Final Exam - DSC 10, Summer Session I 2024

Full Name:

PID:

Instructions:

- This exam consists of 9 questions, worth a total of 131.5 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**.
- Aside from the provided Reference Sheet, you may not refer to any other resources or technology during the exam (no notes, no calculators).

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

Signature:

Version B

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

Important: Before proceeding, make sure to rip off the last sheet of paper from this exam packet and read the data description.

Question 1 (23 pts)

a) (6 pts) Fill in the blanks so that the expression below evaluates to the *proportion* of stages won by the country with the most stage wins.

```
stages.groupby(__(i)__).__(ii)__.get("Type").__(iii)__ / stages.shape[0]
```

`(i)` (2 pts):

`(ii)` (2 pts):

`(iii)` (2 pts):

b) (5 pts) The distance of a stage alone does not encapsulate its difficulty, as riders feel more tired as the tour goes on. Because of this, we want to consider "real distance," a measurement of the length of a stage that takes into account how far into the tour the riders are. The "real distance" is calculated with the following process:

(i) Add one to the stage number.
(ii) Take the square root of the result of (i).
(iii) Multiply the result of (ii) by the raw distance of the stage.

Complete the implementation of the function `real_distance`, which takes in `stages` (a DataFrame), `stage` (a string, the name of the column containing stage numbers), and `distance` (a string, the name of the column containing stage distances). `real_distance` returns a Series containing all of the "real distances" of the stages, as calculated above.

```
def real_distance(stages, stage, distance):
    --------
```

c) (6 pts) Sometimes, stages are repeated in different editions of the Tour de France, meaning that there are some pairs of `"Origin"` and `"Destination"` that appear more than once in `stages`. Fill in the blanks so that the expression below evaluates how often the most common `"Origin"` and `"Destination"` pair in the `stages` DataFrame appears.

```
stages.groupby(__(i)__).__(ii)__.sort_values(by = "Date").get("Type").iloc[__(iii)__]
```

`(i)` (2 pts):

`(ii)` (2 pts):

`(iii)` (2 pts):

d) (6 pts) Fill in the blanks so that the value of `mystery_three` is the `"Destination"` of the longest stage before Stage 12.

```
mystery = stages[stages.get(__(i)__) < 12]
mystery_two = mystery.sort_values(by = "Distance", ascending = __(ii)__)
mystery_three = mystery_two.get(__(iii)__).iloc[-1]
```

`(i)` (2 pts):

`(ii)` (2 pts):

`(iii)` (2 pts):

Question 2 (6 pts)

Suppose we run the following code to simulate the winners of the Tour de France.

```
evenepoel_wins = 0
vingegaard_wins = 0
pogacar_wins = 0
for i in np.arange(4):
    result = np.random.multinomial(1, [0.3, 0.3, 0.4])
    if result[0] == 1:
        evenepoel_wins = evenepoel_wins + 1
    elif result[1] == 1:
        vingegaard_wins = vingegaard_wins + 1
    elif result[2] == 1:
        pogacar_wins = pogacar_wins + 1
```

a) (3 pts) What is the probability that `pogacar_wins` is equal to 4 when the code finishes running? Do not simplify your answer.

b) (3 pts) What is the probability that `evenepoel_wins` is at least 1 when the code finishes running? Do not simplify your answer.

PID: _____

Question 3 (12 pts)

We want to estimate the mean distance of Tour de France stages by bootstrapping 10,000 times and constructing a 90% confidence interval for the mean. In this question, suppose `random_stages` is a random sample of size 500 drawn with replacement from `stages`. Identify the line numbers with errors in the code below. In the adjacent box, point out the error by describing the mistake in less than 10 words or writing a code snippet (correct only the part you think is wrong). You may or may not need all the spaces provided below to identify errors.

```
line 1:      means = np.array([])
line 2:
line 3:      for i in 10000:
line 4:          resample = random_stages.sample(10000)
line 5:          resample_mean = resample.get("Distance").mean()
line 6:          np.append(means, resample_mean)
line 7:
line 8:      left_bound = np.percentile(means, 0)
line 9:      right_bound = np.percentile(means, 90)
```

a) | | |

b) | | |

c) | | |

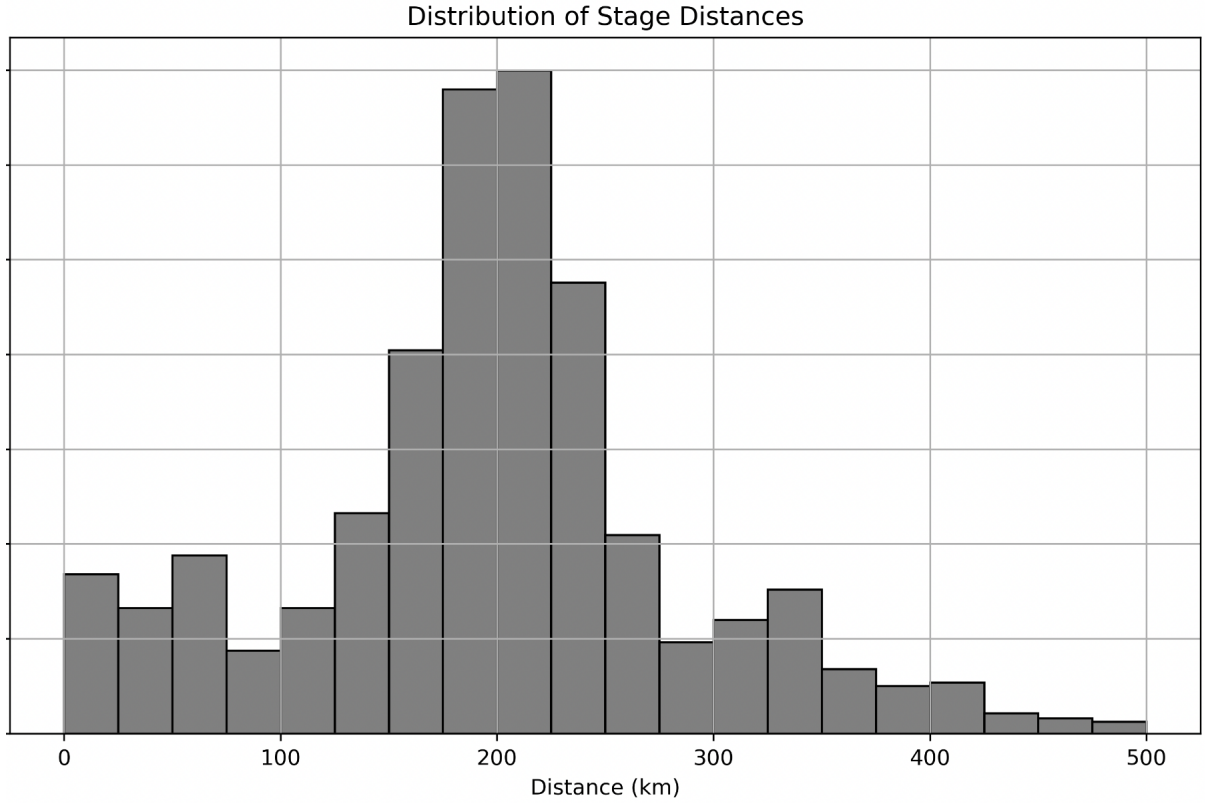
d) | | |

e) | | |

f) | | |

Question 4 (16.5 pts)

Below is a density histogram representing the distribution of randomly sampled stage distances.



a) (1 pt) Which statement below correctly describes the relationship between the mean and the median of the sampled stage distances?

- ◯ The mean is significantly larger than the median.
- ◯ The mean is significantly smaller than the median.
- ◯ The mean is approximately equal to the median.
- ◯ It is impossible to know the relationship between the mean and the median.

b) (2 pts) Assume there are 100 stages in the random sample that generated this plot. If there are 5 stages in the bin `[275, 300)`, approximately how many stages are in the bin `[200, 225)`?

c) (3 pts) Assume the mean distance is 200 km and the standard deviation is 50 km. At least what proportion of stage distances are guaranteed to lie between 0 km and 400 km? Do not simplify your answer.

d) (3 pts) Again, assume the mean stage distance is 200 km and the standard deviation is 50 km. Now, suppose we take a random sample of size 25 from the stage distances, calculate the mean stage distance of this sample, and repeat this process 500 times. What proportion of the means that we calculate will fall between 190 km and 210 km? Do not simplify your answer.

e) (3.5 pts) Assume the mean distance is 200 km and the standard deviation is 50 km. Suppose we use the Central Limit Theorem to generate a 95% confidence interval for the true mean distance of all Tour de France stages, and get the interval [190 km, 210 km]. Which of the following interpretations of this confidence interval are correct?

- ☐ 95% of Tour de France stage distances fall between 190 km and 210 km.
- ☐ There is a 95% chance that the true mean distance of all Tour de France stages is between 190 km and 210 km.
- ☐ We are 95% confident that the true mean distance of all Tour de France stages is between 190 km and 210 km.
- ☐ Our sample is of size 100.
- ☐ Our sample is of size 25.
- ☐ If we collected many original samples and constructed many 95% confidence intervals, then exactly 95% of those intervals would contain the true mean distance.
- ☐ If we collected many original samples and constructed many 95% confidence intervals, then roughly 95% of those intervals would contain the true mean distance.

f) (4 pts) Suppose we take 500 random samples of size 100 from the stage distances, calculate their means, and draw a histogram of the distribution of these sample means. We label this Histogram A. Then, we take 500 random samples of size 1000 from the stage distances, calculate their means, and draw a histogram of the distribution of these sample means. We label this Histogram B. Fill in the blanks so that the sentence below correctly describes how Histogram B looks in comparison to Histogram A.

"Relative to Histogram A, Histogram B would appear __(i)__ and shifted __(ii)__ due to the __(iii)__ mean and the __(iv)__ standard deviation."

(i): ◯ thinner ◯ wider ◯ the same width ◯ unknown

(ii): ◯ left ◯ right ◯ not at all ◯ unknown

(iii): ◯ larger ◯ smaller ◯ unchanged ◯ unknown

(iv): ◯ larger ◯ smaller ◯ unchanged ◯ unknown

Question 5 (22 pts)

In this question, suppose `random_stages` is a random sample of undetermined size drawn with replacement from `stages`. We want to estimate the proportion of stage wins won by each country.

a) (2 pts) Suppose we extract the winning countries and store the resulting Series. Consider the variable `winners` defined below, which you may use throughout this question:

```
winners = random_stages.get("Winner Country")
```

Write a single line of code that evaluates to the proportion of stages in `random_stages` won by France (country code `"FRA"`).

b) (4 pts) We want to generate a 95% confidence interval for the true proportion of wins by France in `stages` by using our random sample `random_stages`. How many rows need to be in `random_stages` for our confidence interval to have width of at most 0.03? Recall that the maximum standard deviation for any series of zeros and ones is 0.5. Do not simplify your answer.

c) (10 pts) Suppose we now want to test the hypothesis that the true proportion of stages won by Italy (`"ITA"`) is 0.2 using a confidence interval and the Central Limit Theorem. We want to conduct our hypothesis test at a significance level of 0.01. Fill in the blanks to construct the confidence interval `[interval_left, interval_right]`. Your answer must use the Central Limit Theorem, not bootstrapping. Assume an integer variable `sample_size = len(winners)` has been defined, regardless of your answer to **b)**.

Hint: `stats.norm.cdf(2.576) - stats.norm.cdf(-2.576) = 0.99`.

```
interval_center = __(i)__
mystery = __(ii)__ * np.std(__(iii)__ ) / __(iv)__
interval_left = interval_center - mystery
interval_right = interval_center + mystery
```

`(i)` (4 pts):

`(ii)` (2 pts):

`(iii)` (2 pts):

`(iv)` (2 pts):

d) (2 pts) What is our null hypothesis?

◯ The true proportion of stages won by Italy is 0.2.
◯ The true proportion of stages won by Italy is not 0.2.
◯ The true proportion of stages won by Italy is greater than 0.2.
◯ The true proportion of stages won by Italy is less than 0.2.

e) (2 pts) What is our alternative hypothesis?

◯ The true proportion of stages won by Italy is 0.2.
◯ The true proportion of stages won by Italy is not 0.2.
◯ The true proportion of stages won by Italy is greater than 0.2.
◯ The true proportion of stages won by Italy is less than 0.2.

f) (2 pts) Suppose we calculated the interval $[0.195, 0.253]$ using the above process. Should we reject or fail to reject our null hypothesis?

◯ Reject ◯ Fail to reject

Question 6 (20 pts)

You want to use the data in `stages` to test the following hypotheses:

- **Null Hypothesis**: In the Tour de France, the mean distance of flat stages is **equal to** the mean distance of mountain stages.
- **Alternative Hypothesis**: In the Tour de France, the mean distance of flat stages is **less than** the mean distance of mountain stages.

For the rest of this problem, assume you have assigned a new column to `stages` called `class`, which categorizes stages into either `flat` or `mountain` stages.

a) (5 pts) Which of the following test statistics could be used to test the given hypothesis? Select all that apply.

☐ The mean distance of flat stages divided by the mean distance of mountain stages.

☐ The difference between the mean distance of mountain stages and the mean distance of flat stages.

☐ The absolute difference between the mean distance of flat stages and the mean distance of mountain stages.

☐ One half of the difference between the mean distance of flat stages and the mean distance of mountain stages.

☐ The squared difference between the mean distance of flat stages and the mean distance of mountain stages.

Assume that for the rest of the question, we will be using the following test statistic:

The difference between the mean distance of flat stages and the mean distance of mountain stages.

b) (3 pts) Fill in the blanks in the code below so that it correctly conducts a hypothesis test of the given hypotheses and returns the p-value.

```
def hypothesis_test(stages):
    means = (stages.get(["class", "Distance"])
                   .groupby("class").mean().get("Distance"))
    observed_stat = means.loc["flat"] - means.loc["mountain"]

    simulated_stats = np.array([])
    for i in np.arange(10000):
        shuffled = stages.assign(shuffled = np.random.__(i)__(stages.get("Distance")))
        shuffled_means = (shuffled.get(["class", "shuffled"])
                                  .groupby("class").mean().get("shuffled"))
        simulated_stat = (shuffled_means.loc["flat"] - shuffled_means.loc["mountain"])
        simulated_stats = __(ii)__(simulated_stats, simulated_stat)

    p_value = np.__(iii)__(simulated_stats <= observed_stat)
    return p_value
```

i) _____

ii) _____

iii) _____

c) (6 pts) Indicate whether each of the following code snippets would correctly calculate `simulated_stat` inside the `for`-loop without errors. Where present, assume the blank `(i)` has been filled in correctly.

```
shuffled = stages.assign(shuffled = np.random.__(i)__(stages.get("Distance")))
shuffled_flat = (shuffled[shuffled.get("class") == "flat"].get("shuffled"))
shuffled_mountain = (shuffled[shuffled.get("class") == "mountain"].get("shuffled"))
simulated_stat = shuffled_flat.mean() - shuffled_mountain.mean()
```

(i) ◯ This code is correct. ◯ This code is incorrect or errors.

```
shuffled = stages.assign(shuffled = np.random.__(i)__(stages.get("class")))
shuffled_flat = (shuffled[shuffled.get("shuffled") == "flat"].get("Distance"))
shuffled_mountain = (shuffled[shuffled.get("shuffled") == "mountain"].get("Distance"))
simulated_stat = shuffled_flat.mean() - shuffled_mountain.mean()
```

(ii) ◯ This code is correct. ◯ This code is incorrect or errors.

```
shuffled = stages.assign(shuffled = np.random.__(i)__(stages.get("Distance")))
shuffled_means = (shuffled.get(["class", "Distance"])
                          .groupby("class").mean())
simulated_stat = (shuffled_means.get("Distance").iloc["flat"] -
                  shuffled_means.get("Distance").iloc["mountain"])
```

(iii) ◯ This code is correct. ◯ This code is incorrect or errors.

d) (2 pts) Assume that the observed statistic for this hypothesis test was equal to -22.5 km. Given that there are 10,000 simulated test statistics generated in the code above, at least how many of those must be equal to -22.5 km in order for us to **reject** the null hypothesis at an 0.05 significance level?

◯ 500 ◯ 5000 ◯ 0 ◯ 9500 ◯ 10000

e) (4 pts) Assume that the code above generated a p-value of 0.03. In the space below, please write your interpretation of this p-value. Your answer should include more than simply "we reject/fail to reject the null hypothesis."

Question 7 (12 pts)

You are analyzing the data in `stages` to see which countries winners come from.

You categorize the countries into four groups: France, Italy, Belgium, and Other. After performing some analysis, you find that the observed distribution of countries of origin for Tour de France stage winners is `[0.3, 0.2, 0.1, 0.4]`; i.e. 30% of stage winners are French, 20% Italian, etc. Based on census information, the expected population distribution is `[0.07, 0.06, 0.01, 0.86]`; that is, France's total population is 7% of the sum of the populations of all countries participating in the Tour, Italy's is 6%, etc.

You conduct a hypothesis test with the following hypotheses:

- **Null Hypothesis**: The distribution of countries of origin for Tour de France stage winners is equivalent to the distribution of the populations of the countries participating in the Tour de France. Any observed difference is due to chance.
- **Alternative Hypothesis**: The distribution of countries of origin for Tour de France stage winners is different from the distribution of the populations of the countries participating in the Tour de France.

a) (4 pts) Which of the following test statistics are appropriate for this hypothesis test? Select all that apply.

- ☐ The absolute difference between the expected proportion of French stage winners and the observed proportion of French stage winners.
- ☐ The sum of the differences between the expected population distribution and the observed distribution of stage winners.
- ☐ The absolute difference between the number of French stage winners and the number of Italian stage winners.
- ☐ The sum of the absolute differences between the expected population distribution and the observed distribution of stage winners.

For the rest of this question, assume that we will be using the Total Variation Distance as our test statistic.

b) (6 pts) Complete the implementation of the `simulate` and `calculate_test_stat` functions so that the code below successfully simulates 10,000 test statistics.

```
expected_dist = [0.07, 0.06, 0.01, 0.86]
observed_dist = [0.3, 0.2, 0.1, 0.4]

def simulate(__(i)__):
    simulated_winners = np.random.__(ii)__(100, __(iii)__)
    return simulated_winners / 100

def calculate_test_stat(__(iv)__, __(v)__):
    return __(vi)__


observed_stat = calculate_test_stat(observed_dist, expected_dist)
simulated_stats = np.array([])
for i in np.arange(10000):
    simulated_dist = simulate(expected_dist)
    simulated_stat = calculate_test_stat(simulated_dist, expected_dist)
    simulated_stats = np.append(simulated_stats, simulated_stat)
```

(i):

(ii):

(iii):

(iv):

(v):

(vi):

c) (2 pts) Fill in the blank in the following code so that `p_value` evaluates to the correct p-value for this hypothesis test:

```
p_value = np.mean(simulated_stats ___ observed_statistic))
```

◯ `>` ◯ `>=` ◯ `<` ◯ `<=` ◯ `=`

PID: _____

Question 8 (8 pts)

Choose the best tool to answer each of the following questions. Note the following:

- By "hypothesis testing," we mean "standard" hypothesis testing, i.e. hypothesis testing that doesn't involve permutation testing or bootstrapping.
- By "bootstrapping," we mean bootstrapping that doesn't involve hypothesis testing.

a) (2 pts) What is the median distance of all Tour de France stages?

◯ Hypothesis Testing
◯ Permutation Testing
◯ Bootstrapping
◯ Central Limit Theorem

b) (2 pts) Is the distribution of Tour de France stage types from before 1960 the same as after 1960?

◯ Hypothesis Testing
◯ Permutation Testing
◯ Bootstrapping
◯ Central Limit Theorem

c) (2 pts) Are there an equal number of destinations that start with letters from the first half of the alphabet and destinations that start with letters from the second half of the alphabet?

◯ Hypothesis Testing
◯ Permutation Testing
◯ Bootstrapping
◯ Central Limit Theorem

d) (2 pts) Are mountain stages with destinations in France from before 1970 longer than flat stages with destinations in Belgium from after 2000?

◯ Hypothesis Testing
◯ Permutation Testing
◯ Bootstrapping
◯ Central Limit Theorem

Question 9 (12 pts)

Suppose the **distance** of a Tour de France stage and the **time** it takes to complete it are linearly associated with correlation coefficient $r = \frac{2}{3}$. Assume distances have a mean of 200 km and a standard deviation of 80 km. Times have a mean of 6 hours.

a) (4 pts) Suppose the regression line to predict the time a stage will take (in hours) based on its length (in km) predicts that a 160 km long stage will take 5 hours. What is the standard deviation of the time it takes to complete a stage?

b) (4 pts) Suppose, regardless of your answer to **a)**, that stage completion times have a standard deviation of 1.5 hours. The other means, SD, and r are unchanged.

Stages in the middle of the Tour tend to be longer than those at the ends. Stage 14 is 60 km longer than stage 20, so we would expect it to take longer based on our linear association. How large will the difference in our predictions of stage completion times be?

c) (2 pts) Suppose a mandatory rest break of 30 minutes (0.5 hours) is implemented for all Tour de France stages. How would the slope of the regression line change?

- ◯ It would decrease.
- ◯ It would stay the same.
- ◯ It would increase.
- ◯ We cannot say for sure.

d) (2 pts) Suppose the means and standard deviations above do not change (continue to assume times have a standard deviation of 1.5 hours), but the correlation coefficient r is different. If we predict a 360 km stage will take 9 hours, what is the value of r? In the box below, write a single number for r or "N/A" if it is not possible to answer.

PID: _____

Feel free to draw us a picture or tell us about your fondest memory from DSC 10 this summer.

Before turning in your exam, please make sure that your PID is on every page.

Congratulations on finishing DSC 10!

 # Tour de France 

In this exam, you'll work with a data set representing the results of the Tour de France, a multi-stage, weeks-long cycling race. The Tour de France takes place over many days each year, and on each day, the riders compete in individual races called `stages`. Each `stage` is a standalone race, and the winner of the entire tour is determined by who performs the best across all of the individual `stages` combined. Each row represents one stage of the Tour (or equivalently, one day of racing). This dataset will be called `stages`.

The columns of `stages` are as follows:

- `"Stage" (int)`: The stage number for the respective year.
- `"Date" (str)`: The day that the stage took place, formatted as "YYYY-MM-DD."
- `"Distance" (float)`: The distance of the stage in kilometers.
- `"Origin" (str)`: The name of the city in which the stage starts.
- `"Destination" (str)`: The name of the city in which the stage ends.
- `"Type" (str)`: The type of the stage.
- `"Winner" (str)`: The name of the rider who won the stage
- `"Winner Country" (str)`: The country from which the winning rider of the stage is from.

The first few rows of `stages` are shown below, though `stages` has many more rows than pictured.

	Stage	Date	Distance	Origin	Destination	Type	Winner	Winner Country
0	1	2017-07-01	14.0	Düsseldorf	Düsseldorf	Individual time trial	Geraint Thomas	GBR
1	2	2017-07-02	203.5	Düsseldorf	Liège	Flat stage	Marcel Kittel	GER
2	3	2017-07-03	212.5	Verviers	Longwy	Medium mountain stage	Peter Sagan	SVK
3	4	2017-07-04	207.5	Mondorf-les-Bains	Vittel	Flat stage	Arnaud Démare	FRA
4	5	2017-07-05	160.5	Vittel	La Planche des Belles Filles	Medium mountain stage	Fabio Aru	ITA
5	6	2017-07-06	216.0	Vesoul	Troyes	Flat stage	Marcel Kittel	GER

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

Assume that we have already run `import babypandas as bpd` and `import numpy as np`.