Introduction to Descriptive Statistics

WIMBA Live Classroom
(www.elms.umd.edu)
Lecture Outline

1. Concepts
2. Methods of descriptive statistics
3. Measures of central tendency
4. Measures of dispersion
5. Measures of shape
1. Concepts
1.1 What is descriptive statistics?

• Statistics that provide a numerical or quantitative summary of the characteristics of a data set.
1.2 Why descriptive statistics?

• Working with a summary of data set is more efficient and effective than with a large group of values.
2. Methods of Descriptive Statistics

• A data set can be summarized with:
  • Measures of central tendency: the **center** of a data distribution (a single value)
  • Measures of dispersion: the **spread** of a data distribution
  • Measures of shape or relative position: the **shape** of a data distribution
2.1 Measure of central tendency: Mode

- Mode: The value that occurs most frequently.
  - Example: 10, 20, 20, 15, 15, 20, 25
- Bimodal: Two peaks (two mode)
- Multimodal: Multiple peaks (multiple mode)
2.2 Measure of central tendency: Median

- Median: The **middle value** from a set of ordered values and is therefore the value with an equal number of data both above it and below it.
How to determine the median?

• Data must always be **reordered** from smallest to largest
• If the sample size $n$ is **odd**, the median is the **middle** number (at the position $(n+1)/2$) in the list of ordered values
• If sample size $n$ is **even**, the median is the **average** of the middle two values
Example

A data set: 10, 15, 15, 20, 20, 20, 25
• n=7, The position of median is \((7+1)/2=4\), the median is 20.

A data set: 10, 15, 15, 20, 20, 20, 25, 30
• n=8, The middle values are 20 (position 4) and 20 (position 5), the median is \((20+20)/2\).
Measure of central tendency: Mean

Mean: Average arithmetic mean, i.e., the sum of the values divided by the number of values.

**Sample mean:** Mean of observations

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

**Population mean:** \( \mu \)

\[
\mu = \frac{\sum_{i=1}^{N} X_i}{N}
\]
Measure of central tendency: Mean (Contd)

Weighted mean: A mean where there is some variation in the relative contribution of individual data values to the mean.

\[
\bar{X}_w = \frac{\sum_{j=1}^{k} X_j f_j}{n}
\]

Where, \( n = \text{total frequency} \)

Example: Histogram data

\[
\bar{X}_w = \frac{15*1 + 25*3 + 35*4 + 45*2}{10}
\]
Summary

• Mode: The most frequent data value.

• Mean: The average of all data values.

• Median: The middle value of a set of ordered values.
3. Measures of dispersion

3.1 Quantile deviation

1. Quantile: Equal portions of a data set.
   - Median: Divide data into 2 equal sets.
   - Quartiles: 4 equal sets.
   - Quintiles: 5 equal sets.
   - Percentile: 100 equal sets.
   - Deciles: 10 equal sets.
3.1 Quantile deviation (Contd)

2. How to find $p$th quantile:
   - Put the data in order, from smallest to largest.
   - Compute $n \times p$, where $n$ is the sample size.
   - If $n \times p$ is an integer, then $p$th quantile is the average of the $(n \times p)^{th}$ and the $[(n \times p) + 1]^{th}$ numbers in the list.
   - If $n \times p$ is not an integer, then round up, and use the number occurs at that place in the list.
Example:

College Park temperature in next 7 days:
90, 85, 88, 83, 79, 81, 80

\[ n = 7 \]

- 1^{st} quartile (25\% quantile) \[ p = \frac{1}{4} \] (\( n \cdot p = 1.75 \sim 2 \))
  
  79, 80, 81, 83, 85, 88, 90

- 3^{rd} quartile (75\% quantile) \[ p = \frac{3}{4} \] (\( n \cdot p = 5.25 \sim 5 \))
  
  79, 80, 81, 83, 85, 88, 90
3. Interquartile range

• The difference between the 1\textsuperscript{st} quartile (25% quantile) value and the 3\textsuperscript{rd} quartile (75% quantile) value
3.1 Quantile deviation (Contd)

4. Box plot

- The **box** spans from the 1\textsuperscript{st} to the 3\textsuperscript{rd} quartile, a **line** is drawn through the box as the median, the “**whiskers**” extend out to the minimum and maximum values.
3.1 Quantile deviation (Contd)

The temperature in next 7 days
3.2 Standard deviation

Sample standard deviation:

\[ s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}} \]

Population standard deviation:

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{N}} \]
3.3 Variance

Sample variance:

\[ s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1} \]

Population variance:

\[ \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{N} \]

Variance: Average squared deviation from the mean
3.4 Coefficient of variation (CV)

Standard deviation is **absolute measure of variability**, which makes it difficult to compare variability among different locations or at different times.

Example: 40 years of annual precipitation

- Buffalo: mean 35.47, standard deviation 4.70
- San Diego: mean 9.62, standard deviation 4.42

**Question**: which city has a higher precipitation variability?
3.4 Coefficient of variation (CV) (Contd)

\[ CV = \frac{s}{\bar{X}} \times 100(\%) \]

CV measures relative variability in the data
- Low CV: Less dispersed
- High CV: More variable

CV for Buffalo data: 13.25
CV for San Diego data: 45.96
4. Measure of shape

4.1 Skewness

Skewness: Measures the degree of symmetry in a frequency distribution.

\[
\text{skewness} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{n \times s^3}
\]

- If skewness = 0, symmetric distribution.
- If skewness > 0, positive skewed (a tail to the right).
- If skewness < 0, negative skewed (a tail to the left).
4.2 Kurtosis

Kurtosis: Measures the flatness in a frequency distribution

\[
\text{Kurtosis} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{n \times s^4} - 3
\]

- Low Kurtosis: A flat distribution, i.e., data are dispersed more evenly over many different portions of the distribution.
- High Kurtosis: Data distribution has a peaked appearance.
- Types: Leptokurtic (Positive), Mesokurtic (Close to zero), and Platykurtic (Negative)
Question: when do you use which measure of central tendency?

- If a distribution is unimodal and symmetric, the mode, median and mean similarly present central tendency;
- If a distribution has some degree of skewness, the mode, median and mean are positioned at different places;
- If a distribution is bimodal or multimodal, the mean and median may not present central tendency;
- The mean is heavily influenced by outliers.

Note: This is important and refer textbook, page 40 for different distributions