Chapter 3 – Numerical Summary Measures

Important features of a numerical data set

- Shape of the distribution – see Chapter 2
- Center of the distribution
- Spread of the distribution

Measures of Center (or central tendency)

Goal – calculate a single number that measures or identifies the middle of a data set … or … identifies the value of a typical observation

- **Sample mean**
    - Let $x_1, x_2, x_3, \ldots, x_n$ denote the n data values in a sample
    - $\bar{x} = \dfrac{x_1 + x_2 + x_3 \cdots + x_n}{n} = \dfrac{\sum x_i}{n}$

- **Sample median**
    - Sort the data values from smallest to largest
    - $\tilde{x} = \begin{cases} \text{middle value} & \text{if } n \text{ is odd} \\ \text{average of middle pair} & \text{if } n \text{ is even} \end{cases}$

- **Sample Mode**
    - M = data value occurring most often
    - If two (or three) values each occur most often the dist'n is bimodal (or trimodal).
    - If every value occurs equally often then the mode does not exist.

Example – A random sample of $n = 12$ tractor trailers was selected from a particular stretch of highway and their speeds recorded. Here are the data

65, 79, 60, 67, 71, 83, 61, 67, 64, 77, 69, 74

(a) Calculate the sample mean ($\bar{x}$), sample median ($\tilde{x}$), and the sample mode (M).

(b) Remove 83 and recalculate $\tilde{x}$.

(c) Change 60 to 61 and recalculate M.

(d) Change one copy of 67 to 68 and recalculate M.

Important Facts

- The sample mean ($\bar{x}$) may not be a good measure of center because … why?

- Comparing the mean ($\bar{x}$) and median ($\tilde{x}$) can reveal the shape of a distribution. When
  - … $\bar{x} > \tilde{x}$, the distribution is likely to have what shape?
  - … $\bar{x} < \tilde{x}$, the distribution is likely to have what shape?
  - … $\bar{x} \approx \tilde{x}$, the distribution is likely to have what shape?

  Question – The distribution of tractor trailer speeds is possibly … what shape?

- $\bar{x}$ = mean of a sample ($\bar{x}$ is an example of a **sample statistic**)
  μ = mean of the entire population (μ is an example of a **population parameter**)
  When μ is unknown we often use $\bar{x}$ as an estimate (or prediction) of μ .

Measures of Spread (or variability)

Goal – calculate a single number that measures the variability among the data values… or … measures how far apart the data values are from each other

For measures of spread, if the answer is …

- … large, then the data is more spread out (has more variability).
- … small, then the data is less spread out (has less variability).

- **Sample range** R = max – min

- **Sample variance**
  - (Definition) $s^2 = \dfrac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1} = \dfrac{\sum (x_i-\bar{x})^2}{n-1}$
  - (Computational or short cut) $s^2 = \dfrac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}{n-1}$

- **Sample standard deviation** $s = \sqrt{s^2}$

- **Interquartile range** $IQR = Q_3 - Q_1$

  Where

  Q_3 = 3rd quartile = 75th percentile
  = median of the upper half of the data

  Q_1 = 1st quartile = 25th percentile
  = median of the lower half of the data

Example – A random sample of $n = 12$ tractor trailers was selected from a particular stretch of highway and their speeds recorded. Here are the data

60, 61, 64, 65, 67, 67, 69, 71, 74, 77, 79, 83

(a) Find the sample range.

(b) Find the first quartile, median, third quartile, and the interquartile range.

(c) Suppose there was a 13th tractor trailer in the sample with x_{13}= 85. Re-do part (b).

Important Fact – Comparing the first quartile (Q_1), the median ($\tilde{x}$), and the third quartile (Q_3) can reveal the shape of a distribution. When …

- … Q_1 is closer to $\tilde{x}$, the distribution is likely to have what shape?

- … Q_3 is closer to $\tilde{x}$, the distribution is likely to have what shape?

- … Q_1 and Q_3 are equally close to $\tilde{x}$, the distribution is likely to have what shape?

Question – The distribution of tractor trailer speeds is possibly … what shape?

Example – Consider the following data 3, 5, 9, 12, 16, 21

(a) Find the sample variance using the definition.

(b) Find the sample variance using the computational (or short cut) formula.

(c) Find the sample standard deviation.

(d) Suppose that the data had units of inches. What would be the units associated with the mean, variance, and standard deviation?

Important Facts

- Standard deviation is often thought of as measuring the typical distance that an observation is from the mean

- Standard deviation is often used as a ruler for judging distances.
  Example – Suppose that Bob took exams in his English and Math classes. His scores together with the class summaries are below. In which class did Bob do better, relative to his classmates?

	English	Math
Bob	80	85
Mean	75	80
Standard Deviation	2.5	5

- s^2 = variance of a sample (s^2 is an example of a sample statistic)
  σ^2 = variance of the entire population (σ^2 is an example of a population parameter)
  When σ^2 is unknown we often use s^2 as an estimate (or prediction) of σ^2.

The mean and the standard deviation can be used together to describe the distribution of a data set more precisely.

Empirical Rule

- Applies only when the shape of the distribution is approximately normal (bell-shaped)
- Approximately 68% of the observations are within 1 standard deviation of the mean
  Approximately 95% of the observations are within 2 standard deviations of the mean
  Approximately 99.7% of the observations are within 3 standard deviations of the mean

Example – In a random sample of people the mean height was $\bar{x} = 66$ inches with a standard deviation of $s = 4$ inches. Assuming the distribution of heights is normal, answer the following.

(a) 68% of the people have heights between _____ & _____.

(b) ____% of the people have heights between 54 and 78.

(c) 95% of the people have heights between _____ & _____.

(d) What percent of the heights is either less than 58 or greater than 74?

(e) What percent of the heights is greater than 78?

(f) What percent of the heights is between 58 and 62?

(g) What percent of the heights is between 62 and 78?

Chebyshev's Rule

- Applies to any data set, regardless of shape
- (Provided $k > 1$) At least $100 \cdot \left(1 - \frac{1}{k^2}\right)\%$ of the data set is within k standard deviations of the mean – i.e. lying in the interval $(\bar{x} - k \cdot s, \bar{x} + k \cdot s)$

Example – In a random sample of people the mean height was $\bar{x} = 66$ inches with a standard deviation of $s = 4$ inches. Without assuming anything about the shape of the distribution of heights, answer the following.

(a) What percent of heights is between 58 and 74?
How does this compare to using the Empirical Rule?

(b) What percent of heights is either less than 58 or greater than 74?

(c) What percent of heights is between 54 and 78?

(d) What percent of heights is between 60 and 72?

(e) At least 84% of the heights is between _____ & _____.