Final Exam - DSC 10, Fall 2024

Full Name:

PID:

Seat you are in:

Instructions:

- This exam consists of 11 questions, worth a total of 180 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.

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

Question 1 (8 pts)

a) (3 pts) Notice that `bookstore` has an index of `"ISBN"` and `sales` does not. Why is that?

◯ There is no good reason. We could have set the index of `sales` to `"ISBN"`.

◯ There can be two different books with the same `"ISBN"`.

◯ `"ISBN"` is already being used as the index of `bookstore`, so it shouldn't also be used as the index of `sales`.

◯ The bookstore can sell multiple copies of the same book.

b) (2 pts) Is `"ISBN"` a numerical or categorical variable?

◯ numerical ◯ categorical

c) (3 pts) Which type of data visualization should be used to compare authors by median rating?

◯ scatter plot ◯ line plot ◯ bar chart ◯ histogram

Question 2 (8 pts)

ISBN numbers for books must follow a very particular format. The first 12 digits encode a lot of information about a book, including the publisher, edition, and physical properties like page count. The 13th and final digit, however, is computed directly from the first 12 digits and is used to verify that a given ISBN number is valid.

In this question, we will define a function to compute the final digit of an ISBN number, given the first 12. First, we need to understand how the last digit is determined. To find the last digit of an ISBN number:

1. Multiply the first digit by 1, the second digit by 3, the third digit by 1, the fourth digit by 3, and so on, alternating 1s and 3s.
2. Add up these products.
3. Find the ones digit of this sum.
4. If the ones digit of the sum is zero, the final digit of the ISBN is zero. Otherwise, the final digit of the ISBN is ten minus the ones digit of the sum.

For example, suppose we have an ISBN number whose first 12 digits are 978186197271. Then the last digit would be calculated as follows:

$$9 \cdot 1 + 7 \cdot 3 + 8 \cdot 1 + 1 \cdot 3 + 8 \cdot 1 + 6 \cdot 3 + 1 \cdot 1 + 9 \cdot 3 + 7 \cdot 1 + 2 \cdot 3 + 7 \cdot 1 + 1 \cdot 3 = 118$$

The ones digit is 8 so the final digit is 2.

Fill in the blanks in the function `calculate_last_digit` below. This function takes as input a **string** representing the first 12 digits of the ISBN number, and should return the final digit of the ISBN number, as an `int`. For example, `calculate_last_digit("978186197271")` should evaluate to `2`.

```
def calculate_last_digit(partial_isbn_string):
    total = 0
    for i in __(a)__:
        digit = int(partial_isbn_string[i])
        if __(b)__:
            total = total + digit
        else:
            total = total + 3 * digit
    ones_digit = __(c)__
    if __(d)__:
        return 10 - ones_digit
    return 0
```

(a):

(b):

(c):

(d):

Note: this procedure may seem complicated, but it is actually how the last digit of an ISBN number is generated in real life!

Question 3 (18 pts)

At the beginning of 2025, avid reader Michelle will join the Blind Date with a Book program sponsored by Bill's Book Bonanza. Once a month from January through December, Michelle will receive a surprise book in the mail from the bookstore. So she'll get 12 books throughout the year.

Each book is randomly chosen among those available at the bookstore (or equivalently, from among those in the `bookstore` DataFrame). Each book is equally likely to be chosen each month, and each month's book is chosen independently of all others, so repeats can occur.

Let p be the proportion of books in `bookstore` from the romance genre ($0 \leq p \leq 1$).

a) (3 pts) What is the probability that the books Michelle receives in February and December are romance, while the books she receives all other months are non-romance? Your answer should be an unsimplified expression in terms of p.

b) (3 pts) What is the probability that at least one of the books Michelle receives in the last four months of the year is romance? Your answer should be an unsimplified expression in terms of p.

c) (3 pts) It is likely Michelle will receive at least one romance book and at least one non-romance book throughout the 12 months of the year. What is the probability that this does not happen? Your answer should be an unsimplified expression in terms of p.

d) (3 pts) Let n be the total number of rows in `bookstore`. What is the probability that Michelle does not receive any duplicate books in the first four months of the year? Your answer should be an unsimplified expression in terms of n.

Now, consider the following lines of code which define variables `i`, `j`, and `k`.

```
def foo(x, y):
    if x == "rating":
        z = bookstore[bookstore.get(x) > y]
    elif x == "genre":
        z = bookstore[bookstore.get(x) == y]
    return z.shape[0]

i = foo("rating", 3)
j = foo("rating", 4)
k = foo("genre", "Romance")
```

For both questions that follow, your answer should be an unsimplified expression in terms of `i`, `j`, `k` only. If you do not have enough information to determine the answer, leave the answer box blank and instead fill in the bubble.

e) (3 pts) If we know in advance that Michelle's January book will have a rating greater than 3, what is the probability that the book's rating is greater than 4?

◯ Not enough information.

f) (3 pts) If we know in advance that Michelle's January book will have a rating greater than 3, what is the probability that the book's genre is romance?

◯ Not enough information.

Question 4 (22 pts)

Matilda has been working at Bill's Book Bonanza since it first opened and her shifts for each week are always randomly scheduled (i.e., she does not work the same shifts each week). Due to the system restrictions at the bookstore, when Matilda logs in with her employee ID, she only has access to the history of sales made during her shifts. Suppose these transactions are stored in a DataFrame called `matilda`, which has the same columns as the `sales` DataFrame that stores all transactions.

Matilda wants to use her random sample to estimate **the mean price of books purchased with cash** at Bill's Book Bonanza. For the purposes of this question, assume Matilda only has access to `matilda`, and not all of `sales`.

a) (8 pts) Complete the code below so that `cash_left` and `cash_right` store the endpoints of an 86% bootstrapped confidence interval for the mean price of books purchased with cash at Bill's Book Bonanza.

```
cash_means = np.array([])
original = __(a)__
for i in np.arange(10000):
    resample = original.sample(__(b)__)
    cash_means = np.append(cash_means, __(c)__)

cash_left = __(d)__
cash_right = __(e)__
```

(a):

(b):

(c):

(d):

(e):

b) (4 pts) Next, Matilda uses the data in `matilda` to construct a 95% CLT-based confidence interval for the same parameter.

Given that there are 400 cash transactions in `matilda` and her confidence interval comes out as $[19.58, 21.58]$, what is the standard deviation of the prices of all cash transactions at Bill's Book Bonanza?

SD = ☐

c) (4 pts) Knowing the endpoints of Matilda's 95% CLT-based confidence interval can actually help us to determine the endpoints of her 86% bootstrapped confidence interval. You may also need to know the following facts:

- `stats.norm.cdf(1.1)` evaluates to `0.86`
- `stats.norm.cdf(1.5)` evaluates to `0.93`

Estimate the value of `cash_left` to one decimal place.

`cash_left` = ☐

d) (3 pts) Which of the following are valid conclusions? **Select all that apply.**

☐ Approximately 95% of the values in `cash_means` fall within the interval $[19.58, 21.58]$.

☐ Approximately 95% of books purchased with cash at Bill's Book Bonanza have a price that falls within the interval $[19.58, 21.58]$.

☐ The actual mean price of books purchased with cash at Bill's Book Bonanza has a 95% chance of falling within the interval $[19.58, 21.58]$.

☐ None of the above.

e) (3 pts) Matilda has been wondering whether the mean price of books purchased with cash at Bill's Book Bonanza is \$20. What can she conclude about this?

◯ The mean price of books purchased with cash at Bill's Book Bonanza is \$20.

◯ The mean price of books purchased with cash at Bill's Book Bonanza could plausibly be \$20.

◯ The mean price of books purchased with cash at Bill's Book Bonanza is not \$20.

◯ The mean price of books purchased with cash at Bill's Book Bonanza is most likely not \$20.

Question 5 (29 pts)

Aathi is a movie enthusiast looking to dive into book-reading. He visits Bill's Book Bonanza and asks for help from the employee on duty, Matilda.

Matilda has access to the `matilda` DataFrame of transactions from her shifts, as well as the `bookstore` DataFrame of all books available in the store. She decides to advise Aathi based on the ratings of books that are popular enough to have been purchased during her shifts. Unfortunately, `matilda` does not contain information about the rating or genre of the books sold. To fix this, she merges the two DataFrames and keeps only the columns that she'll need.

a) (3 pts) Fill in the blanks in the code below so that the resulting `merged` DataFrame has one row for each book in `matilda`, and columns representing the genre and rating of each such book.

```
merged = (matilda.merge(bookstore, __(a)__, __(b)__)
                 .get(["genre", "rating"]))
```

a: [ ] b: [ ]

b) (3 pts) In this example, `matilda` and `merged` have the same number of rows. Why is this the case? Choose the answer which is sufficient alone to guarantee the same number of rows.

◯ Because every book in `bookstore` appears exactly once in `matilda`.

◯ Because every book in `matilda` appears exactly once in `bookstore`.

◯ Because `matilda` has no duplicate rows.

◯ Because `bookstore` has no duplicate rows.

◯ Because the books in `matilda` are a subset of the books in `bookstore`.

Matilda uses some `babypandas` operations on the `merged` DataFrame to create the DataFrame pictured below, which she calls `top_two`.

	rating
genre	
Fantasy	4.65
Mystery	4.53

This DataFrame shows the two genres in `merged` with the highest mean rating. Note that the `"rating"` column in `top_two` represents a mean rating across many books of the same genre, not the rating of any individual book.

c) (5 pts) Write one line of code to define the DataFrame `top_two` as described above. It's okay if you need to write your answer on multiple lines, as long as it represents one line of code.

Based on the data in `top_two`, Matilda recommends that Aathi purchase a fantasy book. However, Aathi is skeptical because he recognizes that the data in `merged` is only a sample from the larger population of all transactions at Bill's Book Bonanza. Before he makes his purchase, he decides to do a permutation test to determine whether fantasy books have higher ratings than mystery books in this larger population.

d) (2 pts) Select the best statement of the null hypothesis for this permutation test.

◯ Among all books sold at Bill's Book Bonanza, fantasy books have a higher rating than mystery books, on average.

◯ Among all books sold at Bill's Book Bonanza, fantasy books do not have a higher rating than mystery books, on average.

◯ Among all books sold at Bill's Book Bonanza, fantasy books have the same rating as mystery books, on average.

e) (2 pts) Select the best statement of the alternative hypothesis for this permutation test.

◯ Among all books sold at Bill's Book Bonanza, fantasy books have a higher rating than mystery books, on average.

◯ Among all books sold at Bill's Book Bonanza, fantasy books do not have a higher rating than mystery books, on average.

◯ Among all books sold at Bill's Book Bonanza, fantasy books have the same rating as mystery books, on average.

f) (3 pts) Aathi decides to use the mean rating for mystery books minus the mean rating for fantasy books as his test statistic. What is his observed value of this statistic?

g) (3 pts) Which of the following best describes how Aathi will interpret the results of his permutation test?

◯ High values of the observed statistic will make him lean towards the alternative hypothesis.

◯ Low values of the observed statistic will make him lean towards the alternative hypothesis.

◯ Both high and low values of the observed statistic will make him lean towards the alternative hypothesis.

h) (5 pts) Fill in the blanks in the following code to perform Aathi's permutation test.

```
just_two = merged[(merged.get("genre") == "Fantasy") |
                  (merged.get("genre") == "Mystery")]

rating_stats = np.array([])
for i in np.arange(10000):
    shuffled = just_two.assign(shuffled = __(a)__)
    grouped = shuffled.groupby("genre").mean().get("shuffled")
    mystery_mean = grouped.iloc[__(b)__]
    fantasy_mean = grouped.iloc[__(c)__]
    rating_stats = np.append(rating_stats, mystery_mean - fantasy_mean)
```

(a):

(b):

(c):

i) (3 pts) Aathi gets a `p_value` of 0.008, and he will base his purchase on this result, using the standard p-value cutoff of 0.05. What will Aathi end up doing?

◯ Aathi fails to reject the null hypothesis and will purchase a fantasy book.

◯ Aathi fails to reject the null hypothesis and will purchase a mystery book.

◯ Aathi rejects the null hypothesis and will purchase a fantasy book.

◯ Aathi rejects the null hypothesis and will purchase a mystery book.

PID: _____

Question 6 (9 pts)

Bill is curious about whether his bookstore is busier on weekends (Saturday, Sunday) or weekdays (Monday, Tuesday, Wednesday, Thursday, Friday). To help figure this out, he wants to define some helpful one-line functions.

The function `find_day` should take in a given `"date"` and return the associated day of the week. For example, `find_day("Saturday, December 7, 2024")` should evaluate to `"Saturday"`.

The function `is_weekend` should take in a given `"date"` and return `True` if that date is on a Saturday or Sunday, and `False` otherwise. For example, `is_weekend("Saturday, December 7, 2024")` should evaluate to `True`.

a) (4 pts) Complete the implementation of both functions below.

```
def find_day(date):
    return date.__(a)__

def weekend(date):
    return find_day(date)[__(b)__] == __(c)__
```

(a): _____

(b): _____

(c): _____

Now, Bill runs the following line of code:

```
sales_day = sales.assign(weekend = sales.get("date").apply(is_weekend))
```

b) (5 pts) Determine which of the following code snippets evaluates to the proportion of book purchases in `sales` that were made on a weekend. **Select all that apply.**

- ☐ `sales_day[sales_day.get("weekend")].count() / sales_day.shape[0]`
- ☐ `sales[sales_day.get("weekend")].shape[0] / sales_day.shape[0]`
- ☐ `sales_day.get("weekend").median()`
- ☐ `sales_day.get("weekend").mean()`
- ☐ `np.count_nonzero(sales_day.get("weekend") > 0.5) / sales_day.shape[0]`

Question 7 (23 pts)

Hargen is an employee at Bill's Book Bonanza who tends to work weekend shifts. He thinks that Fridays, Saturdays, and Sundays are busier than other days, and he proposes the following probability distribution of sales by day:

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
0.2	0.1	0.1	0.1	0.1	0.2	0.2

Let's use the data in `sales` to determine whether Hargen's proposed model could be correct by doing a hypothesis test. The hypotheses are:

- **Null Hypothesis**: Sales at the bookstore are randomly drawn from Hargen's proposed distribution of days.
- **Alternative Hypothesis**: Sales at the bookstore are not randomly drawn from Hargen's proposed distribution of days.

a) (4 pts) Which of the following test statistics could be used to test the given hypothesis? **Select all that apply.**

- ☐ The absolute difference between the proportion of books sold on Saturday and the proposed proportion of books sold on Saturday (0.2).
- ☐ The sum of the differences in proportions between the distribution of books sold by day and Hargen's proposed distribution.
- ☐ The sum of the squared differences in proportions between the distribution of books sold by day and Hargen's proposed distribution.
- ☐ One half of the sum of the absolute differences in proportions between the distribution of books sold by day and Hargen's proposed distribution.

We will use as our test statistic **the mean of the absolute differences in proportions between the distribution of books sold by day and Hargen's proposed distribution.**

b) (3 pts) Suppose the observed distribution of books sold by day was as follows. Calculate the observed statistic in this case.

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
0.34	0.13	0.06	0.07	0.08	0.08	0.24

observed statistic = ☐

c) (7 pts) Let's determine the actual value of the observed statistic based on the data in `sales`. Assume that we have already defined a function called `find_day` that returns the day of the week for a given `"date"`. For example, `find_day("Saturday, December 7, 2024")` evaluates to `"Saturday"`. Fill in the blanks below so that the variable `obs` evaluates to the observed statistic.

```
# in alphabetical order: Fri, Mon, Sat, Sun, Thurs, Tues, Wed
hargen = np.array([0.2, 0.1, 0.2, 0.2, 0.1, 0.1, 0.1])
prop = (sales.assign(day_of_week = __(a)__)
             .groupby(__(b)__).__(c)__.get("ISBN") / sales.shape[0])
obs = __(d)__
```

(a): _____

(b): _____ (c): _____

(d): _____

d) (3 pts) To conduct the hypothesis test, we'll need to generate thousands of simulated day-of-the-week distributions. What will we do with all these simulated distributions?

◯ Use them to determine whether Hargen's proposed distribution looks like a typical simulated distribution.

◯ Use them to determine whether the observed distribution of books sold by day looks like a typical simulated distribution.

◯ Use them to determine whether Hargen's proposed distribution looks like a typical observed distribution of books sold by day.

e) (3 pts) In each iteration of the simulation, we need to collect a sample of size `sales.shape[0]` from which distribution?

◯ Hargen's proposed distribution.

◯ The distribution of data in `sales` by day of the week.

◯ Our original sample's distribution.

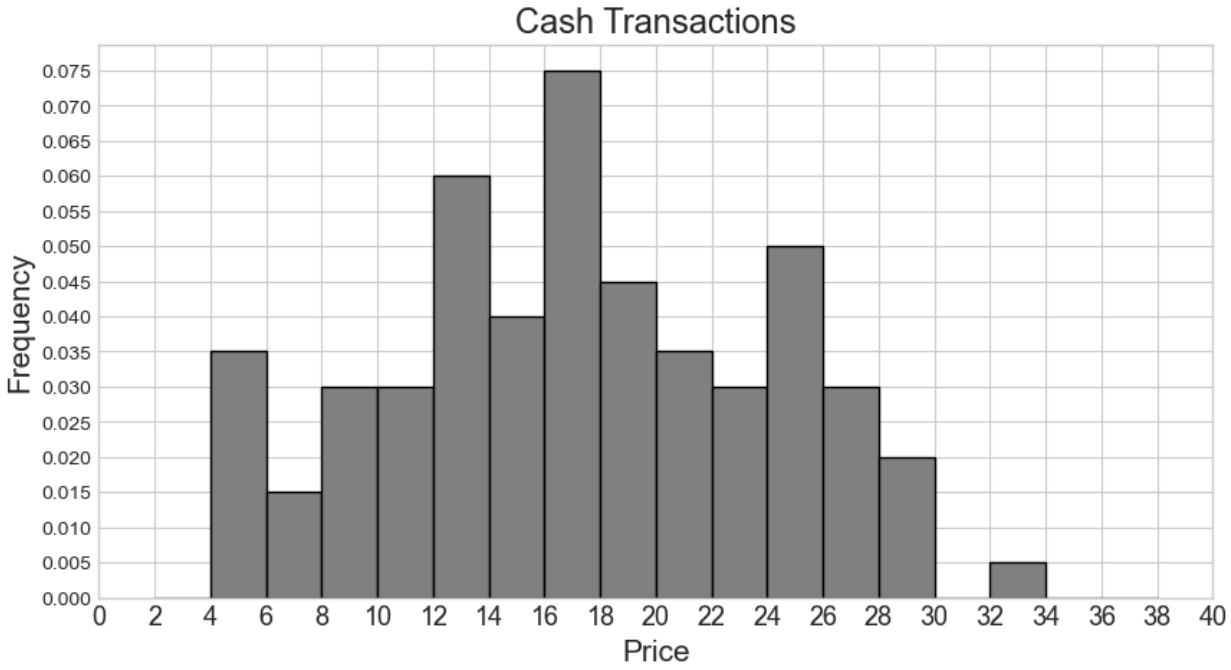
◯ The distribution of possible sample means.

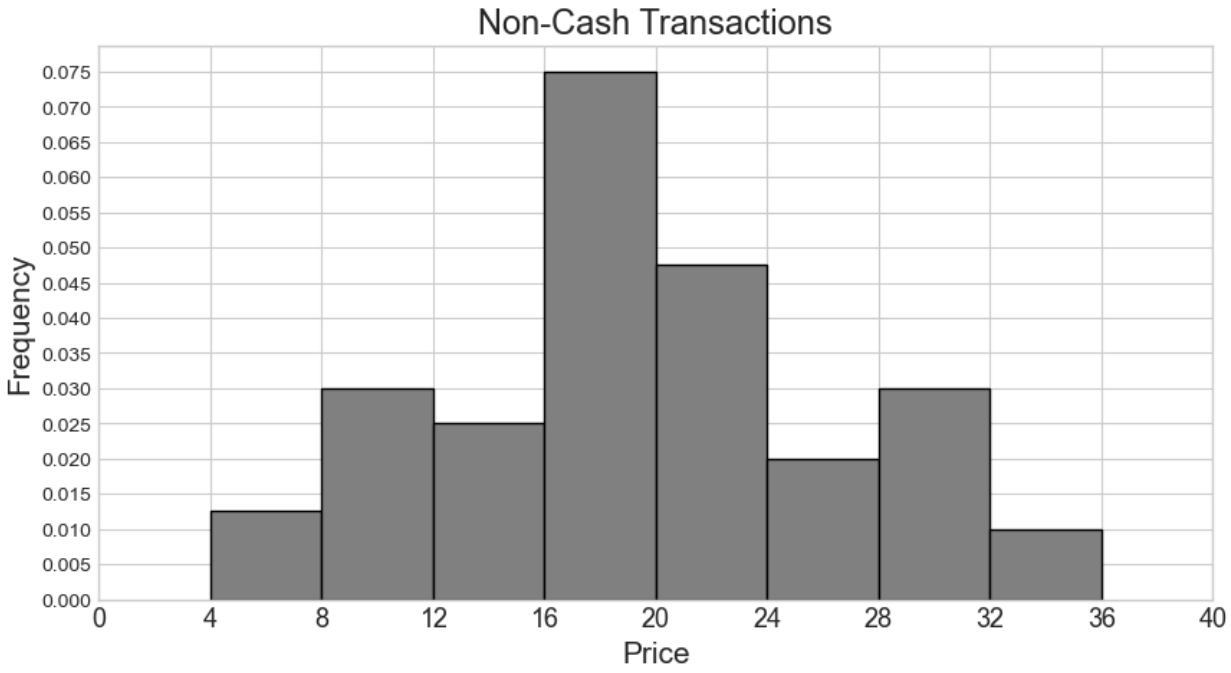
f) (3 pts) Suppose that `obs` comes out to be in the 98th percentile of simulated test statistics. Using the standard p-value cutoff of 0.05, what would Hargen conclude about his original model?

◯ It is likely correct. ◯ It is plausible. ◯ It is likely wrong.

Question 8 (16 pts)

Suppose we are told that `sales` contains 1000 rows, 500 of which represent cash transactions and 500 of which represent non-cash transactions. We are given two density histograms representing the distribution of `"price"` separately for the cash and non-cash transactions.





From these two histograms, we'd like to create a single combined histogram that shows the overall distribution of `"price"` for all 1000 rows in `sales`.

PID: _____

a) (4 pts) How many cash transactions had a `"price"` of \$26 or more?

b) (4 pts) Say our combined histogram has a bin $[4, 6)$. What will the height of this bar be in the combined histogram? Fill in the answer box or select "Not enough information."

height = ☐ ◯ Not enough information.

c) (4 pts) Say our combined histogram has a bin $[24, 28)$. What will the height of this bar be in the combined histogram? Fill in the answer box or select "Not enough information."

height = ☐ ◯ Not enough information.

d) (4 pts) Suppose now that the histograms for cash and non-cash transactions are the same as pictured, but they now represent 600 cash transactions and 400 non-cash transactions. In this situation, say our combined histogram has a bin $[12, 16)$. What will the height of this bar be in the combined histogram? Fill in the answer box or select "Not enough information."

height = ☐ ◯ Not enough information.

Question 9 (18 pts)

As in the previous problem, suppose we are told that `sales` contains 1000 rows, 500 of which represent cash transactions and 500 of which represent non-cash transactions.

This time, instead of being given histograms, we are told that the distribution of `"price"` for cash transactions is roughly normal, with a mean of \$14 and a standard deviation of \$2. We'll call this distribution the cash curve.

Additionally, the distribution of `"price"` for non-cash transactions is roughly normal, with a mean of \$22 and a standard deviation of \$4. We'll call this distribution the non-cash curve.

We want to draw a curve representing the approximate distribution of `"price"` for all transactions combined. We'll call this distribution the combined curve.

a) (4 pts) What is the approximate proportion of area under the **cash curve** between \$10 and \$12? Your answer should be a number between 0 and 1.

b) (4 pts) Fill in the blanks in the code below so that the expression evaluates to the approximate proportion of area under the **cash curve** between \$14.50 and \$17.50. Each answer should be a single number.

```
scipy.stats.norm.cdf(a) - scipy.stats.norm.cdf(b)
```

`a =` ____ `b =` ____

c) (3 pts) Will the combined curve be roughly normal?

◯ Yes, because of the Central Limit Theorem.

◯ Yes, because combining two normal distributions always produces a normal distribution.

◯ No, not necessarily.

d) (4 pts) Fill in the blanks in the code below so that the expression evaluates to the approximate proportion of area under the **combined curve** between \$14 and \$22. Each answer should be a single number.

```
(scipy.stats.norm.cdf(a) - scipy.stats.norm.cdf(b)) / 2
```

`a =` ____ `b =` ____

e) (3 pts) What is the approximate proportion of area under the **combined curve** between \$14 and \$22? Choose the closest answer below.

◯ 0.47 ◯ 0.49 ◯ 0.5 ◯ 0.95 ◯ 0.97

PID: _____

Question 10 (8 pts)

a) (4 pts) Suppose the `bookstore` DataFrame has 10 unique genres, and we are given a sample of 350 books from that DataFrame. Determine the maximum possible total variation distance (TVD) that could occur between our sample's genre distribution and the uniform distribution where each genre occurs with equal probability. Your answer should be a single number.

max possible TVD = $\boxed{}$

b) (2 pts) True or False: If the sample instead had 700 books, then the maximum possible TVD would **increase**.

◯ True ◯ False

c) (2 pts) True or False: If the `bookstore` DataFrame had 11 genres instead of 10, the maximum possible TVD would **increase**.

◯ True ◯ False

Question 11 (21 pts)

Dhruv works at the bookstore, and his job involves pricing new books that come in from the supplier. He prices new books based on the number of pages they have. He does this using linear regression, which he learned about in DSC 10.

To build his regression line, Dhruv gathers the following information about the distinct books currently available at the bookstore:

- The correlation between price and number of pages is 0.6.
- The mean price of all books is \$15, with a standard deviation of \$4.
- The mean number of pages of all books is 500, with a standard deviation of 200.

a) (6 pts) Which of the following statements about Dhruv's regression line are true? **Select all that apply.**

☐ It goes through the point $(500, 15)$.

☐ It goes through the point $(200, 4)$.

☐ Its slope is equal to 0.6.

☐ Its y-intercept is equal to 9.

☐ Its root mean square error is larger than the root mean square error of any other line.

☐ All the books currently available at the bookstore fall on the line.

b) (4 pts) If `"The Martian"` has 30 more pages than `"The Simple Wild"`, and both books are priced according to the regression line, how much more does `"The Martian"` cost than `"The Simple Wild"`? Give your answer as a number, in dollars and cents.

c) (4 pts) A new book added to the inventory is `"The Goldfinch"`, which has 700 pages. How much should Dhruv charge customers for this book, according to the regression line pricing model? Give your answer as a number, in dollars and cents.

d) (3 pts) It turns out that Dhruv had an error in his regression line because he had accidentally recorded the price of one book in the data set, `"Roadside Picnic"`, as $-\$12$ instead of \$12. He builds a new regression line using the correct price for `"Roadside Picnic"` and he finds that his new regression line has a smaller slope than before. What can we conclude about the number of pages in `"Roadside Picnic"` based on this information alone?

◯ `"Roadside Picnic"` has fewer than 500 pages.

◯ `"Roadside Picnic"` has exactly 500 pages.

◯ `"Roadside Picnic"` has more than 500 pages.

◯ Not enough information.

e) (2 pts) Suppose that Dhruv originally based his regression line on a data set which has a single row for each unique book sold at Bill's Book Bonanza. If instead, he had used a dataset with one row for each copy of a book at the bookstore (and there are multiple copies of some books), would his regression line have come out the same?

◯ Yes ◯ No

f) (2 pts) Suppose Dhruv bootstraps his scatterplot 10,000 times and calculates a regression line for each resample. It turns out that 95% of his bootstrapped slopes fall in the interval $[a, b]$ and 95% of his bootstrapped intercepts fall in the interval $[c, d]$. Does this mean that 95% of his predicted prices for a book with 500 pages fall in the interval $[500a + c, 500b + d]$?

◯ Yes ◯ No

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

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

Congratulations on finishing DSC 10!