PSTAT 5A / **FINAL EXAM STUDY PROBLEMS** / Summer 24

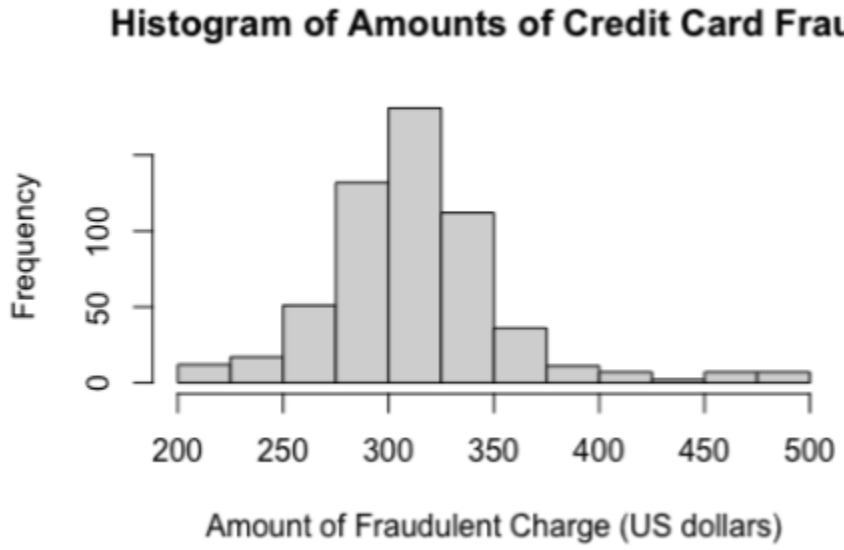
---

**Exam Instructions:**

- You will have 80 minutes to complete the exam.
    - Do not begin until you are instructed to do so.
    - The last 10 minutes of the exam, please remain seated until the examination time concludes.
- The exam has a multiple choice section and a free response section. The multiple choice section has 20 questions and the free response has 2 questions with multiple parts.
- Fill in the bubble corresponding to your answer on the provided scantron. Any work on the multiple choice section of the exam WILL NOT be reviewed.
- <u>You must show your student ID card to turn in your exam.</u>
- The use of a scientific calculator is permitted.
- You are also permitted to one 8.5in by 11in paper, front and back, with handwritten formula.
- You are not allowed to use anything else (no notes, laptop, phones).
- Good luck!

MULTIPLE CHOICE / **FINAL EXAM STUDY PROBLEMS** / Summer 24

Occasionally someone uses another person's credit card (without permission) to purchase things—this is a form of credit card fraud. Recently, I was contacted by my bank to let me know that the bank suspected there were some fraudulent charges made to the account. Sure enough, someone was using our credit card fraudulently. (Everything is fine, the credit card account was closed and the charges were reversed.) In these questions, you will investigate the amounts charged in a random sample of 575 fraudulent credit card charges between $200 and $500.
Consider the following histogram and summary statistics for the amounts of fraudulent credit card charges.



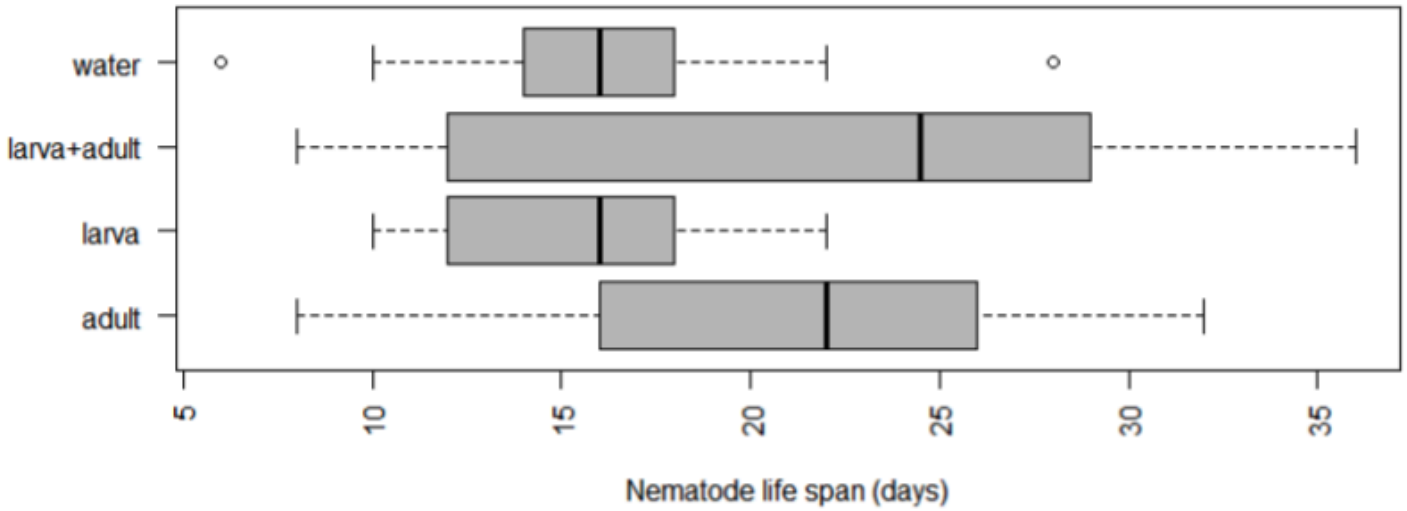

**Problem 1.**
Which of the following is the most reasonable value for the standard deviation for the amounts of the fraudulent charges in this sample?

a. 80
b. 20
c. 50

**Problem 2.**
Suppose that the maximum amount was incorrectly entered as $4988 instead of $498.80. Which of the following is the most reasonable value for the sample mean if this typo was not caught? (Note: You should be able to answer this without doing any calculations.)

a. $321.03
b. $313.20
c. $305.41

As the “baby boom” generation ages, interest in finding treatments that extend lifespan has surged. Experimental research on the aging process primarily uses nonhuman subjects. A notable experiment investigated the effect of trimethadione, an anticonvulsant medication often used in treating epilepsy, on the life span of nematode worms. The plot below shows the life span of 200 nematode worms (in days) that were randomly assigned to one of four treatments of $n$ = 50 worms: three trimethadione treatments (provided at the larval stage, adult stage, or both stages) and a water treatment, which served as a control.




Use this plot to make comparisons about the relationship between Quantity A and Quantity B in Questions 3-6. The questions are independent of one another, so you may use each choice more than once or not at all.

**Problem 3**.
Quantity A: The maximum life span of nematodes in the “larva” treatment group
Quantity B: The median life span of nematodes in the “adult” treatment group

a. Quantity A is greater
b. Quantity B is greater
c. The quantities are about the same
d. The relationship cannot be determined without more information

**Problem 4.**
Quantity A: The range of life spans in the “water” treatment group
Quantity B: The range of life spans in the “larva” treatment group

a. Quantity A is greater
b. Quantity B is greater
c. The quantities are about the same
d. The relationship cannot be determined without more information

**Problem 5**.
Quantity A: The z-score of a nematode that lived to 20 days and was in the “water” treatment group
Quantity B: The z-score of a nematode that lived to 20 days and was in the “larva + adult” treatment group

a. Quantity A is greater
b. Quantity B is greater
c. The quantities are about the same
d. The relationship cannot be determined without more information

**Problem 6**.
Quantity A: The largest IQR across all four treatment groups
Quantity B: The smallest range across all four treatment groups

a. Quantity A is greater
b. Quantity B is greater
c. The quantities are about the same
d. The relationship cannot be determined without more information

**Problem 7**.
Which of the following statements best describes the relationship between a parameter and a statistic?

a. A parameter is used to estimate a statistic.
b. A statistic is used to estimate a parameter.
c. The value of a statistic will always be smaller than the value of a parameter.
d. The value of a parameter is always known, but the value of a statistic is rarely ever known.

**Problem 8**.
Officials at a university want to estimate the percentage of students at the university who earned scholarships outside of scholarships from the university. What statistical method would be most appropriate for this?

a. A hypothesis test for a mean
b. A confidence interval for a mean
c. A hypothesis test for a proportion
d. A confidence interval for a proportion

**Problem 9.**
Researchers at the Pew Research Center are interested in studying teens and video games. They want to determine if teens spend an average of more than 3 hours a day playing video games. The notation for the point estimate is

a. $\hat{p}$
b. $p$

c. $\bar{x}$
d. $\mu$

**Problem 10.**
Researchers at the Pew Research Center are interested in studying teens and video games. They want to estimate the percentage of teens who think that playing video games helps their problem solving skills. The notation for the parameter is

a. $\hat{p}$
b. $p$
c. $\bar{x}$
d. $\mu$

In the United States, tipping is requested (and expected) for completing many jobs, such as delivering food, making a beverage, and personal care such as haircuts. The Pew Research Center is interested in understanding if there is evidence that adults who have worked a job that receives tips are more likely to tip at a coffee shop than adults who have not worked a job that receives tips. Data will be collected by posting a QR code at the checkout counter of coffee shops in various large cities.

**Problem 11.**
The variable workedTipJob is coded yes if the adult has worked a job that receives tips and no if the adult has not worked a job that receives tips. The variable workedTipJob is a

a. Categorical variable.
b. Quantitative variable.

**Problem 12.**
The variable amountTip is the amount (in USD) that the adult tipped on that visit to the coffee shop. The variable amountTip is a

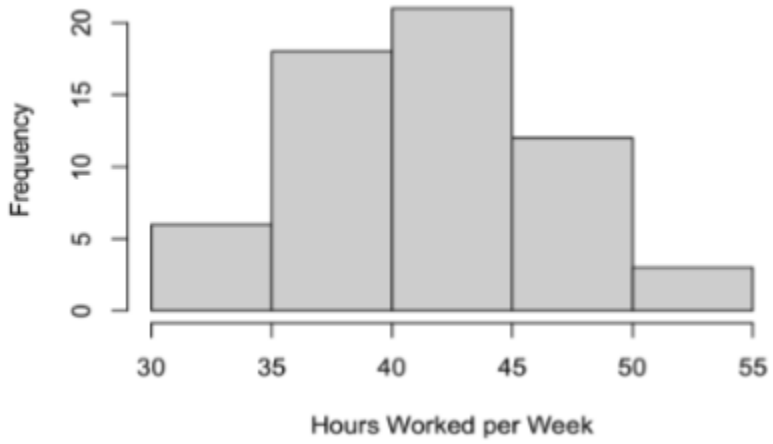
a. Categorical variable.
b. Quantitative variable.

**Problem 13.**
The variable coffeeTip is the variable that indicates whether the adult left a tip when they got coffee. The variable coffeeTip is a(n)

a. Explanatory variable.
b. Response variable.

A typical workweek for full-time workers in the United States consists of 40 hours. Angel, a mental health professional, is concerned that full-time workers spend longer than 40 hours per week working, on average. They collect data on the number of hours worked in a typical week from a random sample of 60 full-time workers in the United States.

Angel used Python to analyze their data. Here are the histogram and summary statistics for Angel's data:




```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   sd(work)
 30.42   38.83   41.81   41.76   45.08   54.99  1] 4.925602
```

**Problem 14.**

The conditions for inference are satisfied because

a. The success-failure condition is satisfied.
b. This was a random sample of full-time workers.
c. The histogram suggests that the population of hours worked per week is approximately normal.
d. Both (a) and (c).
e. Both (b) and (c).

**Problem 15.**

What is the value of the point estimate for the average number of hours worked per week by full-time workers in the United States?

a. 41.81
b. 41.76
c. 40
d. 4.926
e. None of the above

**Problem 16.**

Angel wants to test the hypothesis $H_0$: $\mu = 40$ against the one-sided alternative $H_A$: $\mu > 40$. They ran the hypothesis test in Python and got $TS = 2.762,\ df = 59$, p-value = 0.007647. When Angel was away from their computer, they noticed a minor mistake: they had generated results for the two-sided alternative instead of the one-sided alternative they needed. How will the correct one-sided test affect the value of the test statistic?

a. The test statistic will be negative ($t$ = −2.762)

b. The test statistic will half of the value ($t$ = 1.381)
c. The test statistic will be the same ($t$ = 2. 762)
d. The test statistic will be twice the value ($t$ = 5.524)

**Problem 17**.
As mentioned in Question 16, Angel accidentally ran a two-sided test instead of the one-sided test they were interested in. What will change about the p-value when Angel runs the correct one-sided test?

a. The p-value will be negative (p-value = –0.0076)
b. The p-value will half of the value (p-value = 0.0038)
c. The p-value will be the same (p-value = 0.0076)
d. The p-value will twice the value (p-value = 0.0152)

**Problem 18.**
Angel's results are statistically significant at the $\alpha$ = 0.05 significance level. Which of the following is the best conclusion for Angel's correct one-sided hypothesis test?

a. Full-time workers in the United States do not work more than an average of 40 hours per week.
b. Full-time workers in the United States work more than an average of 40 hours per week.
c. Angel's results do not suggest that full-time workers in the United States work more than an average of 40 hours per week.
d. Angel's results suggest that full-time workers in the United States work more than an average of 40 hours per week.

Suppose that the distribution of heart rates for medium-sized dogs is normally distributed with mean 115 beats per minute and standard deviation 18 beats per minute (bpm).

**Problem 19**.
Approximately what percentage of medium-sized dogs have a heart rate above 97 beats per minute?

a. 16%
b. 32%
c. 50%
d. 68%
e. 84%

**Problem 20**.
Approximately what percentage of medium-sized dogs have a heart rate between 88 and 144 beats per minute?

a. 34.9%
b. 65.1%
c. 94.6%
d. 59.8%

e. None of the above

Note: the final exam will only have 20 multiple choice questions but I'm adding more questions for practice.
You collect data about soda consumption from a large random sample of adults in the United States and calculate a 95% confidence interval for the average number of cans of soda consumed annually per adult in the United States to be 440 cans to 520 cans. Determine if the following statements are correct or incorrect.

**Problem 21**.
We can conclude that 95% of adults in the United States consume between 440 and 520 cans of soda per year. This statement is

a. correct.
b. incorrect.

**Problem 21**.
We are 95% confident that the population mean number of cans of soda consumed annually per adult in the United States is between 440 cans to 520 cans. This statement is

a. correct.
b. incorrect.

**Problem 22**.
There is evidence to suggest that the mean number of cans of soda consumed annually per adult in the United States is less than 500 cans because most of the values in the confidence interval are less than 500. This statement is

a. correct.
b. incorrect.

**Problem 23.**
The 95% confidence interval to estimate the mean number of cans of soda consumed annually per adult in the United States is narrower than the 99% confidence interval calculated using the same data. This statement is

a. correct.
b. incorrect.

**Problem 24.**
A random sample of students in a student success program was taken, and a 95% confidence interval for the mean GPA (grade point average) for all students in program was calculated. The resulting confidence interval was 3.44 to 3.50. Which of the following is not true?

a. The sample mean GPA was 3.47.
b. The probability that the population mean GPA for students in the student success program is in the confidence interval is 0 or 1, but we don't know which.

c. The confidence interval provides a range of plausible values for the population mean GPA for students in the student success program.
d. The confidence interval provides a range of plausible values for the sample mean GPA for students in the student success program.

**Problem 25.**
Which of the following does NOT need to be known in order to compute the p-value?

a. significance level
b. direction of the alternative hypothesis
c. value of the test statistic
d. distribution of the test statistic when the null hypothesis is true

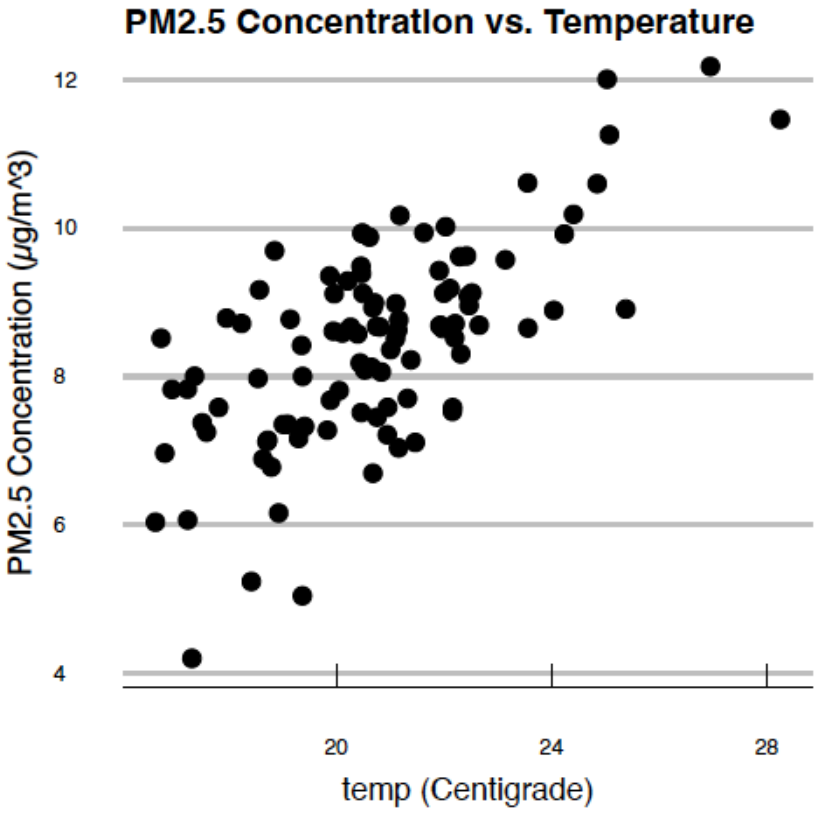
**Problem 26**.
A recent article in an educational research journal reports a correlation of +0.8 between math achievement and overall math aptitude for a large sample of students. It also reports a correlation of –0.8 between math achievement and a math anxiety test for the same group of students. Only students with scores on all three measures were included in the study. Which of the following interpretations is the most correct?

a. The correlation of +0.8 indicates a stronger a relationship than the correlation of –0.8.
b. The correlation of +0.8 is just as strong as the correlation of –0.8.
c. It is impossible to tell which correlation is stronger.

FREE RESPONSE / **FINAL EXAM** / Summer 24

For the following questions, you will not get full credit unless you show your work. Partial credit will be granted based on the work shown.

A researcher is interested in determining the relationship between temperature and the concentration of pollutants in the air. To do so, she takes temperature readings (in Celsius) and PM2.5 concentration readings (PM2.5 concentration can be viewed as a measure of air pollutants) at her house across 100 days. A scatterplot of her data is displayed below:



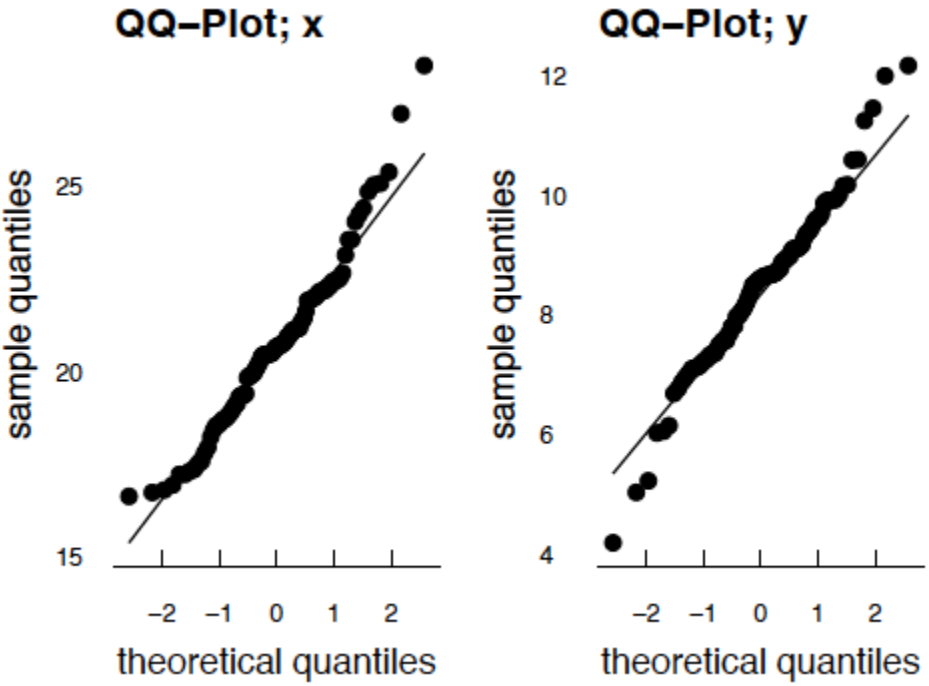

Additionally, the following numerical summaries of her data are provided:

$$\sum_{i=1}^{100} x_i = 2072.653 \quad \sum_{i=1}^{100} (x_i - \overline{x})^2 = 486.1703$$

$$\sum_{i=1}^{100} y_i = 844.7383 \quad \sum_{i=1}^{100} (y_i - \overline{y})^2 = 180.921$$

$$\sum_{i=1}^{100} (x_i - \overline{x})(y_i - \overline{y}) = 200.9171$$

Finally, below are the QQ-plots of temperature and PM2.5 Concentration, respectively:




a. Compute Cor(x,y), the correlation between x (temperature) and y (PM2.5 concentration).

b. Compute $\hat{\beta}_1$, the slope of the OLS regression line.

c. Compute $\hat{\beta}_0$, the intercept of the OLS regression line.

d. Provide an interpretation of your value of $\hat{\beta}_1$.

e. It is found that SD($\hat{\beta}_1$) = 0.0453. Construct a 95% confidence interval for β1, the slope of the true underlying linear relationship between x and y. Interpret your confidence interval.

f. Suppose on day 101 (i.e. the day right after the researcher stops recording data), the temperature at the researcher's house reaches 23 degrees celsius. What is a good estimate for the concentration of PM2.5 on day 101?

g. Is it dangerous to try and use the OLS regression line to predict the PM2.5 concentration of a day in which the temperature is 2 degrees Centigrade? Why or why not? (There is a specific word/term I'm looking for here.)

h. Do you think it is reasonable to assume the datapoints were independent from each other? Why or why not?

i. Assuming the independence assumption was reasonable (which doesn't mean that is the correct answer for the previous part!), do the normality assumptions appear to hold? Why or why not (i.e. what did you look at to answer this question)?

At an animal shelter, 55% of all cats are female. Additionally, 40% of all cats at the shelter are tabby cats. Furthermore, 25% of the tabby cats are female.

a. Define events, and translate the information provided in the problem.

b. What is the probability that a randomly selected cat is both tabby and female?

c. What is the probability that a randomly selected cat is tabby or female?

d. What is the probability that a randomly selected cat is tabby, given that it is female?

There will only be two free response questions on the exam but I am providing more for practice.

A scientist believes that the average amount of air pollution in City A is the same as the average amount of air pollution in City B. As a metric for measuring air pollution, the scientist decides to use the concentration of PM2.5 particles. She then collects 32 measurements from City A and 32 measurements from City B, and computes the following summaries:

| | Sample Mean | Sample Stnd. Dev. |
|---|---|---|
| City A | 8 | 2.5 |
| City B | 7 | 3 |

a. Classify this as either an observational study or an experiment. Explain your reasoning.

b. Which of the three sampling procedures discussed in class (simple random, stratified, and clustered) do you think the researchers used when collecting her data? Explain your reasoning.

Suppose that the scientist now wishes to statistically test her beliefs against a two-sided alternative using a 5% level of significance. Assume all normality and independence assumptions hold. Additionally, let Population 1 refer to City A and Population 2 refer to City B.

c. Define the parameters of interest, $\mu_1$ and $\mu_2$.

d. Write down the null and alternative hypotheses.

e. Compute the value of the test statistic.

f. Assuming the null is correct, what distribution does the test statistic follow? Be sure to include any / all relevant parameters.

g. What is the critical value of the test?

h. Now, conduct the test and phrase your conclusions in the context of the problem.

At a particular salad bar, a salad is created from a choice of base (lettuce, spring mix, or kale), 5 toppings (to be selected from a list of 12 possible toppings, and it is permitted to order multiple servings of the same topping), and a dressing (Caesar, balsamic vinaigrette, or ranch). A single salad (i.e. base + topping + dressing) is to be selected at random from the total set of salads that can be created.

a. If $\Omega$ denotes the outcome space of this experiment, how many elements are in $\Omega$?

b. Compute the probability that the selected salad has ranch dressing.

c. Compute the probability that the selected salad has kale and ranch dressing.