Lecture 27 – Final Review

DSC 10, Fall 2023

Announcements

- The Final Exam is **this Saturday 12/9 from 7-10PM**. See **this post on Ed** for more details, including your assigned room and seat (ignore the room that WebReg says!).
- Lab 7 is due **tomorrow at 11:59PM**.
- You can still submit the Final Project late using slip days.
    - If one or both partners has run out of slip days and you submit the project late, we will reallocate slip days towards the Final Project, away from lesser-weighted assignments. See the **syllabus** for more details.
- If at least 85% 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!
    - As of yesterday, the End-of-Quarter Survey had around a 30% completion rate.

Agenda

- We'll work through selected problems from the Spring 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!**

Spring 2023 Final Exam

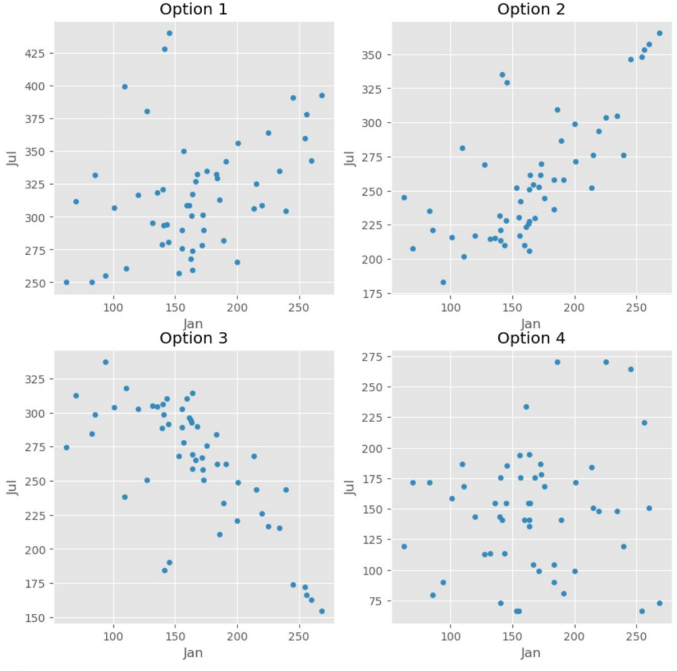
Access the exam **here**. Make sure to read the data info sheet at the top before starting.

Problem 11

Raine finds the regression line that predicts the number of sunshine hours in July (y) for a city given its number of sunshine hours in January (x). In doing so, they find that the correlation between the two variables is $\frac{2}{5}$.

Problem 11.1

Which of these could be a scatter plot of number of sunshine hours in July vs. number of sunshine hours in January?



- Option 1
- Option 2
- Option 3
- Option 4

Problem 11.2

Suppose the standard deviation of the number of sunshine hours in January for cities in California is equal to the standard deviation of the number of sunshine hours in July for cities in California.

Raine's hometown of Santa Clarita saw 60 more sunshine hours in January than the average California city did. How many **more sunshine hours than average** does the regression line predict that Santa Clarita will have in July? Give your answer as a positive integer. *(Hint: You'll need to use the fact that the correlation between the two variables is $\frac{2}{5}$.)*

Problem 11

Raine finds the regression line that predicts the number of sunshine hours in July (y) for a city given its number of sunshine hours in January (x). In doing so, they find that the correlation between the two variables is $\frac{2}{5}$.

To imagine what the dataset may look like in a few years, Anthony subtracts 5 from the number of sunshine hours in both January and July for all California cities in the dataset – i.e., he subtracts 5 from each x value and 5 from each y value in the dataset. He then creates a regression line to use the new xs to predict the new ys.

Problem 11.3

What is the slope of Anthony's new regression line?

Problem 11.4

Suppose the intercept of Raine's original regression line – that is, before Anthony subtracted 5 from each x and each y – was 10. What is the intercept of Anthony's new regression line?

- -7
- -5
- -3
- 0
- 3
- 5
- 7

Problem 11

Jasmine is trying to get as far away from Anthony as possible and has a trip to Chicago planned after finals. Chicago is known for being very warm and sunny in the summer but cold, rainy, and snowy in the winter. She decides to build a regression line that uses month of the year (where 1 is January, 2 is February, 12 is December, etc.) to predict the number of sunshine hours in Chicago.

Problem 11.5

What would you expect to see in a residual plot of Jasmine's regression line?

- ○ A patternless cloud of points
- ○ A distinctive pattern in the residual plot
- ○ Heteroscedasticity (residuals that are not evenly vertically spread)

Problem 10

Costin, a San Francisco native, will be back in San Francisco over the summer, and is curious as to whether it is true that about $\frac{3}{4}$ of days in San Francisco are sunny.

Fast forward to the end of September: Costin counted that of the 30 days in September, 27 were sunny in San Francisco. To test his theory, Costin came up with two pairs of hypotheses.

Pair 1:

- **Null Hypothesis**: The probability that it is sunny on any given day in September in San Francisco is $\frac{3}{4}$, independent of all other days.
- **Alternative Hypothesis**: The probability that it is sunny on any given day in September in San Francisco is **not** $\frac{3}{4}$.

Pair 2:

- **Null Hypothesis**: The probability that it is sunny on any given day in September in San Francisco is $\frac{3}{4}$, independent of all other days.
- **Alternative Hypothesis**: The probability that it is sunny on any given day in September in San Francisco is **greater than** $\frac{3}{4}$.

For each test statistic below, choose whether the test statistic could be used to test Pair 1, Pair 2, both, or neither. Assume that all days are either sunny or cloudy, and that we cannot perform two-tailed hypothesis tests. (If you don't know what those are, you don't need to!)

Problem 10.1

The difference between the number of sunny days and number of cloudy days

- ○ Pair 1
- ○ Pair 2
- ○ Both
- ○ Neither

Problem 10.2

The absolute difference between the number of sunny days and number of cloudy days

Problem 10.3

The difference between the proportion of sunny days and $\frac{1}{4}$

Problem 10.4

The absolute difference between the proportion of cloudy days and $\frac{1}{4}$

Teresa also wants to go to Australia, but can't take time off work in January, and so she plans a trip to The Land Down Under (Australia) in February instead. She finds that the mean number of sunshine hours in February for all 15 Australian cities in `sun` is 250, with a standard deviation of 15.

Problem 8.4

According to Chebyshev's inequality, at least what percentage of Australian cities in `sun` see between 200 and 300 sunshine hours in February?

- 9%
- 30%
- 33.3%
- 91%
- 95%
- 99.73%

Problem 7

Gabriel is originally from Texas and is trying to convince his friends that Texas has better weather than California. Sophia, who is originally from San Diego, is determined to prove Gabriel wrong.

Coincidentally, both are born in February, so they decide to look at the mean number of sunshine hours of all cities in California and Texas in February. They find that the mean number of sunshine hours for California cities in February is 275, while the mean number of sunshine hours for Texas cities in February is 250. They decide to test the following hypotheses:

- **Null Hypothesis**: The distribution of sunshine hours in February for cities in California and Texas are drawn from the same population distribution.
- **Alternative Hypothesis**: The distribution of sunshine hours in February for cities in California and Texas are not drawn from the same population distribution; rather, California cities see more sunshine hours in February on average than Texas cities.

The test statistic they decide to use is:

$$\text{mean sunshine hours in California cities} - \text{mean sunshine hours in Texas cities}$$

The test statistic they decide to use is:

$$\text{mean sunshine hours in California cities} - \text{mean sunshine hours in Texas cities}$$

To simulate data under the null, Sophia proposes the following plan:

1. Count the number of Texas cities, and call that number `t`. Count the total number of cities in both California and Texas, and call that number `n`.
2. Find the total number of sunshine hours across all California and Texas cities in February, and call that number `total`.
3. Take a random sample of `t` sunshine hours from the entire sequence of California and Texas sunshine hours in February in the dataset. Call this random sample `t_samp`.
4. Find the difference between the mean of the values that are not in `t_samp` (the California sample) and the mean of the values that are in `t_samp` (the Texas sample).

Problem 7.1

What type of test is this?

- ○ Hypothesis test
- ○ Permutation test

```
def one_stat(df):
    # You don't need to fill in the ...,
    # assume we've correctly filled them in so that
    # texas_only has only the "Texas" rows from df.
    texas_only = ...
    t = texas_only.shape[0]
    n = df.shape[0]

    total = df.get("Feb").sum()

    t_samp = np.random.choice(df.get("Feb"), t, __(b)__)

    c_mean = __(c)__
    t_mean = t_samp.sum() / t
    return c_mean - t_mean
```

Problem 7.2

What goes in blank (b)?

- ○ `replace=True`
- ○ `replace=False`

Problem 7.3

What goes in blank (c)? *(Hint: Our solution uses 4 of the variables that are defined before `c_mean`.)*

Fill in the blanks below to accurately complete the provided statement.

“If Sophia and Gabriel want to test the null hypothesis that the mean number of sunshine hours in February in the two states is equal using a different tool, they could use bootstrapping to create a confidence interval for the true value of the test statistic they used in the above test and check whether __(d)__ is in the interval."

Problem 7.4

What goes in blank (d)? Your answer should be a specific number.

Problem 6

Oren's favorite bakery in San Diego is Wayfarer. After visiting frequently, he decides to learn how to make croissants and baguettes himself, and to do so, he books a trip to France.

Oren is interested in estimating the mean number of sunshine hours in July across all 10,000+ cities in France. Using the 16 French cities in `sun`, Oren constructs a 95% Central Limit Theorem (CLT)-based confidence interval for the mean sunshine hours of all cities in France. The interval is of the form $[L, R]$, where L and R are positive numbers.

Problem 6.1

Which of the following expressions is equal to the standard deviation of the number of sunshine hours of the 16 French cities in `sun`?

- $R - L$
- $\frac{R-L}{2}$
- $\frac{R-L}{4}$
- $R + L$
- $\frac{R+L}{2}$
- $\frac{R+L}{4}$

Problem 6.2

True or False: There is a 95% chance that the interval $[L, R]$ contains the mean number of sunshine hours in July of all 16 French cities in `sun`.

- True
- False

Problem 6.3

True or False: If we collected 1,000 new samples of 16 French cities and computed the mean of each sample, then about 95% of the new sample means would be contained in $[L, R]$.

- True
- False

Problem 6.4

True or False: If we collected 1,000 new samples of 16 French cities and created a 95% confidence interval using each one, then chose one of the 1,000 new intervals at random, the chance that the randomly chosen interval contains the mean sunshine hours in July across all cities in France is approximately 95%.

- True
- False

Problem 6

Oren's favorite bakery in San Diego is Wayfarer. After visiting frequently, he decides to learn how to make croissants and baguettes himself, and to do so, he books a trip to France.

Oren is interested in estimating the mean number of sunshine hours in July across all 10,000+ cities in France. Using the 16 French cities in `sun`, Oren constructs a 95% Central Limit Theorem (CLT)-based confidence interval for the mean sunshine hours of all cities in France. The interval is of the form $[L, R]$, where L and R are positive numbers.

Problem 6.5

True or False: The interval $[L, R]$ is centered at the mean number of sunshine hours in July across all cities in France.

- ○ True
- ○ False

In addition to creating a 95% CLT-based confidence interval for the mean sunshine hours of all cities in France, Oren would like to create a 72% bootstrap-based confidence interval for the mean sunshine hours of all cities in France.

Oren resamples from the 16 French sunshine hours in `sun` 10,000 times and creates an array named `french_sunshine` containing 10,000 resampled means. He wants to find the left and right endpoints of his 72% confidence interval:

```
boot_left = np.percentile(french_sunshine, __(a)__)
boot_right = np.percentile(french_sunshine, __(b)__)
```

Problem 6.6

Fill in the blanks so that `boot_left` and `boot_right` evaluate to the left and right endpoints of a 72% confidence interval for the mean sunshine hours in July across all cities in France.

What goes in blanks (a) and (b)?

Suppose we are interested in testing the following pair of hypotheses.

- **Null Hypothesis**: The mean number of sunshine hours of all cities in France in July is equal to 225.
- **Alternative Hypothesis**: The mean number of sunshine hours of all cities in France in July is not equal to 225.

Problem 6.7

Suppose that when Oren uses `[boot_left, boot_right]`, his 72% bootstrap-based confidence interval, he fails to reject the null hypothesis above. If that's the case, then when using $[L, R]$, his 95% CLT-based confidence interval, what is the conclusion of his hypothesis test?

- ○ Reject the null
- ○ Fail to reject the null
- ○ Impossible to tell

Problem 6.8

Suppose that Oren also creates a 72% CLT-based confidence interval for the mean sunshine hours of all cities in France in July using the same 16 French cities in `sun` he started with. When using his 72% CLT-based confidence interval, he fails to reject the null hypothesis above. If that's the case, then when using $[L, R]$, what is the conclusion of his hypothesis test?

- ○ Reject the null
- ○ Fail to reject the null
- ○ Impossible to tell

Problem 6.9

True or False: The significance levels of both hypothesis tests described in part (h) are equal.

- ○ True
- ○ False

Problem 5

In some cities, the number of sunshine hours per month is relatively consistent throughout the year. São Paulo, Brazil is one such city; in all months of the year, the number of sunshine hours per month is somewhere between 139 and 173. New York City's, on the other hand, ranges from 139 to 268.

Gina and Abel, both San Diego natives, are interested in assessing how "consistent" the number of sunshine hours per month in San Diego appear to be. Specifically, they'd like to test the following hypotheses:

- **Null Hypothesis**: The number of sunshine hours per month in San Diego is drawn from the uniform distribution, $\left[\frac{1}{12}, \frac{1}{12}, \ldots, \frac{1}{12}\right]$. (In other words, the number of sunshine hours per month in San Diego is equal in all 12 months of the year.)
- **Alternative Hypothesis**: The number of sunshine hours per month in San Diego is not drawn from the uniform distribution.

As their test statistic, Gina and Abel choose the total variation distance. To simulate samples under the null, they will sample from a categorical distribution with 12 categories — January, February, and so on, through December — each of which have an equal probability of being chosen.

Problem 5.1

In order to run their hypothesis test, Gina and Abel need a way to calculate their test statistic. Below is an incomplete implementation of a function that computes the TVD between two arrays of length 12, each of which represent a categorical distribution.

```
def calculate_tvd(dist1, dist2):
    return np.mean(np.abs(dist1 - dist2)) * ____
```

Fill in the blank so that `calculate_tvd` works as intended.

- `1 / 6`
- `1 / 3`
- `1 / 2`
- `2`
- `3`
- `6`

Now, complete the implementation of the function `uniform_test`, which takes in an array `observed_counts` of length 12 containing the number of sunshine hours each month in a city and returns the p-value for the hypothesis test stated at the start of the question.

```
def uniform_test(observed_counts):
    # The values in observed_counts are counts, not proportions!
    total_count = observed_counts.sum()
    uniform_dist = __(b)__
    tvds = np.array([])
    for i in np.arange(10000):
        simulated = __(c)__
        tvd = calculate_tvd(simulated, __(d)__)
        tvds = np.append(tvds, tvd)
    return np.mean(tvds __(e)__ calculate_tvd(uniform_dist, __(f)__))
```

Problem 5.2

What goes in blank (b)? *(Hint: The function `np.ones(k)` returns an array of length `k` in which all elements are `1`.)*

Problem 5.3

What goes in blank (c)?

- `np.random.multinomial(12, uniform_dist)`
- `np.random.multinomial(12, uniform_dist) / 12`
- `np.random.multinomial(12, uniform_dist) / total_count`
- `np.random.multinomial(total_count, uniform_dist)`
- `np.random.multinomial(total_count, uniform_dist) / 12`
- `np.random.multinomial(total_count, uniform_dist) / total_count`