About This Practice Problem Set

This practice problem set is intended to support your preparation for the upcoming midterm exam. The questions reflect the range of difficulty you may encounter on the actual exam; some may be more challenging, while others may be less so.

Scope and Format:

- These problems are **not** representative of the overall length or coverage of the exam.
- The short-answer questions on the actual midterm will generally have **fewer parts** than those in this practice set.

Solutions and Support:

- These problems may be discussed by the TAs during review sessions in the week of October 27–31, this is subject to the wishes of the attendees.
- You may ask questions about these problems during office hours. However, please note that until October 22nd at 8:00 PM (the hard deadline for the 2nd assignment), priority will be given to Assignment 2–related questions.
- You are welcome to discuss these problems on Piazza and work on them with your classmates.

Part A – Multiple Choice

Each multiple choice question is worth 1 mark. For each question, select **one** answer only. There are no penalties for incorrect answers, so it is to your advantage to answer every question.

1. A variable representing the number of siblings a person has is best classified as:

   A. categorical

   B. **quantitative**

   C. correlated

   D. response

   E. explanatory

2. A dataset has 7 observations. When arranged in order, the 3rd value is 28 and the 4th value is 34. The median of this dataset is closest to:

   A. 28 **B.** 31 **C.** **34** **D.** 32 **E.** Cannot be determined

3. A variable has the five-number summary: 12, 20, 35, 48, 90. Using the IQR rule, what is the upper fence for detecting outliers?

   A. 48 **B.** **90** **C.** 70 **D.** 76 **E.** 78

4. A standardized test score has a z-score of -1.5. Which of the following options is closest to the percentile corresponding to this z-score?

   A. 93rd **B.** 50th **C.** **7th** **D.** 2nd **E.** 15th

5. Which of the following statements about correlation is **true**?

   A. **Correlation measures the strength and direction of a linear relationship between two variables.**

   B. Correlation is always positive.

   C. Correlation is affected by changes in units of measurement.

   D. Correlation can only be calculated for categorical variables.

   E. Correlation values can be greater than 1 or less than -1.

6. Two variables have a correlation coefficient of $r = -0.65$. If we multiply all values of one variable by 2, the new correlation will be:

   A. **-0.65**

   B. 0.65

   C. -1.35

   D. 1.35

   E. Cannot be determined

7. The purpose of a residual plot is best described as

   A. to identify the correlation coefficient

   B. **to check if the linear model is appropriate for the data**

   C. to calculate the slope of the regression line

   D. to find outliers in the original data

   E. to determine the mean of the residuals

8. Any regression line must pass through the point

   A. $(0, 0)$

   B. (s_x, s_y)

   C. (r, r)

   D. $(\overline{x}, \overline{y})$

   E. none of the above are correct

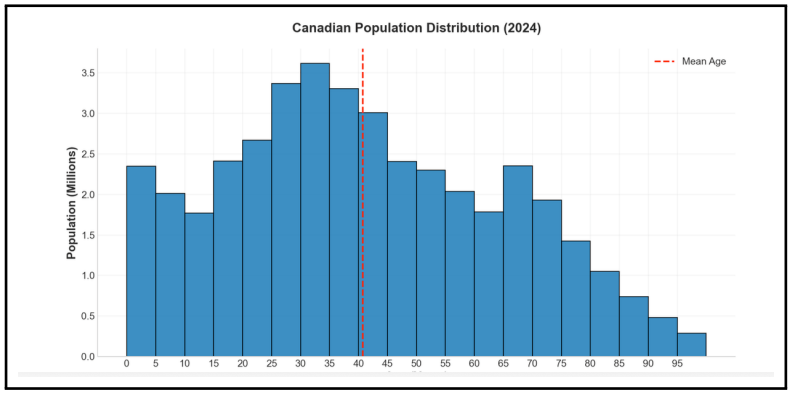
9. Consider the following stemplot where stems represent tens and leaves represent ones:

2	3 5 8
3	1 2 4 7 9
4	0 2 5
5	1

   What is the range of this dataset?

   A. 23 **B.** **28** **C.** 51 **D.** 3 **E.** 25

10. Which of the following is true regarding the Canadian population distribution from 2024?



A. It is left-skewed.

B. There are over 1 million people aged 90–100.

C. **It is right-skewed.**

D. There are several outliers.

E. It appears to be from a normal distribution.

11. A survey involving 35 questions about mental health and social media was given to a sample of 200 Gen Z individuals from a population of 10,000 users on a Discord server.

    How many variables does the data contain?

A. **35**

B. 200

C. 35×200

D. $35 \times 10{,}000$

12. If the number of observations in a dataset is even, how many data points are used to calculate the median?

A. 1 **B.** **2** **C.** 0 **D.** 3 **E.** depends on the variance of the data

13. A teacher returns an exam with possible scores ranging from 0 to 100. The teacher provides a boxplot of the exam scores. The students can get which of the following pieces of information from this plot?

A. the median

B. the IQR

C. the minimum and the maximum

D. **All of the above options are correct.**

14. What is the output of the following Python code?

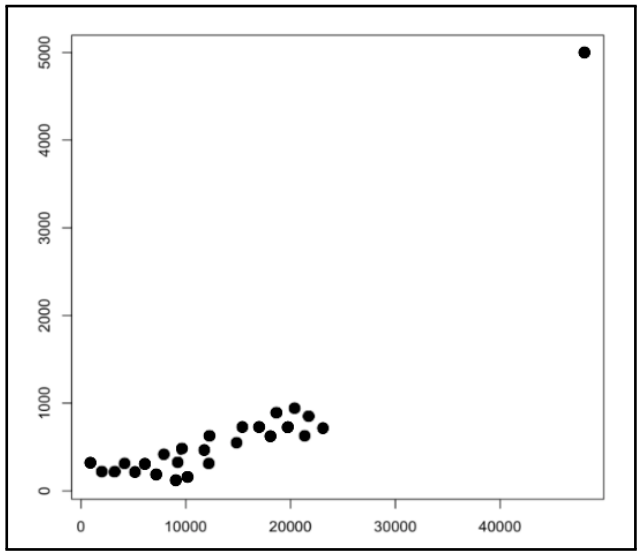
```
data = [10, 20, 30, 40, 50]
print(data[7])
```

A. 20 **B.** 10 **C.** 40 **D.** 50 **E.** **IndexError**

15. In pandas, which method returns the five-number summary ?

A. `df.5numbers()` **B.** `df.summary()` **C.** `df.fivenum()` **D.** `df.describe()` **E.** `df.5num()`

16. The following is a scatterplot for profits versus sales (in tens of thousands of dollars) for a sample of 14 large companies. The correlation between sales and profits is approximately 0.949. If we omit the point with the highest sales value, the slope of the line would



A. increase.

B. decrease.

C. change very little.

D. there is not enough information to determine the answer.

E. then be 0.

17. Which of the following Python code snippets creates a visualization?

A. `plt.stemplot()`

B. `plt.scatter([1,2,3])`

C. `plt.bargraph([1,2,3])`

D. `plt.hist([1,2,3])`

E. `plt.timeseries([1,2,3])`

18. Which of the following is the correct way to import the pandas library in Python?

A. `import pandas as pd`

B. `import pd as pandas`

C. `pandas please`

D. `from pandas import *`

E. `import pd`

19. Consider the following Python code snippet:

```
import numpy as np
import matplotlib.pyplot as plt
x = [1, 2, 3, 4, 5]
```

Which of the following code segments correctly computes the mean of the list `x` and creates an appropriate visualization to display the distribution of the data?

A.
```
print(np.mean(x))
plt.plot(x)
```

B.
```
print(mean(x))
plt.hist(x)
```

C.
```
print(np.average(x))
plt.scatter(x)
```

D.
```
print(np.mean(x))
plt.hist(x)
```

E.
```
print(np.median(x))
plt.hist(x)
```

20. Which of the following code snippets correctly filters a dataframe `df` to show only rows where `age` is less than 30?

A. `df[df['age'] < 30]`

B. `df.filter(age < 30)`

C. `df.query['age' < 30]`

D. `df[age < 30]`

E. `df.select(df['age'] < 30)`

21. A regression equation is: $\hat{y} = 100 + 25x$. Which statement is correct?

A. For each unit increase in x, y increases by 25 units on average

B. For each unit increase in y, x increases by 100 units on average

C. The intercept is the predicted value of y when $x = \bar{x}$

D. The slope is 100

E. $R^2 = 5$

22. A study of 200 students recorded both their major (Science or Arts) and whether they own a laptop (Yes or No). The table shows:

	Science	Arts	Total
Laptop: Yes	60	80	140
Laptop: No	40	20	60
Total	100	100	200

What proportion of students don't own a laptop?

A. **0.30**

B. 0.43

C. 0.50

D. 0.60

E. 0.70

23. In a class of 100 students, the grades on an accounting test are summarized in the following frequency table:

Grade	**Frequency**
91–100	11
81–90	31
71–80	42
61–70	16

The first quartile would fall into what grade range:

A. 91–100

B. 81–90

C. **71–80**

D. 61–70

E. none of the above

24. The exam scores (out of 100 points) for all students taking an introductory statistics course are examined. If 5 points were added to each score, then standard deviation of the new scores would:

A. be increased by 5.

B. be increased by 25.

C. be decreased by 5.

D. **remain unchanged.**

E. be 0.

25. Scores on the SAT math test in recent years approximately follow a normal distribution with a mean of 515 and a standard deviation of 109. The proportion of students scoring between 460 and 550 is closest to:

A. 0.309 **B.** **0.317** **C.** 0.626 **D.** 0.681 **E.** 0.223

26. A company produces boxes of soap powder labeled "Giant Size: 32 Ounces." The actual weight of soap powder in such a box has a Normal distribution, with a mean of 33 oz and a standard deviation of 0.7 oz. To avoid losing money, the company labels the top 5% (the heaviest 5%) overweight. How heavy does a box have to be for it to be labeled overweight?

A. 31.60 oz **B.** 31.85 oz **C.** **34.15 oz** **D.** 34.40 oz **E.** cannot be determined

27. You recently took a statistics exam in a large class. The instructor tells the class that the scores were Normally distributed, with a mean of 72 (out of 100) and a standard deviation of 12. Your score was 90. Your friend took the same statistics course but with another instructor. Your friend had a score of 75 on a test, where the test had a mean of 60 and a standard deviation of 10. What can you conclude from this comparison?

A. You clearly ranked better.

B. **You and your friend ranked equally well.**

C. Your friend actually ranked better.

D. Nothing; the tests cannot be compared.

E. You got the highest score in the class.

28. A least squares regression of profits y (in tens of thousands of dollars) versus sales x (in tens of thousands of dollars) for a sample of 14 large companies is fit and gives the following information:

- For 13 companies: sales range from 0 to 20, profits range from 0 to 10
- For 1 company: sales = 49, profit = 5
- Correlation for all 14 companies: $r = 0.949$

The company with sales $x = 49$ would be considered:

A. an outlier in the y direction only

B. an outlier in both x and y directions

C. **a high leverage point**

D. an influential point

E. a resistant point

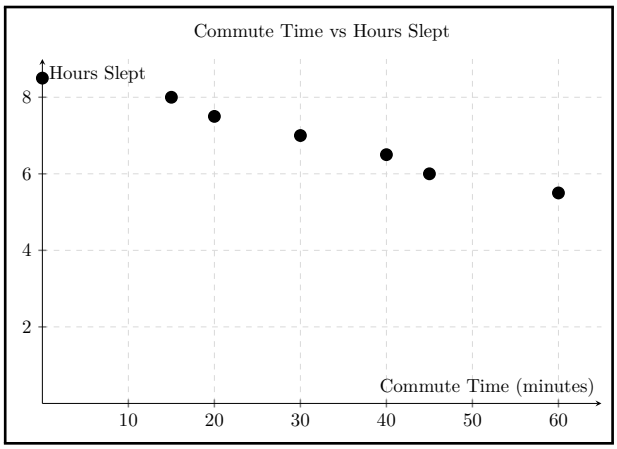
29. Given the following information:

$$\overline{x} = 15, \quad \overline{y} = 20, \quad s_x = 1, \quad s_y = 2, \quad r = 0.95$$

What is the intercept of the least-squares regression line:

A. 48.5 **B.** 1.9 **C.** **-8.5** **D.** 12.875 **E.** cannot be determined

30. The correlation between hours slept and commute time displayed in the scatterplot below is approximately:



A. -1

B. 0.9

C. 0.81

D. **-0.9**

E. 0.33

31. Consider a (random) variable that follows a normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 2$. Which of the following Python expressions correctly calculates the proportion of observations of the variable that are greater than or equal to 0.5?

A. `norm.cdf(0.5, loc = 1, scale = 2)`

B. `1 - norm.cdf(0.5, loc = 1, scale = 2)`

C. `norm.ppf(0.5, loc = 1, scale = 2)`

D. `1 - norm.ppf(0.5, loc = 1, scale = 2)`

E. `norm.prob(0.5, loc = 1, scale = 2)`

32. A study investigates whether smoking is associated with lung cancer. The 2x2 table below shows the number of individuals who either developed lung cancer (Yes) or did not develop lung cancer (No), with or without a smoking history.

Lung Cancer	**Yes**	**No**	**Total**
Smokers	70	120	190
Non-Smokers	30	180	210
Total	100	300	400

Calculate the relative risk of developing lung cancer for a smoker versus a non-smoker.

A. 37% **B.** 18% **C.** **2.6** **D.** 2.3 **E.** -1.3

33. What does `print()` do in Python?

A. It uses the printer to print a document.

B. It creates a plot.

C. It imports a library.

D. It sorts a list.

E. **It displays output to the console.**

34. Which of the following is not a Python module that we have worked with?

A. pandas

B. numpy

C. **datascience**

D. matplotlib

E. seaborn

35. Which of the following values is not returned by the pandas method `describe()`?

A. Mean **B.** Median **C.** **Mode** **D.** Maximum **E.** Standard deviation

Part B – Short Answer

Question 1

A fitness instructor recorded the weekly gym visits for a sample of 10 clients:

1	3	3	3	4	5	5	6	6	14

(a) Find the five-number summary for this dataset.

Solution: **Five-number summary:**

- Minimum = 1
- First quartile (Q1) = 3
- Median = 4.5
- Third quartile (Q3) = 6
- Maximum = 14

(b) Determine if any potential outliers are present.

Solution: The IQR is given by

$$\text{IQR} = Q3 - Q1 = 6 - 3 = 3,$$

and hence the fences are:

$$\text{Lower fence} = Q1 - 1.5 \times \text{IQR} = 3 - 1.5(3) = -1.5$$
$$\text{Upper fence} = Q3 + 1.5 \times \text{IQR} = 6 + 1.5(3) = 10.5$$

Note that

- All values except 14 fall within the range $[-1.5, 10.5]$
- Since $14 > 10.5$, the value 14 exceeds the upper fence

The value 14 is a potential outlier.

(c) Draw the modified boxplot for this data

Solution: The modified boxplot should display the following features:

- A box spanning from Q1 (3) to Q3 (6), with a line at the median (4.5)
- A left whisker extending from Q1 (3) to the minimum value (1)
- A right whisker extending from Q3 (6) to the largest value within the upper fence (6)
- The value 14 plotted as an individual point beyond the right whisker, indicating an outlier

Note: In a modified boxplot, whiskers extend to the most extreme data points that are not outliers. Since 14 exceeds the upper fence (10.5), it is marked separately as an outlier. All other values fall within the fences, so the left whisker reaches the minimum (1) and the right whisker reaches the maximum non-outlier value (6).

Question 2

A sample of daily step counts (in thousands) for four office workers was collected using a pedometer. The observed step counts are:

6	7	9	8

(a) Name two graphical displays that would be appropriate for visualizing this variable.

Solution: Any two of the following would be appropriate:

- Histogram
- Boxplot
- Stemplot

(b) Find the mean and standard deviation of this sample.

Solution: The mean is given by

$$\overline{x} = \frac{6+7+9+8}{4} = \frac{30}{4} = \boxed{7.5 \text{ thousand steps}}$$

The standard deviation is given by

$$\begin{aligned} s &= \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}} \\ &= \sqrt{\frac{(6-7.5)^2 + (7-7.5)^2 + (9-7.5)^2 + (8-7.5)^2}{3}} \\ &= \sqrt{\frac{2.25 + 0.25 + 2.25 + 0.25}{3}} \\ &= \sqrt{\frac{5}{3}} \approx \boxed{1.29 \text{ thousand steps}} \end{aligned}$$

(c) It turned out that the pedometer was malfunctioning and undercounted each worker's steps by 500. If we add 0.5 to each observation, find the new mean and standard deviation.

Solution: When we add a constant to each observation,

- The mean increases by that constant
- The standard deviation remains unchanged

Therefore, the new mean is $\overline{x}_{\text{new}} = 7.5 + 0.5 = \boxed{8.0}$.

The new standard deviation is $s_{\text{new}} = \boxed{1.29}$ thousand steps (unchanged).

Question 3

The five-number summary for the exam scores is:

- Minimum: 45
- First Quartile (Q1): 50
- Median: 75
- Third Quartile (Q3): 85
- Maximum: 98

(a) Calculate the IQR and determine the lower and upper fences for outlier detection.

Solution: IQR calculation:

$$\text{IQR} = Q3 - Q1 = 85 - 50 = 35$$

Fences:

- Lower fence: $Q1 - 1.5 \times \text{IQR} = 50 - 1.5(35) = 50 - 52.5 = \boxed{-2.5}$
- Upper fence: $Q3 + 1.5 \times \text{IQR} = 85 + 1.5(35) = 85 + 52.5 = \boxed{137.5}$

(b) Are there any potential outliers in this dataset? Justify your answer.

Solution: There are no potential outliers in this dataset. To see this, note that

- Minimum (45) $>$ Lower fence (-2.5)
- Maximum (98) $<$ Upper fence (137.5)

Hence, all values lie in the inner fence.

(c) Is the distribution symmetric? Justify your answer.

Solution: This distribution is not symmetric. To see this, note that

- Distance from median to Q1:

$$75 - 50 = 25$$

- Distance from median to Q3:

$$85 - 75 = 10$$

Since 25 is considerably larger than 10 (with respect to the scale of our data), the distribution is not symmetric.

Question 4

A particular brand of LED light bulb comes with a warranty for 800 hours of use. It is known that the life of such a bulb has a normal distribution with a mean of 1,440 hours and a standard deviation of 240 hours.

(a) Compute the z-score corresponding to the warranty period. Explain what this z-score indicates about how the warranty compares to the average bulb lifespan.

Solution:

$$z = \frac{x - \mu}{\sigma} = \frac{800 - 1440}{240} = \frac{-640}{240} = \boxed{-2.67}$$

This z-score of -2.67 indicates that the warranty period of 800 hours falls 2.67 standard deviations below the mean bulb lifespan. In other words, the warranty threshold is set well below the average lifespan, meaning the company will replace only those bulbs whose lifespans are at least 2.67 standard deviations below the mean.

(b) What percentage of bulbs will last through the entire warranty period without failing?

Solution: We need $P(X > 800)$:

$$P(X > 800) = P(Z > -2.67) = 1 - P(Z < -2.67) = 1 - 0.0038 = \boxed{0.9962}$$

Approximately 99.62% of bulbs will last through the warranty period.

(c) Find the probability that a randomly selected bulb will have a lifespan between 1,200 and 1,800 hours.

Solution: We need to find $P(1200 < X < 1800)$.

First, we calculate the z-scores for both values:

$$z_1 = \frac{1200 - 1440}{240} = \frac{-240}{240} = -1$$

$$z_2 = \frac{1800 - 1440}{240} = \frac{360}{240} = 1.5$$

Thus,

$$\begin{aligned} P(1200 < X < 1800) &= P(-1 < Z < 1.5) \\ &= P(Z < 1.5) - P(Z < -1) \\ &= 0.9332 - 0.1587 \\ &= \boxed{0.7745} \end{aligned}$$

Approximately 77.45% of bulbs will have lifespans between 1,200 and 1,800 hours.

(d) The company wants to set a warranty period such that only 10% of bulbs will need to be replaced under warranty. How many hours should the warranty period be?

Solution: We need to find x such that $P(X < x) = 0.10$.

From the standard normal table, $P(Z < -1.28) = 0.10$.

$$-1.28 = \frac{x - 1440}{240}$$

$$\implies x = 1440 + (-1.28)(240) = 1440 - 307.2 = \boxed{1132.8}$$

The warranty should be set at approximately 1132.8 hours.

(e) Determine the two lifespan values such that the middle 75% of all bulbs will have lifespans between these two values.

Solution: To find the middle 75% of bulbs, we need to determine the values that leave 12.5% in each tail of the distribution.

From the standard normal table:

- $P(Z < -1.15) = 0.125$ (lower tail)
- $P(Z < 1.15) = 0.875$ (upper tail)

Converting these z-scores to lifespan values:

Lower bound:

$$x_1 = \mu + z_1 \cdot \sigma = 1440 + (-1.15)(240) = 1440 - 276 = 1164 \text{ hours}$$

Upper bound:

$$x_2 = \mu + z_2 \cdot \sigma = 1440 + (1.15)(240) = 1440 + 276 = 1716 \text{ hours}$$

$$\boxed{\text{The middle 75\% of bulbs will have lifespans between 1,164 and 1,716 hours.}}$$

Question 5

A study examined the relationship between study hours per week (x) and final exam scores (y) for 500 students. The least-squares regression line was found to be:

$$\hat{y} = 40 + 6x$$

Additional information: $\overline{x} = 10$ hours, $s_x = 3$ hours, $\overline{y} = 100$ points, and $r = 0.8$.

(a) Using the regression equation and the given information, calculate s_y, the standard deviation of y.

Solution: We use the relationship between the slope, correlation, and standard deviations:

$$b = r \cdot \frac{s_y}{s_x}$$

From the regression equation, we know $b = 6$. Substituting the known values:

$$6 = 0.8 \cdot \frac{s_y}{3}$$

Solving for s_y

$$s_y = \frac{6 \times 3}{0.8} = \frac{18}{0.8} = \boxed{22.5}$$

(b) What is the coefficient of determination (R^2) for this regression?

Solution: The coefficient of determination is the square of the correlation coefficient:

$$R^2 = r^2 = (0.8)^2 = 0.64$$

$$\boxed{R^2 = 0.64}$$

(c) Interpret the R^2 value in the context of this problem.

Solution: An R^2 value of 0.64 indicates that 64% of the variation in final exam scores can be explained by the linear relationship with study hours per week. The remaining 36% of variation is due to other factors not captured by this model.

(d) Interpret the slope of the regression line in context.

Solution: The slope of 6 means that for each additional hour per week a student studies, the final exam score is expected to increase by 6 points on average.

(e) A student who studied 15 hours per week scored 85 on the exam. Calculate and interpret the residual for this student.

Solution: First, we calculate the predicted score using the regression equation and find

$$\hat{y} = 40 + 6(15) = 40 + 90 = 130$$

As the residual is given by the actual score minus the predicted score, we have

$$e = y - \hat{y} = 85 - 130 = \boxed{-45}$$

This negative residual indicates that the student scored 45 points lower than predicted by the regression model. In other words, the student underperformed by 45 points relative to what the model predicted based on their study hours.

Question 6

A university is comparing the success rates of two learning strategies (Active Recall and Rote Memorization) across two different class sizes.

For small classes (20 students or fewer):

	Active Recall	Rote Memorization
Pass	9	72
Fail	1	18
Total	10	90

For large classes (more than 20 students):

	Active Recall	Rote Memorization
Pass	63	6
Fail	27	4
Total	90	10

(a) Calculate the pass rates for each strategy within each class size. Which strategy has a higher pass rate in small classes? Which strategy has a higher pass rate in large classes?

Solution: **Small classes:**

- Active Recall: $\frac{9}{10} = 0.90$ or 90%
- Rote Memorization: $\frac{72}{90} = 0.80$ or 80%

Active Recall has a higher pass rate in small classes.

Large classes:

- Active Recall: $\frac{63}{90} = 0.70$ or 70%
- Rote Memorization: $\frac{6}{10} = 0.60$ or 60%

Active Recall has a higher pass rate in large classes.

(b) Obtain the overall pass rates for each strategy by using the following table:

	Active Recall	Rote Memorization
Pass	72	78
Fail	28	22
Total	100	100

Which strategy appears better overall?

Solution: **Overall pass rates:**

- Active Recall: $\frac{72}{100} = 0.72$ or 72%
- Rote Memorization: $\frac{78}{100} = 0.78$ or 78%

Rote Memorization appears better overall with a 78% pass rate compared to Active Recall's 72% pass rate.

(c) Is this an example of Simpson's Paradox? Explain your reasoning.

Solution: Yes, this is Simpson's Paradox. Active Recall has a higher pass rate in both small classes (90% vs 80%) and large classes (70% vs 60%), but when the data is combined, Rote Memorization appears to have a higher overall pass rate (78% vs 72%). This reversal occurs because class size acts as a lurking/confounding variable (depending on if we are assuming it is available at the time of analysis) that affects both the choice of strategy and the pass rates.

(d) Find the probability of passing the course given that a student used Rote Memorization.

Solution:

$$P(\text{Pass}|\text{Rote Memorization}) = \frac{\text{Number who passed using Rote Memorization}}{\text{Total using Rote Memorization}} = \frac{78}{100} = 0.78$$

(e) Find the relative risk of passing the course when using Rote Memorization compared to Active Recall.

Solution:

$$\text{Relative Risk} = \frac{P(\text{Pass}|\text{Rote Memorization})}{P(\text{Pass}|\text{Active Recall})} = \frac{0.78}{0.72} = 1.083$$

Students using Rote Memorization are approximately 1.08 times as likely to pass compared to those using Active Recall (based on overall rates).

Question 7

(a) Explain why the median is a resistant measure of center while the mean is not. Provide a specific example to illustrate your answer.

Solution: The median is resistant because it is not affected by extreme values (outliers), while the mean is not resistant because it uses all values in its calculation and can be heavily influenced by outliers.

Example: Consider the dataset: 1, 2, 3, 4, 5

- Mean = 3, Median = 3

Now change the last value to 100: 1, 2, 3, 4, 100

- Mean = 22 (drastically changed)
- Median = 3 (unchanged)

The median remains at 3 while the mean jumps to 22, demonstrating the median's resistance to outliers.

(b) How can we find the shape of a distribution from a stemplot?

Solution: To determine the shape of a distribution from a stemplot, rotate the stemplot 90 degrees counterclockwise so that it resembles a histogram. The stems become the horizontal axis and the leaves stack to show frequency. From this view, we can identify the shape based on the tails as we would for a histogram (or a density plot).

(c) What is a scatterplot?

Solution: A scatterplot is a graphical display that shows the relationship between two quantitative variables by plotting ordered pairs (x, y) as points on a coordinate plane. The horizontal axis represents one variable (explanatory) and the vertical axis represents the other variable (response).

(d) What is the ecological fallacy?

Solution: The ecological fallacy is the error of drawing conclusions about individuals based solely on group-level data. It occurs when we incorrectly assume that relationships observed at the group level necessarily hold for individuals within those groups.

For example, concluding that because a city with higher average income has lower crime rates, that wealthy individuals commit fewer crimes (when the relationship might be due to other group-level factors).

(e) Define lurking variable and provide an example.

Solution: A lurking variable is a variable that is not included in a study but affects both the explanatory and response variables, potentially creating a misleading association between them.

Example: A study finds that ice cream sales and drowning deaths are positively correlated. However, temperature is a lurking variable: warm weather increases both ice cream sales and swimming activity (which increases drowning risk).

Both variables are influenced by the lurking variable (temperature) and hence we cannot conclude causality.

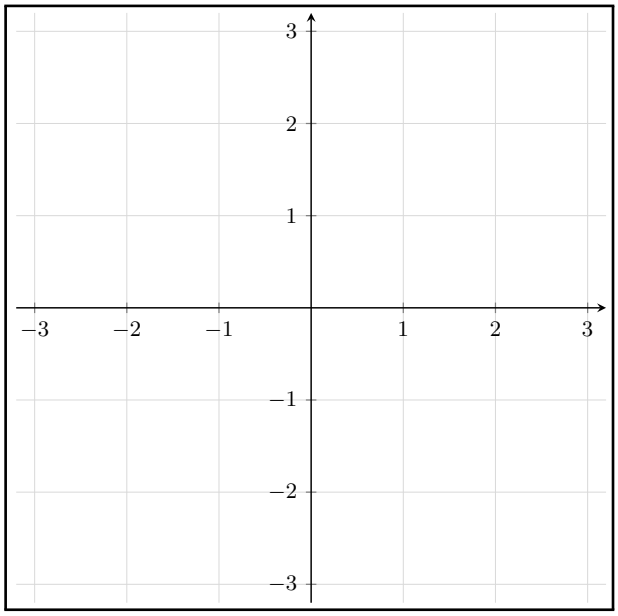
(f) What is Simpson's paradox?

Solution: Simpson's Paradox occurs when a trend that appears in different groups of data disappears or reverses when the groups are combined. This happens when a lurking variable (often a grouping variable) affects the relationship between two variables differently across subgroups. The overall aggregate data can show the opposite pattern from what is observed in each individual subgroup.

Question 8

(a) Plot the following data on a scatterplot.

x	y
-1	-2
0	-1
1	-1



Solution: Plot the three points: $(-1,-2)$, $(0,-1)$, and $(1,-1)$ on the coordinate plane.

(b) Calculate the correlation of x and y.

Solution: First, we calculate the means

$$\overline{x} = \frac{-1+0+1}{3} = 0, \quad \overline{y} = \frac{-2+(-1)+(-1)}{3} = \frac{-4}{3}$$

Next, we find the standard deviations

$$s_x = \sqrt{\frac{(-1-0)^2+(0-0)^2+(1-0)^2}{2}} = \sqrt{\frac{2}{2}} = 1$$

$$s_y = \sqrt{\frac{(-2+\frac{4}{3})^2+(-1+\frac{4}{3})^2+(-1+\frac{4}{3})^2}{2}}$$

$$= \sqrt{\frac{\frac{4}{9}+\frac{1}{9}+\frac{1}{9}}{2}} = \sqrt{\frac{6}{18}} = \sqrt{\frac{1}{3}} = \frac{1}{\sqrt{3}}$$

Calculate correlation:

$$r = \frac{1}{n-1}\sum \frac{(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y}$$

$$= \frac{1}{2}\left[\frac{(-1)(-\frac{2}{3})}{1\cdot\frac{1}{\sqrt{3}}} + \frac{(0)(\frac{1}{3})}{1\cdot\frac{1}{\sqrt{3}}} + \frac{(1)(\frac{1}{3})}{1\cdot\frac{1}{\sqrt{3}}}\right]$$

$$= \frac{1}{2}\left[\frac{2\sqrt{3}}{3} + 0 + \frac{\sqrt{3}}{3}\right]$$

$$= \frac{1}{2}\cdot\frac{3\sqrt{3}}{3}$$

$$= \boxed{\frac{\sqrt{3}}{2} \approx 0.866}$$

(c) Calculate the correlation of x and x.

Solution: The correlation of any variable with itself is always 1. This is because the variable is perfectly linearly related to itself (every point lies exactly on the line $y = x$).