Final Exam - DSC 10, Winter 2026

Full Name:

PID:

Section: ◯ 9am ◯ 11am

Instructions:

- This exam consists of 12 questions, worth a total of 100 points.
- Write your PID in the top right corner of each page in the space provided.
- Please write **clearly** in the provided answer boxes; we will not grade work that appears elsewhere. Completely fill in bubbles and square boxes; if we cannot tell which option(s) you selected, you may lose points.

  ◯ A bubble means that you should only **select one choice**.

  ☐ A square box means you should **select all that apply**.
- For full credit, your solutions must use methods of the course.
- You may use one page of double-sided handwritten notes. Aside from this, you may not refer to any other resources or technology during the exam. No calculators!

Full name of person to your left:

Full name of person to your right:

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

Signature:

Version A

Please do not open your exam until instructed to do so.

Important: Before proceeding, make sure that you have the data description reference sheet.

This sheet is intentionally blank. Feel free to use it as scratch paper.

PID: _____

Question 1 (8 pts)

A silly capybara has scrambled up Jeffrey's Python code! Unscramble the code so that after it runs, `jeffrey` is **a single string containing the name of the exhibit with the largest mean reptile weight**.

```
A. jeffrey = jeffrey.get(['exhibit', 'weight_lb'])
B. jeffrey = jeffrey.get(['kind', 'weight_lb'])
C. jeffrey = jeffrey[jeffrey.get('kind') == 'Reptile']
D. jeffrey = jeffrey[jeffrey.get('species') == 'Reptile']
E. jeffrey = zoo
F. jeffrey = jeffrey.loc[0]
G. jeffrey = jeffrey.iloc[0]
H. jeffrey = jeffrey.index[0]
I. jeffrey = jeffrey.max()
J. jeffrey = jeffrey.mean()
K. jeffrey = jeffrey.reset_index()
L. jeffrey = jeffrey.sort_values(by='weight_lb', ascending=False)
M. jeffrey = jeffrey.sort_values(by='weight_lb', ascending=True)
N. jeffrey = jeffrey.groupby('exhibit')
```

For each of Lines 1–9, select a line from A–N such that the code produces the desired result when run in order. Your code may not need all 9 lines — fill in "Not used" for any remaining lines at the end. Line 1 has been filled in for you. Not all of the code above will be used, and some lines may be used more than once.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	Not used
Line 1	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○
Line 2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Line 3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Line 4	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Line 5	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Line 6	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Line 7	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Line 8	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Line 9	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Question 2 (4 pts)

Kate is the San Diego Zoo's newest zookeeper! She creates the DataFrames `first_5` and `first_7` using the code below.

```
first_5 = zoo.take(np.arange(5))
first_7 = zoo.take(np.arange(7))
```

a) Kate's first merge call is shown below.

```
merged_a = first_7.merge(first_7, on='species')
```

How many rows does `merged_a` have? Write your answer as a **single number** in the box, or fill in the bubble for "Not Enough Information."

◯ Not Enough Information

b) Kate's second merge call is shown below.

```
merged_b = merged_a.merge(first_5, on='species')
```

How many rows does `merged_b` have? Write your answer as a **single number** in the box, or fill in the bubble for "Not Enough Information."

◯ Not Enough Information

PID: _____

Question 3 (4 pts)

For this problem, let `weight` and `age` be defined by:

```
weight = np.array(zoo.get('weight_lb'))
age = np.array(zoo.get('age'))
```

a) Avi wants to quickly estimate each animal's weight in kilograms instead of pounds using the shortcut

$$\text{kg} \approx 0.45 \cdot \text{lbs}.$$

Which of the following expressions output an array of the animal weights in kilograms using this shortcut? Select all that apply.

```
A. weight * 0.45
B. weight - 0.55
C. weight * 0.90 / 2
D. weight / 0.45
```

☐ A ☐ B ☐ C ☐ D

b) Michelle wants to find the range of `age`, i.e., the age of the oldest zoo animal minus the age of the youngest zoo animal.

Which of the following expressions **incorrectly** computes this value?

```
A. age.max() - age.min()
B. (age - age.max()).min()
C. (age.max() - age).max()
D. (age - age.min()).max()
```

◯ A ◯ B ◯ C ◯ D

Question 4 (8 pts)

Bianca is writing helper functions for the San Diego Zoo's animal tracking tools. Complete the code below so that `mystery` returns a string made of the first letter of each word in a species name. Then, Bianca can use blank `(b)` to create `bianca`, a copy of `zoo` with one extra column called `initials`. For example, `mystery("Giant Panda")` should return `"GP"`.

```
def mystery(value):
    _____(a)_____
    return result
bianca = zoo.assign(initials=_____(b)_____)
```

a) Which snippets could replace blank `(a)`? Select all that apply.

☐
```
result = ""
for word in value:
    result = result + word[0]
```

☐
```
words = value.split()
result = ""
for word in words:
    result = result + word[0]
```

☐
```
temp = bpd.DataFrame().assign(word=value.split())
result = temp.get('word').get(0)
```

☐
```
words = np.array(value.split())
for i in np.arange(len(words)):
    words[i] = words[i][0]
result = ''.join(words)
```

b) Which snippets could replace blank `(b)`? Select all that apply.

☐ `zoo.get('species').apply(mystery)`

☐ `zoo.apply(mystery).get('species')`

☐ `mystery(zoo.get('species'))`

☐ `zoo.mystery('species')`

PID: _____

Question 5 (4 pts)

Sam has subsetted `zoo` to include only the animals whose `status` is `"Endangered"`.

Suppose that in Sam's subset:

- $\frac{2}{5}$ of the animals are mammals,
- $\frac{1}{4}$ of the animals live in the Asian Passage, and
- $\frac{1}{10}$ of the animals are mammals that live in the Asian Passage.

a) What is the probability that a randomly selected animal from Sam's sample lives in the Asian Passage but is **not** a mammal? Give your answer as a simplified fraction.

☐ ◯ Not Enough Information

b) Sam chooses two animals uniformly at random with replacement from his sample. What is the probability that **at least one** of the two animals is a mammal living in the Asian Passage? Give your answer as a simplified fraction.

☐ ◯ Not Enough Information

Question 6 (8 pts)

Sofia gets a summer internship at the San Diego Zoo. Her first task is to study the relationship between conservation status and how much food an animal eats each day. She treats `zoo` as a SRS of animals from zoos worldwide. Complete the code below so that `bootstrap(zoo)` returns a 75% bootstrap confidence interval for the following statistic:

median daily food of Endangered animals − median daily food of Vulnerable animals.

```
def bootstrap(df):
    vulnerable = df[df.get('status') == 'Vulnerable']
    endangered = df[df.get('status') == 'Endangered']

    median_diffs = []

    for i in range(1000):
        _____(a)_____
    return np.percentile(median_diffs, _____(b)_____)
```

a) Which snippets could replace blank `(a)`? Select all that apply.

☐
```
sample_v = df.sample(vulnerable.shape[0], replace=True)
sample_e = df.sample(endangered.shape[0], replace=True)
median_diffs.append(sample_e.get('daily_food_lb').median()
                    - sample_v.get('daily_food_lb').median())
```

☐
```
sample_v = vulnerable.sample(vulnerable.shape[0], replace=True)
sample_e = endangered.sample(endangered.shape[0], replace=True)
median_diffs.append(sample_e.get('daily_food_lb').median()
                    - sample_v.get('daily_food_lb').median())
```

☐
```
v_foods = np.array(vulnerable.get('daily_food_lb'))
e_foods = np.array(endangered.get('daily_food_lb'))
sample_v = np.random.choice(v_foods, len(v_foods), replace=True)
sample_e = np.random.choice(e_foods, len(e_foods), replace=True)
median_diffs.append(np.median(sample_e) - np.median(sample_v))
```

☐ None of the above

b) Which expression should replace blank `(b)`?

◯ `75` ◯ `87.5` ◯ `[25, 75]` ◯ `[12.5, 87.5]` ◯ `[75, 25]`

PID: _____

Question 7 (12 pts)

For this question, assume that Peter ran bootstrap code similar to what we saw in lecture to bootstrap the `age` column from `zoo` 5,000 times and computed the median ages of each bootstrap sample in the variable `ages5000`. Then, he bootstraps the median age from `zoo` 1,000 additional times and stores the ages in the variable `ages1000`:

```
>>> ages1000
array([5, 9, 12, ..., 16, 19, 12], shape=(1000,))

>>> ages5000
array([12, 7, 13, ..., 11, 11, 16], shape=(5000,))
```

a) Suppose that Peter creates a 95% bootstrap confidence interval using `ages5000`. Select all true statements about this confidence interval.

- ☐ The total width of this confidence interval can be calculated using the Central Limit Theorem as described in class, without bootstrapping.
- ☐ The midpoint of the confidence interval is exactly equal to: `np.median(zoo.get('age'))`.
- ☐ This confidence interval gives a valid estimate of the median age of animals in all zoos even if the San Diego Zoo is a nonrandom sample of animals in all zoos.
- ☐ If Peter's target population is the animals in the San Diego Zoo, this confidence interval isn't useful.
- ☐ None of the above

b) Select all true statements about `ages5000` and `ages1000`.

- ☐ Because `ages5000` contains more bootstrap medians, its variance should be smaller than the variance of `ages1000`.
- ☐ Because `ages5000` contains more bootstrap medians, a 95% confidence interval created using `ages5000` will be narrower than a 95% confidence interval created using `ages1000`.
- ☐ The variances of `ages1000` and `ages5000` should be similar because both arrays were generated by bootstrapping the same sample and computing the same statistic.
- ☐ The variances of `ages1000` and `ages5000` might not be exactly equal because the bootstrap procedure involves randomness.
- ☐ None of the above

Question 8 (8 pts)

Raymond wants to estimate the average daily food consumption, in pounds, of animals at the San Diego Zoo. He runs `my_sample = zoo.sample(30)`. His sample consumes `1200` pounds of food per day in total, with a standard deviation of approximately `88`.

a) A 95% CLT-based confidence interval for the average daily amount of food consumed can be approximately expressed as $[x, 9x]$, where x is a positive number. What is x? Give your answer as a single simplified fraction or whole number.
(Hint: $\sqrt{30} \approx \frac{11}{2}$.)

$\square$ ◯ Not Enough Information

b) Let z be a positive real number so that some percentage of values on a normal distribution fall within $\pm z$ standard deviations of the mean. Using the same sample as in part (a), Raymond makes a valid CLT-based confidence interval for the average daily amount of food consumed where the right endpoint is $\frac{5}{4}$ times the midpoint of the interval.

What is the approximate value of z used to create this interval? Give your answer as a simplified fraction or whole number.
(Hint: you may again use $\sqrt{30} \approx \frac{11}{2}$.)

$\square$ ◯ Not Enough Information

c) Ella takes a simple random sample from a different zoo and creates a CLT-based 95% confidence interval for the average daily amount of food consumed. Raymond does the same using his sample at 90% confidence. Ella's interval is narrower than Raymond's. Could this be reasonable? Fill in the bubble next to the best explanation.

◯ No. A smaller confidence level results in a narrower confidence interval.
◯ Yes. A smaller confidence level results in a wider confidence interval.
◯ Yes. Ella might have a smaller sample size than Raymond.
◯ Yes. Ella's sample might have a smaller standard deviation.

d) For this part, assume that there are exactly 12,000 animals at the San Diego Zoo. Raymond now wants to estimate the **total** amount of food all of them consume per day. If a CLT-based confidence interval for the *average* daily amount of food is $[a, b]$, what is the corresponding confidence interval for the *total* daily amount of food?

◯ $[30a, 30b]$
◯ $[\sqrt{12000} \cdot a, \sqrt{12000} \cdot b]$
◯ $[12000a, 12000b]$
◯ $[a + 12000, b + 12000]$
◯ This cannot be determined from the information given.

Question 9 (12 pts)

Punch is a baby Japanese macaque at Ichikawa City Zoo in Japan who went viral in early 2026 after being abandoned by his mother. Zookeepers hand-reared him and gave him a stuffed orangutan plush toy for comfort. As Punch began integrating into his troop, zookeepers tracked his interactions with the other macaques. They recorded his last 50 interactions and classified each as either **friendly** or **unfriendly**. Out of 50 interactions, only 20 were friendly.

For a typical macaque in this troop, zookeepers have observed that 50% of interactions are friendly. Minchan wants to test whether Punch is being treated the same as a typical macaque, or whether he is being treated worse than the other macaques. He performs a simulation-based hypothesis test with the following hypotheses:

> Null: The probability that any given interaction with Punch is friendly is `0.5`.
>
> Alternative: The probability is less than `0.5`.

Help Minchan debug the code below so that after making the changes in questions b)–e) and running the code, `p_value` contains a valid p-value for the hypothesis test. Assume that `total_variation_distance(dist1, dist2)` computes the TVD as defined in class.

```
A. expected = 50 * 0.5
B. obs = abs(20 - expected)
C. stats = np.array([])
D. for i in np.arange(1000):
E.     n = np.random.multinomial(50, [0.5, 0.5])[0]
F.     stat = abs(n - expected)
G.     stats = np.append(stats, stat)
H. p_value = np.count_nonzero(stats >= obs) / 1000
```

a) Suppose that Minchan runs the code as-is without making any changes. Select all true statements.

- ☐ A histogram of `stats` would be centered at 0.
- ☐ A histogram of `stats` would have a peak at 5.
- ☐ `p_value` is approximately twice as large as it should be.
- ☐ `p_value` is more than twice as large as it should be.
- ☐ None of the above are true.

b) Line B should be:

- ◯ Left as-is.
- ◯ `obs = 20 - expected`
- ◯ `obs = 20 / 50 - expected`
- ◯ `obs = abs(20 / 50 - expected / 50)`
- ◯ `obs = total_variation_distance([0.4, 0.6], [0.5, 0.5])`

Minchan's code is repeated here for convenience:

```
A. expected = 50 * 0.5
B. obs = abs(20 - expected)
C. stats = np.array([])
D. for i in np.arange(1000):
E.     n = np.random.multinomial(50, [0.5, 0.5])[0]
F.     stat = abs(n - expected)
G.     stats = np.append(stats, stat)
H. p_value = np.count_nonzero(stats >= obs) / 1000
```

c) Line E should be:

- ◯ Left as-is.
- ◯ `n = np.random.choice(['Friendly', 'Unfriendly'], 50)`
- ◯ `n = np.random.multinomial(50, [0.4, 0.6])[0]`
- ◯ `n = np.random.multinomial(20, [0.5, 0.5])[0]`
- ◯ `n = np.random.multinomial(50, [0.4, 0.6])[0]`
- ◯ `n = np.random.multinomial(50, [0.5, 0.5])`

d) Line F should be:

- ◯ Left as-is.
- ◯ `stat = obs`
- ◯ `stat = abs(n - 20)`
- ◯ `stat = total_variation_distance([n/50, 1 - n/50], [0.5, 0.5])`
- ◯ `stat = n - expected`
- ◯ `stat = 20 - expected`
- ◯ `stat = n - 20`

e) Line H should be:

- ◯ Left as-is.
- ◯ `np.count_nonzero(stats >= obs) / 50`
- ◯ `np.count_nonzero(stats <= obs) / 50`
- ◯ `np.count_nonzero(stats <= obs) / 1000`

PID: _____

Question 10 (20 pts)

Austin wants to test whether mammals and birds have the same proportion of "fast" animals, where "fast" means having a maximum speed above 40 mph. He performs a permutation test using a DataFrame called `df` with two columns:

- `kind`: either `"Mammal"` or `"Bird"`.
- `is_fast`: **`True`** if the animal's maximum speed is above 40 mph, **`False`** otherwise.

`df` contains only the rows of `zoo` where `kind` is either `"Mammal"` or `"Bird"`. For this question, assume that the animals in `df` are a representative sample of all mammals and birds.

a) Select all valid statements of the null hypothesis for Austin's permutation test.

- ☐ The distribution of `is_fast` is the same for mammals and birds.
- ☐ The proportion of fast animals is the same among mammals and birds.
- ☐ The proportion of fast animals among mammals and birds in `df` is the same.
- ☐ The average maximum speed is the same for mammals and birds.
- ☐ There is no association between `kind` and `is_fast`.
- ☐ The proportion of fast animals among both mammals and birds is `0.5`.
- ☐ None of the above.

b) Select the valid statement of the alternative hypothesis for Austin's test.

- ◯ The proportion of fast mammals minus the proportion of fast birds $\neq 0$.
- ◯ The distribution of `kind` is different for fast and non-fast animals.
- ◯ The proportion of fast animals among both mammals and birds is not `0.5`.
- ◯ None of the choices above are valid.

c) Select the test statistic that would be valid for this hypothesis test.

- ◯ Proportion of fast mammals minus the proportion of fast birds.
- ◯ Total variation distance between the distribution of `kind` among fast animals and the distribution of `kind` among non-fast animals.
- ◯ Absolute difference between the number of fast mammals and the number of fast birds.
- ◯ None of the choices above are valid.

d) Which of the following is the correct way to simulate new samples under the null hypothesis?

- ◯ Shuffle the `kind` column of `df` and recompute the test statistic.
- ◯ Resample rows of `df` with replacement and recompute the test statistic.
- ◯ For each animal in `df`, flip a fair coin to assign `is_fast` to **`True`** or **`False`**, then recompute the test statistic.
- ◯ Separately resample the mammals and birds in `df` with replacement, then recompute the test statistic.

(For convenience, we've repeated the question's information here:)

Austin wants to test whether mammals and birds have the same proportion of "fast" animals, where "fast" means having a maximum speed above 40 mph. He performs a permutation test using a DataFrame called `df` with two columns:

- `kind`: either `"Mammal"` or `"Bird"`.
- `is_fast`: `True` if the animal's maximum speed is above 40 mph, `False` otherwise.

`df` contains only the rows of `zoo` where `kind` is either `"Mammal"` or `"Bird"`. For this question, assume that the animals in `df` are a representative sample of all mammals and birds.

e) Consider the function defined below.

```
def stat(tbl):
    props = tbl.groupby("kind").mean().get("is_fast")
    return np.std(props)
```

Suppose Austin uses this function as his test statistic and simulates 1,000 values under the null hypothesis. Select all true statements about the resulting empirical distribution.

☐ The distribution is symmetric and centered at 0.

☐ The distribution has a peak at 0.

☐ The distribution contains negative values.

☐ `stat(df)` would be a valid observed test statistic for this test.

☐ If the same shuffles are used, computing the test statistic with `stat` yields exactly the same p-value as using the absolute difference in proportions.

☐ If the same shuffles are used, computing the test statistic with `stat` yields exactly the same p-value as using the TVD between the distribution of `is_fast` among mammals and the distribution of `is_fast` among birds.

☐ None of the above.

f) Austin wants to estimate the proportion of mammals that are fast using a 95% CLT-based confidence interval. He wants the width of the interval to be at most 0.1, regardless of the true proportion. What is the minimum number of mammals he needs to sample? Answer with a single integer.

PID: _____

Question 11 (12 pts)

Hemanth finds the regression line predicting `daily_food_lb` from `weight_lb`:

$$\text{predicted } \texttt{daily\_food\_lb} = 0.04\ \texttt{weight\_lb}\ + 2$$

For this problem, $r = 0.8$, the mean of `weight_lb` is 200 pounds, the standard deviation of `weight_lb` is 100 pounds, and the standard deviation of `daily_food_lb` is 5 pounds.

a) Select all statements that must be true about Hemanth's regression line.

☐ Approximately 80% of the data points in `zoo` fall on or near the regression line.

☐ It's possible to find a different line with a lower average of squared residuals.

☐ It is not possible to bootstrap the dataset and get the exact same regression line.

☐ None of the above.

b) What is the mean of `daily_food_lb`? Answer with a single number.

c) Hemanth converts both `weight_lb` and `daily_food_lb` to kilograms (1 lb $\approx$ 0.45 kg), then fits a new regression line predicting `daily_food_kg` from `weight_kg`. Compared to the original regression line above, does each of the following increase, decrease, or stay the same?

Slope:	◯ Increase	◯ Decrease	◯ Stay the same
Intercept:	◯ Increase	◯ Decrease	◯ Stay the same
r:	◯ Increase	◯ Decrease	◯ Stay the same

d) Hemanth standardizes `weight_lb`, then refits the regression model predicting `daily_food_lb`. Compared to the original regression line above, does each of the following increase, decrease, or stay the same?

Slope:	◯ Increase	◯ Decrease	◯ Stay the same
Intercept:	◯ Increase	◯ Decrease	◯ Stay the same
r:	◯ Increase	◯ Decrease	◯ Stay the same

e) Hemanth bootstraps the dataset 1000 times and computes a regression line for each resample. He uses each bootstrapped line to predict `daily_food_lb` for a 150-pound animal and for a 400-pound animal. For which animal will the 1000 predictions vary more?

◯ The 150-pound animal.

◯ The 400-pound animal.

◯ The predictions will vary equally for both.

◯ Need more information.

Question 12 (0 pts)

Congratulations on finishing DSC 10! Before you turn in your exam, a few reminders:

1. Make sure you've written your student ID (PID) on each page of the exam.
2. Fill out bubbles and squares **completely** — no check marks.

If you'd like, draw a picture about DSC 10 in the box below.