CH 3: Descriptive Statistics: Numerical Measures Part 1

1. Measure of Locations

(A) Observation Notation \( x_i \): the \( i \)th observation in the list of observations.

(B) Summation Notation \( \sum \) (“Sigma”–Computing the sum):
We write \( \sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \)

(C) Sample Mean (Notation: \( \bar{x} \))
\[
\bar{x} = \frac{\sum x_i}{n} \quad (\text{eq} 3.1)
\]
EX 1 Given a set of data with \( n = 5 \) (the birth weights): 9.2, 6.4, 10.5, 8.1, 7.8. Find the mean.

(D) The Population Mean (Notation: \( \mu \))
\[
\mu = \frac{\sum x_i}{N} \quad (\text{eq} 3.2)
\]

(E) Median: the middle value when the observations are arranged in ascending order (smallest value to largest value).

Note 1: For an odd number of observations, the median is the middle value; for an even number of observations, the median is the average of the two middle values.

EX 1 (cont.) Find the median.

EX 2 Find the mean and median of the data set: \( n = 6 \) 15, 3, 46, 623, 126, 64. Find the mean and the median.

2. Measures of Variability

(A) Variance (Notation: Sample Variance \( S^2 \), Population Variance \( \sigma^2 \))
\[
\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad (\text{eq} 3.7)
\]
\[
S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad (\text{eq} 3.8)
\]

(B) Standard Deviation ((Notation: Sample Variance \( s \), Population Variance \( \sigma \))
\[
s = \sqrt{\sigma^2} \quad (\text{eq} 3.9)
\]
\[
\sigma = \sqrt{\sigma^2} \quad (\text{eq} 3.10)
\]
EX 6 Given a set of data: \( n = 5 \): 3, 7, 5, 8, 7. Find the variance and the standard deviation.

Step 1: Find \( \bar{x} \)
\[
\bar{x} = \frac{\sum x_i}{n} \quad (\text{eq} 3.1)
\]
Step 2: Set up a table to find \( (x_i - \bar{x})^2 \)
Step 3: Sample Variance
\[
S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 4
\]
Step 4: Standard Deviation
\[
S = \sqrt{\sigma^2} = \sqrt{4} = 2
\]
CH 3: Descriptive Statistics: Numerical Measures Part 2

(C) Range

Range = Largest value – Smallest value

EX 6 cont. Find the range of the data set: (n = 5: 3, 7, 5, 8, 7).

(D) Interquartile Range

Interquartile Range = \( Q_3 - Q_1 \) (eq3.6)

1. Quartiles: dividing the ordered data into four portions.
2. \( Q_1 \): the first quartile (25th percentile).
3. \( Q_2 \): the second quartile (the median, 50th percentile).
4. \( Q_3 \): the third quartile (the 75th percentile).

EX 5 (cont.) Given a set of data: 15, 20, 25, 25, 27, 28, 30, 34. Find \( Q_1 \), median(\( Q_2 \)), and \( Q_3 \) and find the interquartile range.

(E) Coefficient of Variation

\[
\text{CV} = \left( \frac{\text{Standard deviation}}{\text{Mean}} \times 100 \right) \% = \frac{s}{\bar{x}} \times 100\% \text{ (eq3.11)}
\]

CV is used in comparing two or more sets of data measured in different units

3. Five Number Summary and the Boxplot

(A) The five-number summary: smallest value, \( Q_1 \), \( Q_2 \) (median), \( Q_3 \), largest value

(B) Boxplot: A graphic display of the Five-Number Summary

EX 5 (cont.) Construct the Boxplot of the given data set.

(C) Distribution Shape based on Boxplot:

Case 1: Symmetric

Case 2: Left-skewed

Case 3: Right-skewed

EX 5 (cont.) Find the distribution shape of the data set.

Note: An important numerical measure of the shape of a distribution is called Skewness. Case 1 symmetric (skewness = 0);
Case 2 Left-skewed (skewness < 0);
Case 3 Right-skewness (skewness > 0)

4. \( z \) Scores

(1) \( z \)-Score

\[
z_i = \frac{x_i - \bar{x}}{s} \text{ (eq3.12)}
\]

(2) \( z \)-score is often called the standardized value.

(3) \( z \)-score reflects how many standard deviations above or below the population mean an observation is. For instance, on a scale that has a mean of 500 and a standard deviation of 100, a value of 450 would equal \( z \) score of \( 450-500)/100 = -50/100 = -0.50 \), which indicates that the value is half a standard deviation below the mean.

5. The Empirical Rule:

For a “Bell-Shaped” normal distribution. About 68% (2/3 of the data) lie within one standard deviation of the mean; about 95% of the data lie within two standard deviation of the mean; Almost all (about 99.7%) of the data lie within three standard deviation of the mean.
5. The Empirical Rule:
   For a “Bell-Shaped” normal distribution. About 68% (2/3 of the data) lie within one standard deviation of the mean; about 95% of the data lie within two standard deviation of the mean; Almost all (about 99.7%) of the data lie within three standard deviation of the mean.

6. Measures of Association Between Two Variables
   (A) Scatter diagram: Given paired observations \((x_i, y_i)\) (i.e. data set that is concerning with two measurement variables \(x\) and \(y\)), a scatter diagram uses the \(x\) and \(y\) axis to represent the data.

   (B) The Covariance:
   (1) The covariance measures the strength of the linear relationship between two numerical variables \((x\) and \(y)\).
   (2) The sample covariance is computed from the following equation:
   \[
   s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}
   \]

   (C) The correlation Coefficient
   (1) The correlation coefficient measures the strength of the linear relationship between two numerical variables \((x\) and \(y)\).
   (2) The sample correlation coefficient is computed from the following equation:
   \[
   r_{xy} = \frac{s_{xy}}{s_x s_y}
   \]
   where \(s_x\) is the sample standard deviation of \(x\) and \(s_y\) is the sample standard deviation of \(y\).
   (3) In particular, \(-1 \leq r_{xy} \leq 1\).

EX 7 Given a set of paired observations with \(n = 4\): \((2, 5), (1, 3), (5, 6), (0, 2)\)

   (1) Obtain the scatter diagram.

   (2) Compute the covariance \(s_{xy}\).

   (3) Compute the sample standard deviations \(s_x\) and \(s_y\).

   (4) Compute the correlation coefficient \(r_{xy}\).

   (5) Interpret the result.