

Lecture 26 – Final Review

DSC 10, Winter 2024

Announcements

- Lab 7 is due **tomorrow at 11:59PM**.
- The Final Exam is **this Saturday 3/16 from 7-10PM**.
 - All sections will take the exam in Catalyst 0125. Seating assignments will be released Friday.
- Collaborative study session on **Friday 3/15 from 5-8PM** in Solis 104.
- If at least 75% of the class fills out both [SETs](#) and the DSC 10-specific [End-of-Quarter Survey](#), then the entire class will have **1% of extra credit added to their overall grade**. We value your feedback!

Agenda

- We'll work through selected problems from the Fall 2023 Final Exam.
- We won't write any code, since you can't run code during the exam. Instead, we'll try to think like the computer ourselves.
- These annotated slides will be posted after lecture is over, as will the solutions to the entire exam.
- **Try the problems with us!**

Fall 2023 Final Exam

Access the exam [here](#). Make sure to read the data info sheet at the top before starting.

Problem 6

Aaron wants to explore the discrepancy in fraud rates between "discover" transactions and "mastercard" transactions. To do so, he creates the DataFrame `ds_mc`, which only contains the rows in `txn` corresponding to "mastercard" or "discover" transactions.

After he creates `ds_mc`, Aaron groups `ds_mc` on the "card" column using two different aggregation methods. The relevant columns in the resulting DataFrames are shown below.

```
ds_mc.groupby("card").sum()
```

	is_fraud
card	
discover	160
mastercard	4000

```
ds_mc.groupby("card").count()
```

	is_fraud
card	
discover	2000
mastercard	40000

Aaron decides to perform a test of the following pair of hypotheses:

- **Null Hypothesis:** The proportion of fraudulent "mastercard" transactions is **the same as** the proportion of fraudulent "discover" transactions.
- **Alternative Hypothesis:** The proportion of fraudulent "mastercard" transactions is **less than** the proportion of fraudulent "discover" transactions.

As his test statistic, Aaron chooses the **difference in proportion of transactions that are fraudulent**, in the order "mastercard" minus "discover".

Problem 6

Problem 6.1

What type of statistical test is Aaron performing?

- Standard hypothesis test
- Permutation test

Problem 6.2

What is the value of the observed statistic? Give your answer either as an exact decimal or simplified fraction.

```
ds_mc.groupby("card").sum()
```

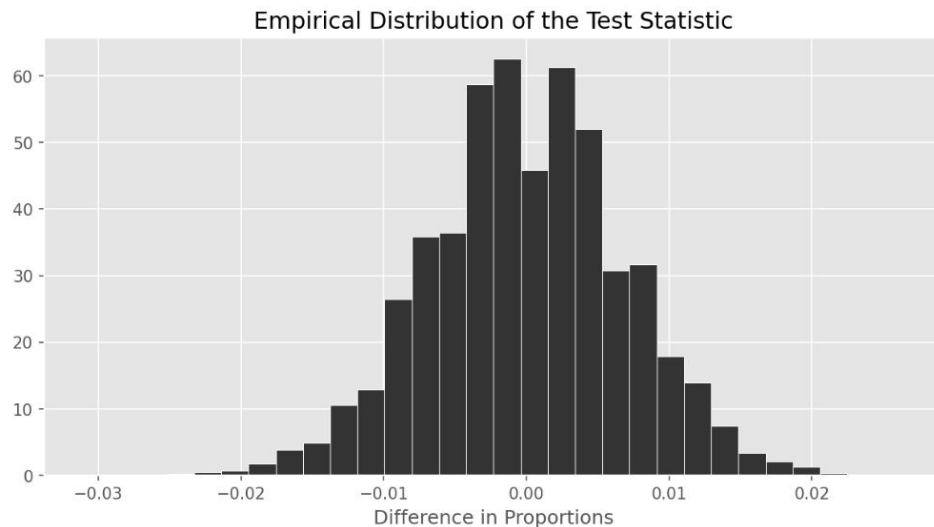
	is_fraud
card	
discover	160
mastercard	4000

```
ds_mc.groupby("card").count()
```

	is_fraud
card	
discover	2000
mastercard	40000

Problem 6

The empirical distribution of Aaron's chosen test statistic is shown below.



Problem 6.3

Which of the following is closest to the p-value of Aaron's test?

- 0.001
- 0.37
- 0.63
- 0.94
- 0.999

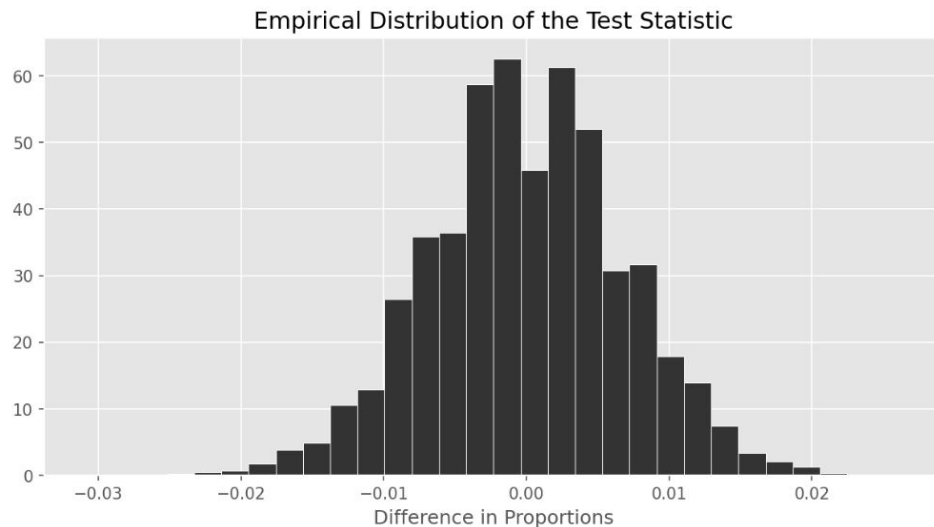
Problem 6.4

What is the conclusion of Aaron's test?

- The proportion of fraudulent "mastercard" transactions is **less than** the proportion of fraudulent "discover" transactions.
- The proportion of fraudulent "mastercard" transactions is **greater than** the proportion of fraudulent "discover" transactions.
- The test results are inconclusive.
- None of the above.

Problem 6

The empirical distribution of Aaron's chosen test statistic is shown below.



Aaron now decides to test a slightly different pair of hypotheses.

- **Null Hypothesis:** The proportion of fraudulent "mastercard" transactions is **the same as** the proportion of fraudulent "discover" transactions.
- **Alternative Hypothesis:** The proportion of fraudulent "mastercard" transactions is **greater than** the proportion of fraudulent "discover" transactions.

He uses the same test statistic as before.

Problem 6.5

Which of the following is closest to the p-value of Aaron's new test?

- 0.001
- 0.06
- 0.37
- 0.63
- 0.94
- 0.999

Problem 9

The DataFrame `ten_txns`, displayed in its entirety below, contains a simple random sample of 10 rows from `txn`.

<code>transaction_id</code>	<code>is_fraud</code>	<code>amount</code>	<code>method</code>	<code>card</code>	<code>lifetime</code>	<code>browser</code>
<code>3169166</code>	True	100.00	credit	visa	532601.00	chrome 63.0
<code>3093921</code>	False	100.00	debit	mastercard	173276.00	mobile safari 10.0
<code>3137058</code>	False	100.00	debit	visa	120000.00	chrome 63.0
<code>3063164</code>	False	100.00	debit	visa	141342.00	mobile safari 11.0
<code>3051461</code>	False	75.00	credit	visa	153871.00	ie 11.0 for desktop
<code>3171154</code>	False	25.00	debit	visa	182654.00	mobile safari generic
<code>3222420</code>	False	25.00	credit	visa	50199.00	safari generic
<code>3226397</code>	False	22.96	debit	mastercard	122352.00	safari generic
<code>3073572</code>	False	9.33	credit	mastercard	66703.00	mobile safari 11.0
<code>3253371</code>	False	5.00	debit	visa	3007.00	chrome 64.0

Problem 9.1

Suppose we randomly select one transaction from `ten_txns`. What is the probability that the selected transaction is made with a `"card"` of `"mastercard"` or a `"method"` of `"debit"`?

Problem 9

The DataFrame `ten_txns`, displayed in its entirety below, contains a simple random sample of 10 rows from `txn`.

<code>transaction_id</code>	<code>is_fraud</code>	<code>amount</code>	<code>method</code>	<code>card</code>	<code>lifetime</code>	<code>browser</code>
<code>3169166</code>	True	100.00	credit	visa	532601.00	chrome 63.0
<code>3093921</code>	False	100.00	debit	mastercard	173276.00	mobile safari 10.0
<code>3137058</code>	False	100.00	debit	visa	120000.00	chrome 63.0
<code>3063164</code>	False	100.00	debit	visa	141342.00	mobile safari 11.0
<code>3051461</code>	False	75.00	credit	visa	153871.00	ie 11.0 for desktop
<code>3171154</code>	False	25.00	debit	visa	182654.00	mobile safari generic
<code>3222420</code>	False	25.00	credit	visa	50199.00	safari generic
<code>3226397</code>	False	22.96	debit	mastercard	122352.00	safari generic
<code>3073572</code>	False	9.33	credit	mastercard	66703.00	mobile safari 11.0
<code>3253371</code>	False	5.00	debit	visa	3007.00	chrome 64.0

Problem 9.2

Suppose we randomly select two transactions from `ten_txns`, without replacement, and learn that neither of the selected transactions is for an amount of 100 dollars. Given this information, what is the probability that:

- the first transaction is made with a `card` of `visa` and a `method` of `debit`, and
- the second transaction is made with a `card` of `visa` and a `method` of `credit`?

Problem 11

On Reddit, Yutian read that 22% of all online transactions are fraudulent. She decides to test the following hypotheses:

- **Null Hypothesis:** The proportion of online transactions that are fraudulent is **0.22**.
- **Alternative Hypothesis:** The proportion of online transactions that are fraudulent is not **0.22**.

To test her hypotheses, she decides to create a **95%** confidence interval for the proportion of online transactions that are fraudulent using the Central Limit Theorem.

Unfortunately, she doesn't have access to the entire `txn` DataFrame; rather, she has access to a simple random sample of `txn` of size n . In her sample, the proportion of transactions that are fraudulent is **0.2** (or equivalently, $\frac{1}{5}$).

Problem 11.1

The width of Yutian's confidence interval is of the form

$$\frac{c}{5\sqrt{n}}$$

where n is the size of her sample and c is some positive integer.

What is the value of c ? Give your answer as an integer.

Hint: Use the fact that in a collection of 0s and 1s, if the proportion of values that are 1 is p , the standard deviation of the collection is $\sqrt{p(1-p)}$.

Problem 11

On Reddit, Yutian read that 22% of all online transactions are fraudulent. She decides to test the following hypotheses:

- **Null Hypothesis:** The proportion of online transactions that are fraudulent is **0.22**.
- **Alternative Hypothesis:** The proportion of online transactions that are fraudulent is not **0.22**.

To test her hypotheses, she decides to create a **95%** confidence interval for the proportion of online transactions that are fraudulent using the Central Limit Theorem.

Unfortunately, she doesn't have access to the entire `txn` DataFrame; rather, she has access to a simple random sample of `txn` of size n . In her sample, the proportion of transactions that are fraudulent is **0.2** (or equivalently, $\frac{1}{5}$).

Problem 11.2

There is a positive integer J such that:

- If $n < J$, Yutian will fail to reject her null hypothesis at the **0.05** significance level.
- If $n > J$, Yutian will reject her null hypothesis at the **0.05** significance level.

What is the value of J ? Give your answer as an integer.

Problem 12

On Reddit, Keenan also read that 22% of all online transactions are fraudulent. He decides to test the following hypotheses at the **0.16 significance level**:

- **Null Hypothesis:** The proportion of online transactions that are fraudulent is **0.22**.
- **Alternative Hypothesis:** The proportion of online transactions that are fraudulent is not **0.22**.

Keenan has access to a simple random sample of `txn` of size **500**. In his sample, the proportion of transactions that are fraudulent is **0.23**.

Below is an incomplete implementation of the function `reject_null`, which creates a bootstrap-based confidence interval and returns **True** if the conclusion of Keenan's test is to **reject** the null hypothesis, and **False** if the conclusion is to **fail to reject** the null hypothesis, all at the **0.16** significance level.

```
def reject_null():
    fraud_counts = np.array([])
    for i in np.arange(10000):
        fraud_count = np.random.multinomial(500, __a__)[0]
        fraud_counts = np.append(fraud_counts, fraud_count)

    L = np.percentile(fraud_counts, __b__)
    R = np.percentile(fraud_counts, __c__)

    if __d__ < L or __d__ > R:
        # Return True if we REJECT the null.
        return True
    else:
        # Return False if we FAIL to reject the null.
        return False
```

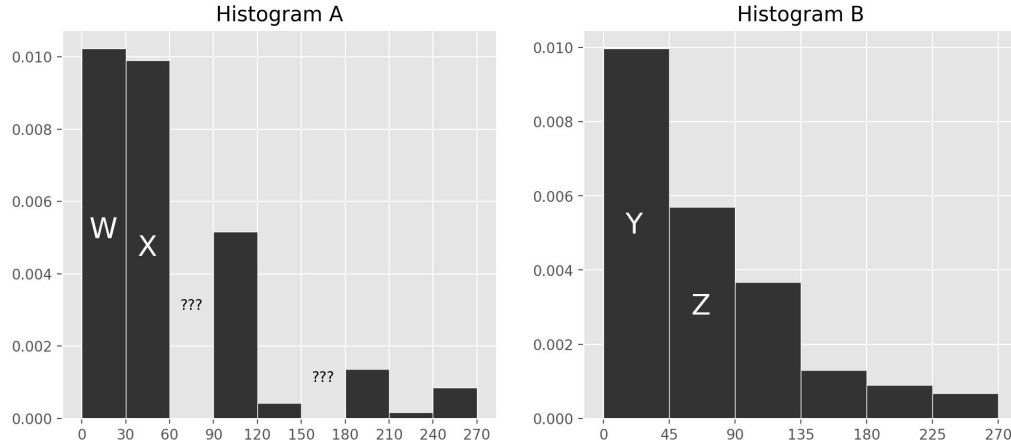
Fill in the blanks so that `reject_null` works as intended.

Hint: Your answer to (d) should be an integer greater than 50.

Problem 13

Ashley doesn't have access to the entire `txn` DataFrame; instead, she has access to a simple random sample of **400** rows of `txn`.

She draws two histograms, each of which depicts the distribution of the `"amount"` column in her sample, using different bins.



Unfortunately, DataHub is being finicky and so two of the bars in Histogram A are deleted after it is created.

Problem 13.1

In Histogram A, which of the following bins contains approximately 60 transactions?

- [30, 60)
- [90, 120)
- [120, 150)
- [180, 210)

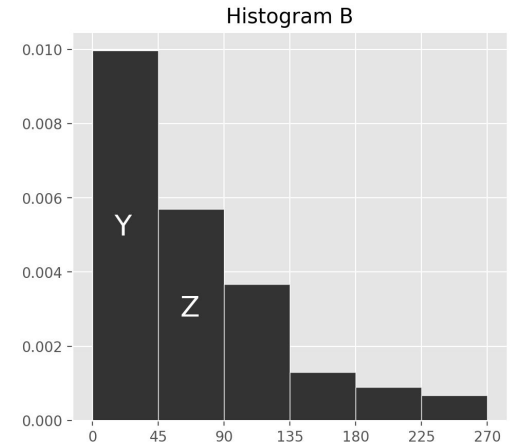
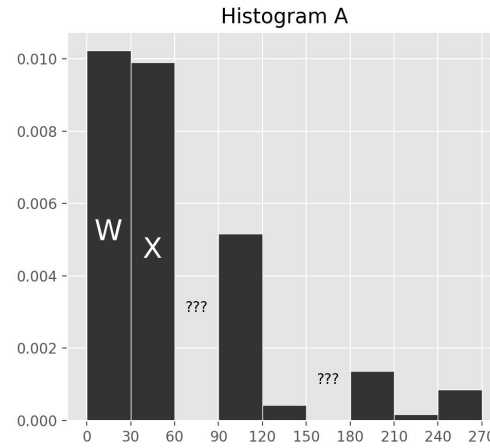
Problem 13

Problem 13.2

Let w , x , y , and z be the heights of bars W , X , Y , and Z , respectively. For instance, y is about 0.01.

Which of the following expressions gives the height of the bar corresponding to the $[60, 90)$ bin in Histogram A?

- $(y + z) - (w + x)$
- $(w + x) - (y + z)$
- $\frac{3}{2}(y + z) - (w + x)$
- $(y + z) - \frac{3}{2}(w + x)$
- $3(y + z) - 2(w + x)$
- $2(y + z) - 3(w + x)$
- None of the above.



Problem 14

As mentioned in the previous problem, Ashley has sample of 400 rows of `txn`. Coincidentally, in Ashley's sample of 400 transactions, the mean and standard deviation of the `"amount"` column both come out to 70 dollars.

Problem 14.1

Fill in the blank:

"According to Chebyshev's inequality, at most 25 transactions in Ashley's sample are above ___ dollars; the rest must be below ___ dollars."

What goes in the blank? Give your answer as an **integer**. Both blanks are filled in with the same number.

Problem 14

As mentioned in the previous problem, Ashley has sample of 400 rows of `txn`. Coincidentally, in Ashley's sample of 400 transactions, the mean and standard deviation of the `"amount"` column both come out to 70 dollars.

Problem 14.3

The predicted lifetime spending, in **dollars**, of a card with a transaction amount of 280 dollars is of the form $f \cdot c$, where f is a fraction. What is the value of f ? Give your answer as a simplified fraction.

Problem 14

As mentioned in the previous problem, Ashley has sample of 400 rows of `txn`. Coincidentally, in Ashley's sample of 400 transactions, the mean and standard deviation of the `"amount"` column both come out to 70 dollars.

Problem 14.4

Suppose the intercept of the regression line, when both transaction amounts and lifetime spending are measured in **dollars**, is 40. What is the value of c ? Give your answer as an integer.

Problem 1

Problem 1.1

Nate's favorite number is 5. He calls a number "lucky" if it's greater than 500 or if it contains a 5 anywhere in its representation. For example, 1000.04 and 5.23 are both lucky numbers.

Complete the implementation of the function `check_lucky`, which takes in a number as a float and returns `True` if it is lucky and `False` otherwise. Then, add a column named `"is_lucky"` to `txn` that contains `True` for lucky transaction amounts and `False` for all other transaction amounts, and save the resulting DataFrame to the variable `luck`.

```
def check_lucky(x):  
    return __ (a) __
```

```
luck = txn.assign(is_lucky = __ (b) __)
```

1. What goes in blank (a)?
2. What goes in blank (b)?

Problem 1

Problem 1.2

Fill in the blanks below so that `lucky_prop` evaluates to the proportion of fraudulent `"visa"` card transactions whose transaction amounts are lucky.

```
visa_fraud = __(a)__
```

```
lucky_prop = visa_fraud.__(b)__.mean()
```

1. What goes in blank (a)?
2. What goes in blank (b)?

Problem 1

Problem 1.3

Fill in the blanks below so that `lucky_prop` is one value in the Series `many_props`.

```
many_props = luck.groupby(__(a)__).mean().get(__(b)__)
```

1. What goes in blank (a)?
2. What goes in blank (b)?

Problem 2

Consider the DataFrame `combo`, defined below.

```
combo = txn.groupby(["is_fraud", "method", "card"]).mean()
```

Problem 2.1

What is the maximum possible value of `combo.shape[0]`? Give your answer as an integer.

Problem 10

As a senior suffering from senioritis, Weiyue has plenty of time on his hands. 1,000 times, he repeats the following process, creating 1,000 confidence intervals:

1. Collect a simple random sample of 100 rows from `txn`.
2. Resample from his sample 10,000 times, computing the mean transaction amount in each resample.
3. Create a 95% confidence interval by taking the middle 95% of resample means.

He then computes the width of each confidence interval by subtracting its left endpoint from its right endpoint; e.g. if $[2, 5]$ is a confidence interval, its width is 3. This gives him 1,000 widths. Let M be the mean of these 1,000 widths.

Problem 10.1

Select the true statement below.

- About 950 of Weiyue's intervals will contain the mean transaction amount of all transactions ever.
- About 950 of Weiyue's intervals will contain the mean transaction amount of all transactions in `txn`.
- About 950 of Weiyue's intervals will contain the mean transaction amount of all transactions in the first random sample of 100 rows of `txn` Weiyue took.
- About 950 of Weiyue's intervals will contain M .