permutation testing
merge
probability
how to
choose
test stat

Final Exam - DSC 10, Spring 2025

Full Name:

PID:

Seat you are in:

Instructions:

- This exam consists of 9 questions, worth a total of 117 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**.
- 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!

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.

 **The Hunger Games** 

The Hunger Games is a a young adult dystopian fictional novel. The events take place in the future in the fictional country of Panem, which consists of 12 impoverished districts and a wealthy metropolitan area called the Capitol.

The plot centers around an annual televised competition called the Hunger Games, in which children from the districts are forced to compete in a battle to the death. The participants, called tributes, are randomly selected via a lottery system.

The competition takes place in an arena that is specially designed for this purpose, and usually lasts several days or weeks, until only one tribute survives. The entire event is turned into a spectacle, which is broadcast throughout Panem as a way for the oppressive government to remind district residents of their powerlessness.

The main character and protagonist of *The Hunger Games* is Katniss Everdeen, a 16-year-old girl from District 12 who competes in the 74th annual Hunger Games. After a life of poverty and near starvation, her experience in the Hunger Games arena further fuels her hatred of the government and lights a fire in her to fight back against the oppression.

Notes:

- Throughout this exam, assume we have already run `import babypandas as bpd`, `import numpy as np`, and `import scipy`.
- At any point, feel free to use functions and variables that you defined in earlier subparts of the same question.

PID: _____

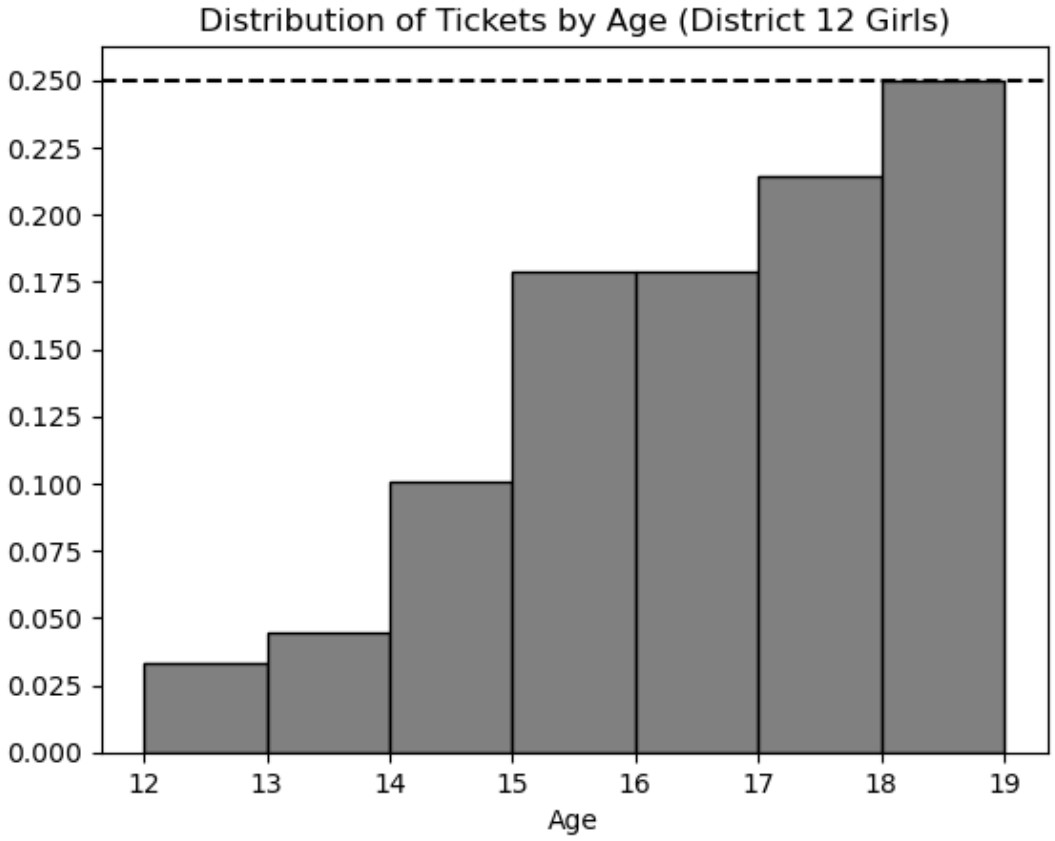
Question 1 (11 pts)

In an annual ceremony known as the reaping, tributes are selected to represent their district in the Hunger Games. One male and one female tribute from each district are randomly selected via a lottery drawing.

Every child between the ages of 12 and 18 (inclusive) has tickets entered into the drawing for their sex and district (e.g. girls from District 12). The number of tickets entered is dependent on age.

Starting at age 12, each child receives one ticket in the lottery. For each year after that, they receive one additional ticket, added to the total from the previous year. For example, 13-year-olds have two tickets, 14-year-olds have three tickets, and so on.

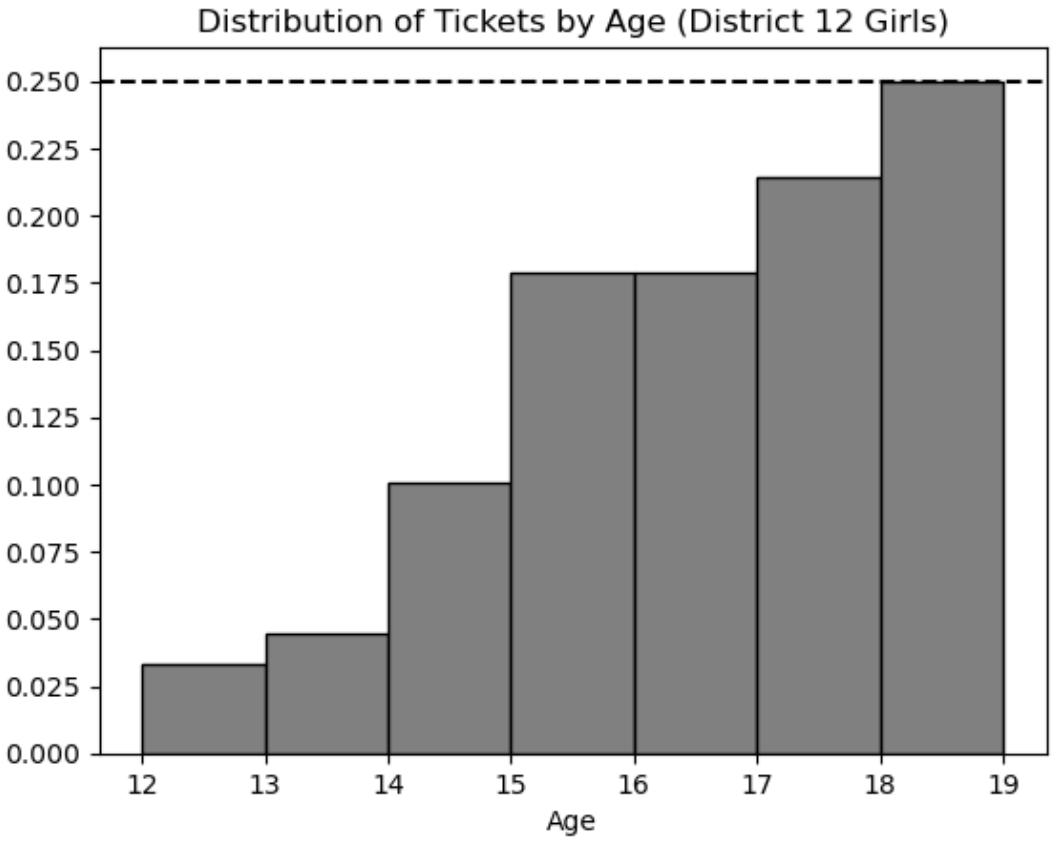
In this problem, we will consider only tickets corresponding to girls from District 12, and look at the distribution of these tickets according to the age of the person they represent. A density histogram for these tickets is shown below.



a) (3 pts) Which of the following statements about this distribution is correct?

- ◯ The mean is less than the median.
- ◯ The mean is the same as the median.
- ◯ The mean is greater than the median.
- ◯ It is impossible to determine the relationship between the mean and the median with the given information.

The histogram from the previous page is repeated below for your reference.



b) (4 pts) Suppose the rules of the Hunger Games were changed to eliminate 18-year-olds. If we plotted a new density histogram of the distribution of ages for tickets corresponding to girls from District 12 aged 12 to 17, how would the height of the `[13, 14)` bar change?

Let h be the height of this bar in the original histogram. Give its height in the new histogram in terms of h.

c) (4 pts) What is the **most common age** among girls from District 12 aged 12 to 18? Remember, the distribution above is for all tickets, and older girls have more tickets.

◯ 12 ◯ 13 ◯ 14 ◯ 15 ◯ 16 ◯ 17 ◯ 18

PID: _____

Question 2 (11 pts)

As we saw in the last problem, children aged 12 to 18 (inclusive) have tickets entered into a drawing at the reaping. 12-year-olds have one ticket, 13-year-olds have two tickets, 14-year-olds have three tickets, and so on, gaining one ticket per year of age.

In this problem, we'll look at the ages of all boys from District 3 and determine the probability that a boy of a certain age is selected in the drawing.

a) (4 pts) Suppose that there are only five boys from District 3 and their ages are as follows (in no particular order):

$$17, 12, 15, 14, 12.$$

Determine the probability that a 17-year-old is selected in the drawing.

Give your answer as an **unsimplified fraction** where the numerator is the number of tickets corresponding to a 17-year-old and the denominator is the total number of tickets.

b) (4 pts) Now, we'll solve the problem more generally. Fill in the blanks below to define a function `pick_prob` that takes as input an array containing the ages of all boys in District 3, and a single age between 12 and 18 (inclusive). The function should return the probability of randomly selecting a boy of that age during the reaping.

```
def pick_prob(ages, one_age):
    age_tickets = __(a)__
    total_tickets = __(b)__
    return age_tickets / total_tickets
```

`(a)`:

`(b)`:

c) (3 pts) Using `pick_prob`, write one line of code that evaluates to the probability that a 14-year-old boy is **not chosen** during the reaping if the boys in District 3 are aged 12, 14, 14, 15, 17, and 18.

Question 3 (12 pts)

So far, we have seen one way that children have tickets entered into the reaping: they receive one ticket when they are 12 years old, and then each year thereafter, an additional ticket is added onto the previous year's total. This means 13-year-olds have two tickets, 14-year-olds have three tickets, and so on. We'll call these tickets **age tickets**.

In this problem only, we'll consider another way that a child may *choose* to enter tickets into the reaping in addition to the mandatory age tickets. If a child wishes, they can guarantee food rations for their family members, including themselves, at the price of **one ticket per person**. We'll call these tickets **food tickets**. Like age tickets, food tickets are compounded each year, adding onto last year's total.

As an example, let's calculate the number of tickets that Katniss Everdeen has entered into the drawing at the reaping. Katniss is 16 years old, and every year, she has bought food for 3 family members (herself, her mother, and her sister Prim). This means:

- At age 12, Katniss had one age ticket and three food tickets, making 4 tickets total.
- At age 13, Katniss had one age ticket and three food tickets **in addition to** the 4 tickets from the year before, making 8 tickets total.
- At age 14, Katniss again had one age ticket and three food tickets **in addition to** the 8 tickets from the year before, making 12 tickets total.

This pattern continues, and by the time Katniss is 16, she has 20 tickets.

In other words, Katniss had 4 tickets entered when she was 12 years, and 4 more with each passing year. The array `np.arange(4, 24, 4)` contains the number of tickets Katniss entered each year, starting at age 12, up to and including her current age of 16 years old.

a) (5 pts) Fill in the blanks below to define the function `tix_array` which takes in a child's current `age` between 12 and 18 (inclusive) and a number of family members, `k`. The function returns an array similar Katniss's array above, representing the number of tickets they entered into the reaping each year since they were 12 years old, assuming that they buy food for their whole family every year.

Tip: `tix_array(16, 3)` should be the same as the array `np.arange(4, 24, 4)`.

```
def tix_array(age, k):
    return np.arange(__(a)__, __(b)__, __(c)__)
```

(a):

(b):

(c):

b) (4 pts) The DataFrame `reaping` contains information on the children of District 12 between the ages of 12 and 18. For each child, we have their `"name"`, `"age"`, `"family_size"` which includes themselves, and a boolean variable `"buying_food"`. A value of `True` means the child always buys food for their entire family, and `False` means the child never buys food for anyone. The first few rows of `reaping` are shown below, but there are many more rows than pictured.

	name	age	family_size	buying_food
0	Gale Hawthorne	18	5	True
1	Madge Undersee	16	3	False
2	Primrose Everdeen	12	3	False

Fill in the blanks in the code below to add a new column, `"tickets"`, to `reaping` that contains the number of tickets that the child will have entered into the drawing in the current year.

Hint: In Python, `True` is treated as 1 and `False` is treated as 0 when doing arithmetic!

```
tickets_per_year = __(d)__ * __(e)__ + 1
current_tickets = tickets_per_year * (__(f)__)
reaping = reaping.assign(tickets = current_tickets)
```

(d): _____

(e): _____

(f): _____

c) (3 pts) For this subpart, assume that the `tix_array` function was defined correctly in part (*a*), and that the `"tickets"` column was added correctly to the `reaping` DataFrame in part (*b*). Fill in the blanks in the code below so that the following expression evaluates to `True`.

```
reaping.get("tickets").iloc[7] == tix_array(__(g)__, __(h)__)[-1]
```

(g): _____

(h): _____

Question 4 (10 pts)

After being selected at the reaping, tributes are transported to the Capitol to prepare for the Hunger Games. While they are there, they attend a training camp to practice skills that might be helpful in the arena. At the training camp, there are 8 different stations such as camouflage, knife throwing, archery, plant identification, etc. At each of the 8 stations, tributes are scored on their skills from 1 to 10.

These 8 scores are combined into an overall score as follows:

- Count the number of stations at which the tribute scored **more than 5**, demonstrating basic proficiency.
- Count the number of stations at which the tribute scored **more than 8**, demonstrating expertise.
- Add these counts together, capping the overall score at 12. This means that if the sum is larger than 12, the tribute earns the maximum possible score of 12.

Overall scores therefore range from 0 to 12. Which of the following functions takes as input an array containing a tribute's 8 scores from the stations and correctly outputs their overall score? Select all that apply.

Hint: In Python, `True + True` evaluates to 2.

☐
```
def function1(stations):
    overall = 0
    for score in stations:
        if score > 5:
            overall = overall + 1
        if score > 8:
            overall = overall + 1
        if overall >= 12:
            return 12
    return overall
```

☐
```
def function2(stations):
    overall = 0
    for score in stations:
        if score > 5:
            overall = overall + 1
        elif score > 8:
            overall = overall + 2
    return min(overall, 12)
```

☐
```
def function3(stations):
    overall = 0
    for i in np.arange(8):
        if stations[i] > 8:
            overall = overall + 2
        elif stations[i] > 5:
            overall = overall + 1
    return min(overall, 12)
```

☐
```
def function4(stations):
    overall = 0
    for score in stations:
        add = score > 5
        add = (score > 8) + add
        overall = overall + add
    return min(overall, 12)
```

☐
```
def function5(stations):
    return min(12, np.count_nonzero(stations > 5) +
                   np.count_nonzero(stations > 8))
```

PID: _____

Question 5 (21 pts)

The night before the Hunger Games begins, each tribute is interviewed in front of a live audience. During this interview, the host asks each tribute a few personal questions and reveals their overall score from the training camp. These interviews are broadcast across the country, so that the residents of Panem can get to know the tributes better and form opinions about who they want to win.

The Capitol wants to understand public perceptions of the tributes after the interviews for the 74th Hunger Games. They conduct a survey of a sample of residents from all 12 districts, asking them two questions:

1. "What district do you live in?"
2. "Who do you think will win this year's Hunger Games?"

The survey results are in the DataFrame `survey`, with columns `"District"` and `"Tribute"` which contain each person's answers to the two questions above. The first few rows of `survey` are shown below.

	District	Tribute
0	7	Glimmer
1	3	Clove
2	1	Katniss
3	2	Clove
4	12	Peeta
5	10	Thresh

In this problem, we will try to estimate the proportion of residents from a given district who think a certain tribute will win the Hunger Games.

a) (4 pts) What proportion of residents in District 11 think Peeta will win? Write **one line of code** that evaluates to this proportion **in our sample**, based on the data in `survey`.

b) (5 pts) Next, we want to create a 95% confidence interval for the proportion of **all** residents from a given district who think a certain tribute will win. Fill in the blanks in the function `win_CI` below. This function takes the name of a tribute and the number of a district and returns the endpoints of a 95% bootstrapped confidence interval for the proportion of all residents of that district who think that tribute will win, based on the data in `survey`.

For example `win_ci("Peeta", 11)` returns the endpoints of a 95% confidence interval for the proportion of all residents from District 11 who think Peeta will win.

```
def win_ci(tribute, district):
    only_district = survey[survey.get("District") == district]
    props = np.array([])
    for i in np.arange(10000):
        resample = __(a)__
        tribute_count = __(b)__
        boot_prop = tribute_count / __(c)__
        props = np.append(props, boot_prop)
    return [np.percentile(props, 2.5), np.percentile(props, 97.5)]
```

(a):

(b):

(c):

c) (4 pts) Suppose we were to plot a histogram of `props` within the function `win_CI`. Which of the following best describes this histogram?

◯ The histogram reflects the shape of the population.

◯ The histogram reflects the shape of the data in `survey`.

◯ The histogram reflects the shape of the data in `survey` which corresponds to the given district.

◯ The histogram is roughly normal because of the Central Limit Theorem (CLT).

◯ The histogram is roughly normal, but not because of the CLT.

d) (4 pts) Suppose we now compute the following:

```
>>> win_ci("Katniss", 4)
[0.25, 0.72]

>>> win_ci("Katniss", 12)
[0.50, 0.70]
```

Which of the following reasons best explains why the second interval is narrower than the first?

◯ More people live in District 12 than District 4.
◯ More people live in District 4 than District 12.
◯ A greater fraction of District 12 residents than District 4 residents think Katniss will win.
◯ A greater fraction of District 4 residents than District 12 residents think Katniss will win.
◯ There are more survey participants from District 12 than District 4.
◯ There are more survey participants from District 4 than District 12.

e) (4 pts) Suppose we want to redo our survey so that our confidence interval for the proportion of District 12 residents who think Katniss will win has a width of at most 0.10. We will assume that our new sample's standard deviation will be the same as our original sample's standard deviation. Which of the following best describes how to achieve this?

◯ Our new sample should have twice as many people from District 12. It doesn't matter how many people the sample contains overall.
◯ Our new sample should have four times as many people from District 12. It doesn't matter how many people the sample contains overall.
◯ Our new sample should have twice as many people overall. It doesn't matter how many of them are from District 12.
◯ Our new sample should have four times as many people overall. It doesn't matter how many of them are from District 12.

Question 6 (14 pts)

Residents of Panem not participating in the Hunger Games can sponsor tributes to help them survive. Sponsors purchase supplies and have them delivered to tributes in the arena via parachute. Haymitch is the mentor for the tributes from District 12, Katniss and Peeta. Part of his job is to recruit sponsors to buy necessary supplies for Katniss and Peeta while they are in the arena.

In his advertising to potential sponsors, Haymitch claims that in 100 randomly selected parachutes delivered to tributes in past Hunger Games, the supplies were distributed into categories as follows:

Haymitch's Distribution

Category	Medical	Food	Weapons	Shelter	Other	**Total**
Count	40	30	10	15	5	100

Data scientists at the Capitol know that across all past Hunger Games, there have been 2000 parachute deliveries of supplies to tributes. Further, they have recorded the following count of how many of these parachutes' supplies fell into each category:

The Capitol's Distribution

Category	Medical	Food	Weapons	Shelter	Other	**Total**
Count	700	500	300	300	200	2000

In this problem, we will assess whether Haymitch is making an accurate claim by determining whether his sample of 100 parachutes looks like a random sample from the Capitol's distribution.

a) (5 pts) Which of the following are appropriate test statistics for this hypothesis test? Select all that apply.

- ☐ The mean difference in proportions between Haymitch's distribution and the Capitol's distribution.
- ☐ The maximum absolute difference in proportions between Haymitch's distribution and the Capitol's distribution.
- ☐ The average absolute difference in proportions between Haymitch's distribution and the Capitol's distribution.
- ☐ The sum of squared differences in proportions between Haymitch's distribution and the Capitol's distribution.
- ☐ The correlation coefficient between Haymitch's proportion and the Capitol's proportion, for each category.
- ☐ None of the above.

PID: _____

The distributions from the previous page are repeated below for your reference.

Haymitch's Distribution

Category	Medical	Food	Weapons	Shelter	Other	**Total**
Count	40	30	10	15	5	100

The Capitol's Distribution

Category	Medical	Food	Weapons	Shelter	Other	**Total**
Count	700	500	300	300	200	2000

b) (4 pts) Suppose we decide to use total variation distance (TVD) as the test statistic for this hypothesis test. Calculate the TVD between Haymitch's distribution and the Capitol's distribution. Give your answer as an exact decimal.

$0.2/2 =$ **0.1**

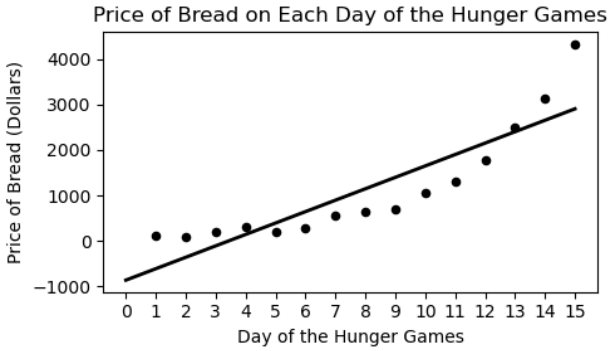
9am

c) (5 pts) Suppose you run a simulation to generate 1000 TVDs between the Capitol's distribution and samples of size 100 randomly drawn from that distribution. You determine that the TVD you calculated in part (*b*) is in the 96th percentile of your simulated TVDs. Which of the following correctly interprets this result? Select all that apply.

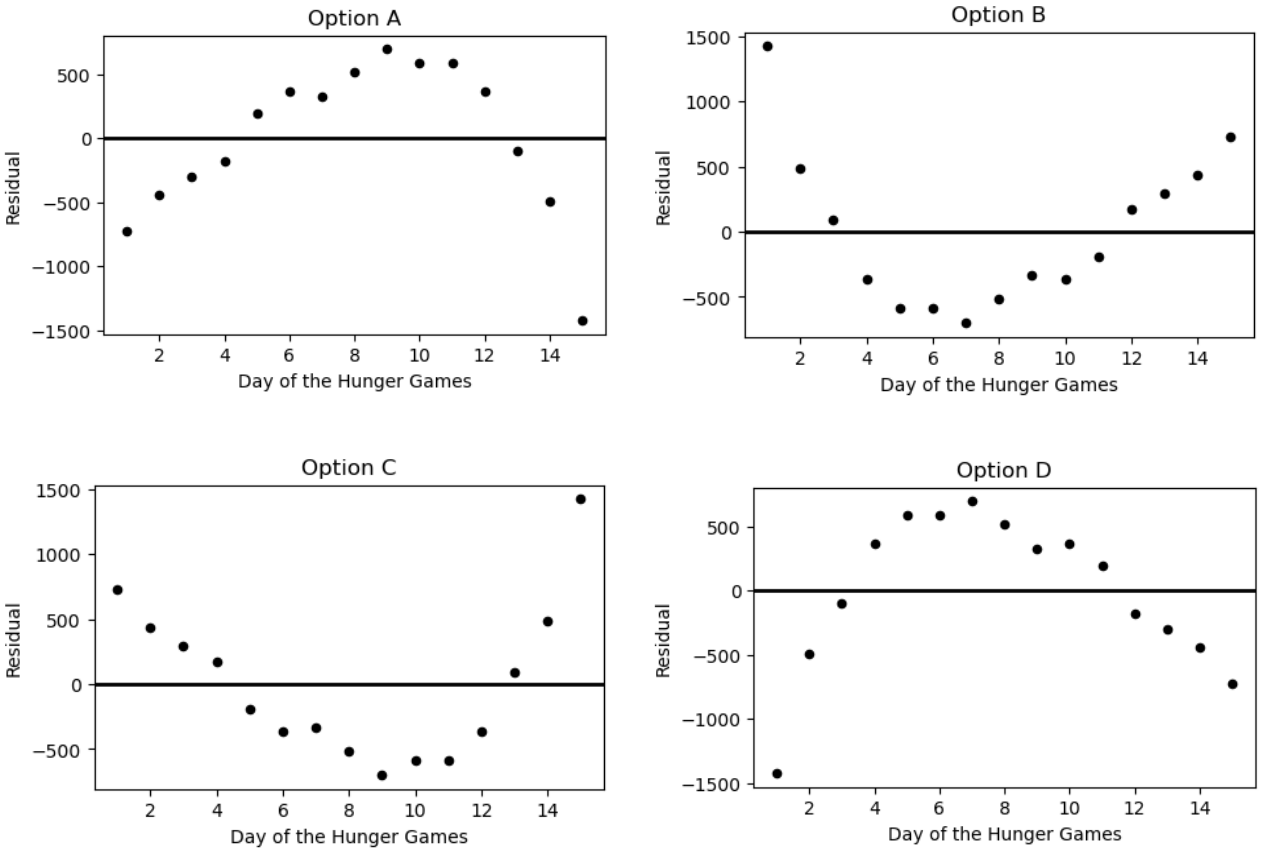
- ☐ About 96% of simulated TVDs are less than the one you calculated in part (*b*).
- ☐ Haymitch's distribution is 96% more accurate than the Capitol's distribution.
- ☐ There is a 96% probability that Haymitch's sample was a random sample from the Capitol's distribution.
- ☐ If Haymitch's sample were a random sample from the Capitol's distribution, getting a TVD greater than or equal to the one you calculated in part (*b*) would happen about 4% of the time.
- ☐ If Haymitch's sample were a random sample from the Capitol's distribution, getting a TVD greater than or equal to the one you calculated in part (*b*) would happen about 96% of the time.
- ☐ None of the above.

Question 7 (12 pts)

During the Hunger Games, sponsors may purchase supplies for the tributes competing. However, supplies increase in price as the Hunger Games progress. Haymitch collects data on the price, in dollars, of purchasing bread for a tribute over the first 15 days of the Hunger Games competition. He uses linear regression to predict the price of bread based on the day of the competition. His regression line is shown below on a scatterplot of the data.



a) (3 pts) Which of the following plots is the residual plot for the data above?



◯ Option A ◯ Option B ◯ Option C ◯ Option D

b) (3 pts) What conclusions can Haymitch draw from looking at the residual plot of his regression line? Select all that apply.

- ☐ The correlation coefficient between these variables is weak ($r < 0.5$).
- ☐ A line is not the best choice to model the relationship between these variables.
- ☐ There is a different line that fits the data better than this line.
- ☐ None of the above.

c) (6 pts) Haymitch wants to present the scatter plot to potential sponsors, but he wants to give the prices in thousands of dollars instead of dollars. He divides each of the 15 prices in his data set by 1000, then recalculates the regression line. Which of the following statements are correct? Select all that apply.

- ☐ The mean of the prices will be divided by 1000.
- ☐ The standard deviation of the prices will be divided by 1000.
- ☐ The slope of the regression line predicting price from day will be divided by 1000.
- ☐ The intercept of the regression line predicting price from day will be divided by 1000.
- ☐ The slope of the regression line predicting price in standard units from day in standard units will be divided by 1000.
- ☐ The root mean square error (RMSE) of the regression line will be divided by 1000.
- ☐ None of the above.

Question 8 (13 pts)

In certain districts (1, 2, and 4), children spend years training for the Hunger Games and frequently volunteer to participate in them. Tributes that come from these districts are known as **Career tributes**. Many residents of Panem believe that Career tributes generally fare better in the Hunger Games because of their extensive training.

We'll test this claim using historical data. The DataFrame `survival` has a row for each tribute who participated in one of the first 74 Hunger Games. The columns are as follows:

- `"Tribute"`: The name of the tribute.
- `"District"`: Their home district (1–12).
- `"Days"`: The number of days they stayed alive in the arena.
- `"Game"`: The Hunger Games edition they competed in (1–74).

A few rows of `survival` are shown below:

Tribute	District	Days	Game
Jessup Diggs	12	2	10
Glimmer Belcourt	1	5	74
Rue Stenberg	11	9	74

We'll use this data to test the following pair of hypotheses:

- **Null Hypothesis**: On average, Career tributes and non-Career tributes survive an equal amount of time in the arena.
- **Alternative Hypothesis**: On average, Career tributes survive **longer** in the arena than non-Career tributes.

Our test statistic will be the mean survival time of Career tributes minus the mean survival time of non-Career tributes

PID: _____

a) (5 pts) Write code to create a DataFrame called `tributes` that has all the data in `survival` plus an additional column called `"Career"`. This column should contain boolean values indicating whether each tribute is considered a Career tribute. Feel free to define intermediate variables and functions as needed, and to do this in multiple lines of code.

districts 1, 2, 4

b) (4 pts) Fill in the blanks in the code below so that `statistics` evaluates to an array with 10000 simulated values of the test statistic under the null hypothesis.

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

for i in np.arange(10000):
    shuffled = tributes.assign(__(a)__)
    means = shuffled.groupby("Career").mean().get("Days")
    stat = __(b)__
    statistics = np.append(statistics, stat)
```

add shuffled column

mean me ... C − N

mean mean T − F

(a): Days = np.random.permutation(tributes.get('Days'))

(b): means.loc[True] − means.loc[False]

c) (4 pts) The output of `tributes.groupby("Career").mean()` is shown below.

original data ⇒ info to calculate obs stat

Career	District	Days	Game
False	7.89	4.6	37.5
True	2.33	5.2	37.5

= 5.2 − 4.6 = 0.6

Fill in the blank below to compute the p-value of this test.

```
p_value = (__(a)__).mean()
```

(a): statistics >= 0.6

p-val

Series of T/F values

= 0.6 ← np.count_nonzero(statistics >= 0.6) / 10000

Question 9 (13 pts)

Before getting selected for the Hunger Games, Katniss often spent her days hunting with her friend Gale. Hunting is illegal in Panem, so Katniss and Gale sold their poached game at a black market known as The Hob, always splitting the profits equally, even if one person's kills were worth more than the other's.

a) (6 pts) Suppose Katniss and Gale hunted together three times. The values of each person's kills are recorded in `katniss_sample` and `gale_sample`, respectively. Their individual profits are recorded in `average_sample`. Calculate the mean and variance of each of these three samples. Give all of your answers as integers.

Sample	**Mean**	**Variance**
`katniss_sample = [47, 44, 50]`		
`gale_sample = [25, 28, 28]`		
`average_sample = [36, 36, 39]`		

b) (4 pts) Suppose that the value of Katniss's kills are normally distributed with mean \$50 and SD \$3, and the value of Gale's kills are independently normally distributed with mean \$30 and SD \$2.

On one hunting trip, Katniss's kills are worth twice as much as Gale's, but the value of her kills in standard units is the same as the value of Gale's kills in standard units. Determine the value of Gale's kills on this hunting trip. Give your answer as an integer.

c) (3 pts) Now, suppose that we no longer know whether the distribution of the value of Katniss's kills is normal. All we know about this distribution is that it has mean \$50 and SD \$3.

Which of the following statements are true? Select all that apply.

- ☐ It is possible that Katniss's kills are never valued between \$48 and \$52.
- ☐ No more than 75% of Katniss's kills are between \$44 and \$56 in value.
- ☐ `scipy.stats.norm.cdf(50)` gives an approximation for the fraction of Katniss's kills that are below \$50 in value.
- ☐ None of these.