function takes in an array of numbers and outputs the cumulative sum for each item in the array. The cumulative sum for an element is the current element plus the sum of all the previous elements in the array.

For example:

```
>>> sums(np.array([1, 2, 3, 4, 5]))
array([1, 3, 6, 10, 15])

>>> sums(np.array([100, 1, 1]))
array([100, 101, 102])
```

The incomplete definition of `sums` is shown below.

```
def sums(arr):
    res = _________
              (a)
    res = np.append(res, arr[0])
    for i in _________:
                 (b)
        res = np.append(res, _________)
                                   (c)
    return res
```

a) Fill in blank (a).

b) Fill in blank (b).

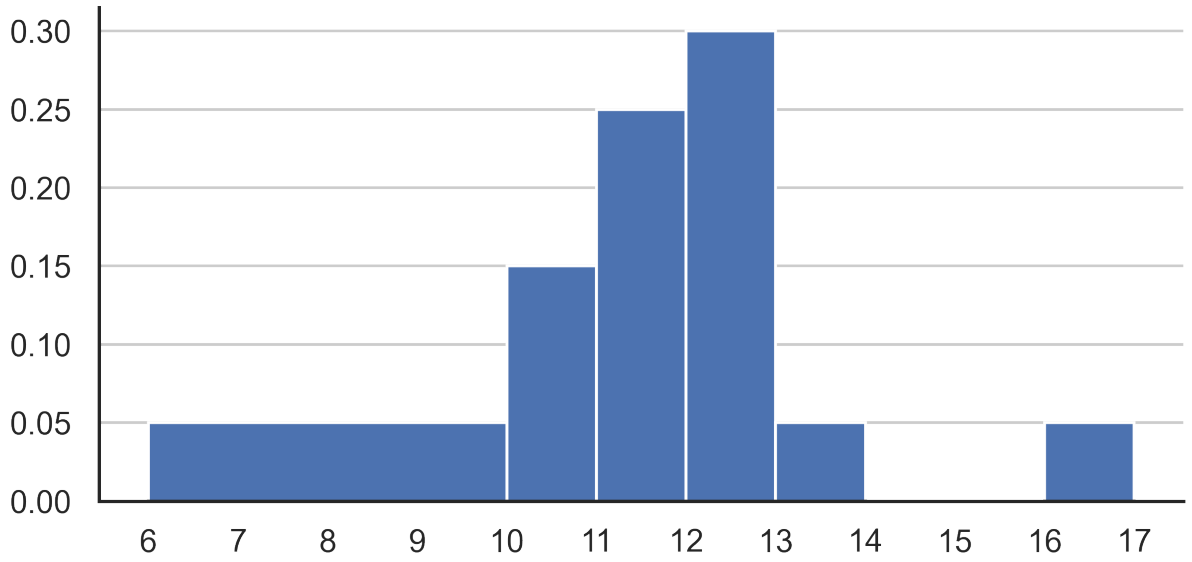
c) Fill in blank (c).

Question 5 (9 points)

Shivani wrote a function called `doggos` defined as follows:

```
def doggos(n, lower, upper):
    t = df.sample(n, replace=True).get("longevity")
    return sum(lower <= t < upper)
```

This plot shows a density histogram of the `"longevity"` column.



Answer each of these questions by either writing a **single number** in the box or selecting "Not enough information", but **not both**. What is the probability that:

a) `doggos(1, 10, 11) == 1` is `True`?

◯ Not enough information

b) `doggos(2, 0, 12) == 2` is `True`?

◯ Not enough information

c) `doggos(2, 13, 20) > 0` is `True`?

◯ Not enough information

Name: _____

Question 6 (21 points)

Every year, the American Kennel Club holds a Photo Contest for dogs. Eric wants to know whether **toy dogs win disproportionately more often than other kinds of dogs**. He has collected a sample of 500 dogs that have won the Photo Contest. In his sample, 200 dogs were toy dogs.

Eric also knows the distribution of dog kinds in the population:

	herding	hound	toy	terrier	sporting	working
proportion	0.1	0.2	0.3	0.2	0.15	0.05

a) Select all correct statements of the null hypothesis.

☐ The distribution of dogs in the sample is the same as the distribution in the population. Any difference is due to chance.

☐ Every dog in the sample was drawn uniformly at random without replacement from the population.

☐ The number of toy dogs that win is the same as the number of toy dogs in the population.

☐ The proportion of toy dogs that win is the same as the proportion of toy dogs in the population.

☐ The proportion of toy dogs that win is 0.3.

☐ The proportion of toy dogs that win is 0.5.

b) Select the correct statement of the alternative hypothesis.

◯ The model in the null hypothesis underestimates how often toy dogs win.

◯ The model in the null hypothesis overestimates how often toy dogs win.

◯ The distribution of dog kinds in the sample is not the same as the population.

◯ The data were not drawn at random from the population.

c) Select all the test statistics that Eric can use to conduct his hypothesis test.

☐ The proportion of toy dogs in his sample.

☐ The number of toy dogs in his sample.

☐ The absolute difference of the sample proportion of toy dogs and 0.3 .

☐ The absolute difference of the sample proportion of toy dogs and 0.5 .

☐ The TVD between his sample and the population.

d) Eric decides on this test statistic: the proportion of toy dogs minus the proportion of non-toy dogs. What is the observed value of the test statistic?

◯ -0.4 ◯ -0.2 ◯ 0 ◯ 0.2 ◯ 0.4

e) Which snippets of code correctly compute Eric's test statistic on one simulated sample under the null hypothesis? Select all that apply. The result must be stored in the variable `stat`.

☐ Snippet 1:

```
a = np.random.choice([0.3, 0.7])
b = np.random.choice([0.3, 0.7])
stat = a - b
```

☐ Snippet 2:

```
a = np.random.choice([0.1, 0.2, 0.3, 0.2, 0.15, 0.05])
stat = a - (1 - a)
```

☐ Snippet 3:

```
a = np.random.multinomial(500, [0.1, 0.2, 0.3, 0.2, 0.15, 0.05]) / 500
stat = a[2] - (1 - a[2])
```

☐ Snippet 4:

```
a = np.random.multinomial(500, [0.3, 0.7]) / 500
stat = a[0] - (1 - a[0])
```

☐ Snippet 5:

```
a = df.sample(500, replace=True)
b = a[a.get("kind") == "toy"].shape[0] / 500
stat = b - (1 - b)
```

f) After simulating, Eric has an array called `sim` that stores his simulated test statistics, and a variable called `obs` that stores his observed test statistic.

What should go in the blank to compute the p-value?

```
np.mean(sim _______ obs)
```

◯ < ◯ <= ◯ > ◯ >= ◯ ==

g) Eric's p-value is 0.03. If his p-value cutoff is 0.01, what does he conclude?

◯ He rejects the null in favor of the alternative.

◯ He accepts the null.

◯ He accepts the alternative.

◯ He fails to reject the null.

Name: _____

Question 7 (8 points)

Suppose Tiffany has a random sample of dogs. Select the most appropriate technique to answer each of the following questions using Tiffany's dog sample.

a) Do small dogs typically live longer than medium and large dogs?

◯ Standard hypothesis test. ◯ Permutation test. ◯ Bootstrapping.

b) Does Tiffany's sample have an even distribution of dog kinds?

◯ Standard hypothesis test. ◯ Permutation test. ◯ Bootstrapping.

c) What's the median weight for herding dogs?

◯ Standard hypothesis test. ◯ Permutation test. ◯ Bootstrapping.

d) Do dogs live longer than 12 years on average?

◯ Standard hypothesis test. ◯ Permutation test. ◯ Bootstrapping.

Question 8 (24 points)

Oren has a random sample of 200 dog prices in an array called `oren`. He has also bootstrapped his sample 1,000 times and stored the mean of each resample in an array called `boots`.

In this question, assume that the following code has run:

```
a = np.mean(oren)
b = np.std(oren)
c = len(oren)
```

a) What expression best estimates the population's standard deviation?

◯ `b` ◯ `b / c` ◯ `b / np.sqrt(c)` ◯ `b * np.sqrt(c)`

b) Which expression best estimates the mean of `boots`?

◯ `0` ◯ `a` ◯ `(oren - a).mean()` ◯ `(oren - a) / b`

c) What expression best estimates the standard deviation of `boots`?

◯ `b` ◯ `b / c` ◯ `b / np.sqrt(c)` ◯ `(a - b) / np.sqrt(c)`

(continued on next page)

The variables for this question are repeated here for convenience.

```
a = np.mean(oren)
b = np.std(oren)
c = len(oren)
```

d) What is a price of \$560 in standard units?

◯ `(560 - a) / b`

◯ `(560 - a) / (b / np.sqrt(c))`

◯ `(a - 560) / (b / np.sqrt(c))`

◯ `abs(560 - a) / b`

◯ `abs(560 - a) / (b / np.sqrt(c))`

e) The distribution of `boots` is normal because of the Central Limit Theorem.

◯ True ◯ False

f) If Oren's sample was 400 dogs instead of 200, the standard deviation of `boots` will...

◯ Increase by a factor of 2 ◯ Increase by a factor of $\sqrt{2}$

◯ Decrease by a factor of 2 ◯ Decrease by a factor of $\sqrt{2}$

◯ None of the above

g) If Oren took 4,000 bootstrap resamples instead of 1,000, the standard deviation of `boots` will...

◯ Increase by a factor of 4 ◯ Increase by a factor of 2

◯ Decrease by a factor of 4 ◯ Decrease by a factor of 2

◯ None of the above

h) Write one line of code that evaluates to the **right endpoint** of a 92% normal theory confidence interval for the mean dog price. The following expressions may help:

```
stats.norm.cdf(1.75) # => 0.96
stats.norm.cdf(1.4)  # => 0.92
```

Name: _____

Question 9 (3 points)

Suppose you want to estimate the proportion of UCSD students that love dogs using a survey with a yes/no question. If you poll 400 students, what is the widest possible width for a 95% confidence interval?

◯ 0.01 ◯ 0.05 ◯ 0.1 ◯ 0.2 ◯ 0.4 ◯ None of the above

Question 10 (12 points)

Sam wants to fit a linear model to predict a dog's height using its weight.

He first runs the following code:

```
x = df.get('weight')
y = df.get('height')

def su(vals):
    return (vals - vals.mean()) / np.std(vals)
```

a) Select all of the Python snippets that correctly compute the correlation coefficient into the variable `r`.

☐ Snippet 1:

```
r = (su(x) * su(y)).mean()
```

☐ Snippet 2:

```
r = su(x * y).mean()
```

☐ Snippet 3:

```
t = 0
for i in range(len(x)):
    t = t + su(x[i]) * su(y[i])
r = t / len(x)
```

☐ Snippet 4:

```
t = np.array([])
for i in range(len(x)):
    t = np.append(t, su(x)[i] * su(y)[i])
r = t.mean()
```

(continued on next page)

b) Sam computes the following statistics for his sample:

statistic	value
correlation coefficient	0.8
mean dog weight	15
SD of dog weights	8
mean dog height	50
SD of dog heights	??

The best-fit line predicts that a dog with a weight of 10 kg has a height of 45 cm.

What is the SD of dog heights?

◯ 2 ◯ 4.5 ◯ 10 ◯ 25 ◯ 45 ◯ None of the above

c) Assume that the statistics in part b) still hold. Select all of the statements below that are true. (You don't need to finish part b) in order to solve this question.)

☐ The relationship between dog weight and height is linear.

☐ The root mean squared error of the best-fit line is smaller than 5.

☐ The best-fit line predicts that a dog that weighs 15 kg will be 50 cm tall.

☐ The best-fit line predicts that a dog that weighs 10 kg will be shorter than 50 cm.

Question 11 (20 points)

Answer the following true/false questions.

a) When John Snow took off the handle from the Broad Street pump, the number of cholera deaths decreased. Because of this, he could conclude that cholera was caused by dirty water.

◯ True ◯ False

b) If you get a p-value of 0.009 using a permutation test, you can conclude causality.

◯ True ◯ False

c) `df.groupby("kind").mean()` will have 5 columns.

◯ True ◯ False

d) If the 95% confidence interval for dog price is [650, 900], there is a 95% chance that the population dog price is between \$650 and \$900.

◯ True ◯ False

e) For a given sample, an 90% confidence interval is narrower than a 95% confidence interval.

◯ True ◯ False

Name: _____

f) If you tried to bootstrap your sample 1,000 times but accidentally sampled **without** replacement, the standard deviation of your test statistics is always equal to 0.

◯ True ◯ False

g) You run a permutation test and store 500 simulated test statistics in an array called `stats`. You can construct a 95% confidence interval by finding the 2.5th and 97.5th percentiles of `stats`.

◯ True ◯ False

h) The distribution of sample proportions is roughly normal for large samples because of the Central Limit Theorem.

◯ True ◯ False

i) The 20th percentile of the sequence [10, 30, 50, 40, 9, 70] is 30.

◯ True ◯ False

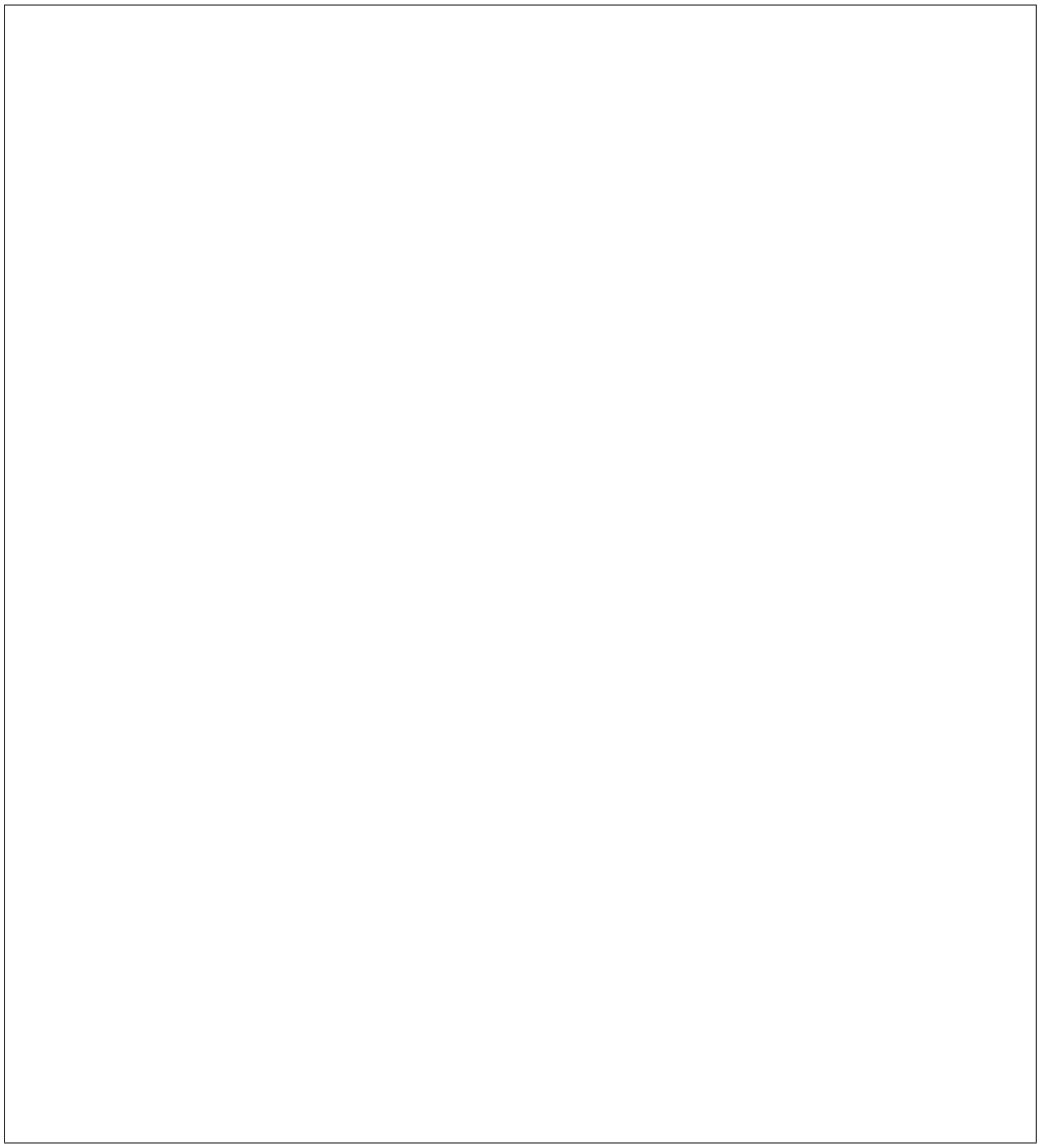
j) Chebyshev's inequality implies that we can always create a 96% confidence interval from a bootstrap distribution using the mean of the distribution $\pm$ 5 standard deviations.

◯ True ◯ False

This is the end of the exam.

Congrats—you're finished with DSC 10!
Before turning in your exam, please make sure that your PID or name is on every page.

Feel free to leave us a drawing here.



Woof!

The American Kennel Club (AKC) organizes information about dog breeds. We've loaded their dataset into a DataFrame called `df`. The index of `df` contains the dog breed names as `str` values.

The columns are:

- `"kind" (str)`: the kind of dog (herding, hound, toy, etc.). There are six total kinds.
- `"size" (str)`: small, medium, or large.
- `"longevity" (float)`: typical lifetime (years).
- `"price" (float)`: average purchase price (dollars).
- `"kids" (int)`: suitability for children. A value of `1` means high suitability, `2` means medium, and `3` means low.
- `"weight" (float)`: typical weight (kg).
- `"height" (float)`: typical height (cm).

The rows of `df` are arranged in **no particular order**. The first five rows of `df` are shown below (though `df` has **many more rows** than pictured here).

breed	kind	size	longevity	price	kids	weight	height
Siberian Husky	working	medium	12.58	650.0	1	22.00	55.0
Shetland Sheepdog	herding	small	12.53	465.0	1	10.15	37.0
Chihuahua	toy	small	16.50	588.0	3	2.00	13.0
Labrador Retriever	sporting	medium	12.04	810.0	1	31.00	58.0
Maltese	toy	small	12.25	650.0	3	2.00	23.0

Assume:

- We have already run `import babypandas as bpd` and `import numpy as np`.

Throughout this exam, we will refer to `df` repeatedly.