PSTAT 5A Practice Worksheet 2 Solutions

Descriptive Statistics

Instructor Solutions

2025-07-23

Table of contents

1 Section A: Basic Descriptive Statistics

1.1 Problem A1: Mean and Standard Deviation

Given:

- 24 students: average = 74, standard deviation = 8.9
- 1 makeup student: score = 64

1.1.1 Part (a): Does the new student's score increase or decrease the average score?

Solution: DECREASE

Since $64 < 74$ (the current average), adding this score will pull the average down.

1.1.2 Part (b): What is the new average?

Solution:

```
New average = (Sum of all 25 scores) / 25
Sum of first 24 scores = 24 × 74 = 1,776
Total sum = 1,776 + 64 = 1,840
New average = 1,840 / 25 = 73.6 points
```

1.1.3 Part (c): Does the new student's score increase or decrease the standard deviation?

Solution: INCREASE

The score of 64 is more than one standard deviation below the original mean (74 - 8.9 = 65.1). This adds more variability to the dataset, increasing the standard deviation.

1.2 Problem A2: Distribution Shape Analysis

Given: - TV watching hours per week - Mean = 4.71 hours - Standard deviation = 4.18 hours

Question: Is the distribution symmetric? What shape? Explain reasoning.

Solution: NOT SYMMETRIC - RIGHT-SKEWED

Reasoning: 1. The standard deviation (4.18) is nearly as large as the mean (4.71)

2. Since hours cannot be negative, there's a natural lower bound at 0
3. Some students likely watch much more TV than others, creating a long right tail
4. The large standard deviation relative to the mean suggests high variability
5. In a right-skewed distribution, a few high values (heavy TV watchers) pull the mean higher

2 Section B: Data Interpretation and Graphical Analysis

2.1 Problem B1: Interpreting Histograms

Context: Infant mortality histogram shows right-skewed distribution with:

- Highest bar at 0-10 range (about 38% of countries)
- Decreasing bars: 10-20 (23%), 20-30 (11%)
- Long right tail with few countries having high rates

2.1.1 Part (a): Estimate Q1, the median, and Q3 from the histogram.

Solution: Looking at cumulative percentages:

- **Q1 (25th percentile)** 8 deaths per 1,000 live births
- **Median (50th percentile)** 15 deaths per 1,000 live births
- **Q3 (75th percentile)** 35 deaths per 1,000 live births

2.1.2 Part (b): Would you expect the mean to be smaller or larger than the median? Explain.

Solution: MEAN > MEDIAN

Reasoning: - The distribution is right-skewed

- The long right tail contains countries with very high infant mortality rates
- These extreme values pull the mean higher than the median
- In right-skewed distributions, the mean is always greater than the median
- The median is resistant to outliers, but the mean is affected by them

2.2 Problem B2: Comparing Distributions

Based on the plots showing Gain vs No Gain counties:

2.2.1 Center:

- **Gain group** has higher median household income (~$55,000)
- **No Gain group** has lower median household income (~$45,000)

2.2.2 Variability:

- **Gain group** shows less variability (tighter distribution)
- **No Gain group** shows greater variability (wider spread)

2.2.3 Shape:

- Both groups are right-skewed
- Shape is relatively consistent between groups
- Both have longer right tails

2.2.4 Modes:

- Each group has one prominent mode
- **Gain group:** mode around $50,000-$55,000
- **No Gain group:** mode around $40,000-$45,000

3 Section C: Variance Calculations Practice

3.1 Problem C1: Basic Variance Calculations

Data: 3, 7, 2, 8, 5, 6, 4, 9

3.1.1 Part (a): Calculate the sample mean $\bar{x}$.

Solution:

```
x̄ = (3 + 7 + 2 + 8 + 5 + 6 + 4 + 9) / 8
x̄ = 44 / 8 = 5.5
```

3.1.2 Part (b): Calculate the sample variance s^2 using (n-1).

Solution:

```
s² = Σ(xi - x̄)² / (n-1)

Deviations from mean:
(3-5.5)² = (-2.5)² = 6.25
(7-5.5)² = (1.5)² = 2.25
(2-5.5)² = (-3.5)² = 12.25
(8-5.5)² = (2.5)² = 6.25
(5-5.5)² = (-0.5)² = 0.25
(6-5.5)² = (0.5)² = 0.25
(4-5.5)² = (-1.5)² = 2.25
(9-5.5)² = (3.5)² = 12.25

Sum = 6.25 + 2.25 + 12.25 + 6.25 + 0.25 + 0.25 + 2.25 + 12.25 = 42

s² = 42 / (8-1) = 42 / 7 = 6.0
```

3.1.3 Part (c): Calculate the sample standard deviation s.

Solution:

```
s = √s² = √6 = 2.4495
```

3.1.4 Part (d): Population variance ^{2} if treated as complete population.

Solution:

```
 ² = Σ(xi -  )² / N
 ² = 42 / 8 = 5.25
```

3.1.5 Part (e): Why divide by (n-1) for sample variance instead of n?

Solution: We use (n-1) because of degrees of freedom. When we use the sample mean $\bar{x}$ to calculate deviations, we "use up" one degree of freedom. The sample mean constrains the data - if we know (n-1) deviations and the sample mean, the last deviation is determined. This makes s^2 an unbiased estimator of the population variance ^{2}.

3.2 Problem C2: Comparing Variability

Data Sets: - Set A: 10, 12, 14, 16, 18 - Set B: 5, 10, 14, 18, 23

3.2.1 Part (a): Calculate the mean for each set.

Solution:

```
Set A: x̄_A = (10 + 12 + 14 + 16 + 18) / 5 = 70 / 5 = 14
Set B: x̄_B = (5 + 10 + 14 + 18 + 23) / 5 = 70 / 5 = 14
```

3.2.2 Part (b): Calculate the sample variance for each set.

Solution:

Set A:

```
Deviations: (10-14)²=16, (12-14)²=4, (14-14)²=0, (16-14)²=4, (18-14)²=16
Sum = 16 + 4 + 0 + 4 + 16 = 40
s²_A = 40 / (5-1) = 40 / 4 = 10
```

Set B:

```
Deviations: (5-14)²=81, (10-14)²=16, (14-14)²=0, (18-14)²=16, (23-14)²=81
Sum = 81 + 16 + 0 + 16 + 81 = 194
s²_B = 194 / (5-1) = 194 / 4 = 48.5
```

3.2.3 Part (c): Which set has greater variability?

Solution: SET B has greater variability

Set B has variance = 48.5 vs Set A variance = 10

3.2.4 Part (d): Calculate coefficient of variation for each set. Which has greater relative variability?

Solution:

```
CV_A = s_A / x̄_A = √10 / 14 = 3.162 / 14 = 0.2259
CV_B = s_B / x̄_B = √48.5 / 14 = 6.964 / 14 = 0.4974
```

SET B has greater relative variability (CV_B = 0.4974 > CV_A = 0.2259)

Note: The coefficient of variation measures variability relative to the mean, making it useful for comparing datasets with different units or scales.
