Chapter 3
Descriptive Statistics: Numerical Measures

Slide 1

Learning objectives

1. Single variable – Part I (Basic)
   1.1. How to calculate and use the measures of location
   1.2. How to calculate and use the measures of variability

2. Single variable – Part II (Application)
   2.1. Understand what the measures of location (e.g., mean, median, mode) tell us about distribution shape
      - Discuss its use in manipulating simulated experiments
   2.2. How to detect outliers using z-score and empirical rule
   2.3. How to use Box plot to explore data
   2.4. How to calculate weighted mean
   2.5. How to calculate mean and variance for grouped data

3. Two variables
   3.1. How to calculate and use the measures of association
      - Covariance, Correlation coefficient

Slide 2

L.O. 1. Numerical measures – Part I

- Numerical measures
- Measures of Location
  - Mean, median, mode, percentiles, quartiles
- Measures of Variability
  - Range, interquartile range, variance, standard deviation, coefficient of variation

Slide 3

Numerical Measures

If the measures are computed for data from a sample, they are called <u>sample statistics</u>.

If the measures are computed for data from a population, they are called <u>population parameters</u>.

A sample statistic is referred to as the <u>point estimator</u> of the corresponding population parameter.

Slide 4

•L.O. 1.1.
•Mean
•Median
•Mode
•Percentile
•Quartile

Mean

- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean *m*.

$$\bar{x} = \frac{\sum x_i}{n}$$

Sum of the values of the ***n*** observations

Number of observations in the sample

$$m = \frac{\sum x_i}{N}$$

Sum of the values of the ***N*** observations

Number of observations in the population

Slide 5

•L.O. 1.1.
•Mean
•Median
•Mode
•Percentile
•Quartile

Median

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
  - For odd number of observations:
    - the median is the middle value
  - For even number of observations:
    - the median is the average of the middle two values.
- Whenever a data set has extreme values, the median is the preferred measure of central location.
  - Often used in annual income and property value data

Slide 6

Mode

- The mode of a data set is the value that occurs with the greatest frequency.
- The greatest frequency can occur at two or more different values.
  - If the data have exactly two modes, the data are bimodal.
  - If the data have more than two modes, the data are multimodal.

Example

- Q4 (p. 84)

Compute the mean, median, and mode of the following sample:

53, 55, 70, 58, 64, 57, 53, 69, 57, 68, 53

- Mean = 59.727
- Median = 57
- Mode = 53

- What is the median, if 59 is added to the data?

- Median = 57.5
  = (57+58) / 2

Percentiles

•L.O. 1.1.
•Mean
•Median
•Mode
•Percentile
•Quartile

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
  - Admission test scores for colleges and universities are frequently reported in terms of percentiles.

- The ***p*th percentile** of a data set is a value such that at least ***p*** percent of the items take on this value or less and at least (100 - ***p***) percent of the items take on this value or more.

Slide 9

Percentiles

•L.O. 1.1.
•Mean
•Median
•Mode
•Percentile
•Quartile

- Arrange the data in ascending order.
- Compute index ***i***, the position of the ***p***th percentile.

  $$\boldsymbol{i} = (\boldsymbol{p}/100)\boldsymbol{n}$$

- If ***i*** is not an integer, round up. The ***p***th percentile is the value in the ***i***th position.
- If ***i*** is an integer, the ***p***th percentile is the average of the values in positions ***i*** and ***i***+1.

Slide 10

Quartiles

•L.O. 1.1.
•Mean
•Median
•Mode
•Percentile
•Quartile

- Quartiles are specific percentiles.
- First Quartile = 25th Percentile
- Second Quartile = 50th Percentile = Median
- Third Quartile = 75th Percentile

Slide 11

Example: Percentiles and Quartiles

•L.O. 1.1.
•Mean
•Median
•Mode
•Percentile
•Quartile

- Q4 (p. 84)

Find 25^{th} and 75^{th} percentiles from the sample below:

53, 55, 70, 58, 64, 57, 53, 69, 57, 68, 53

- 25^{th} percentile = First quartile = 53
- 75^{th} percentile = Third quartile = 68

Slide 12

Measures of Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
  - For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.

- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

Slide 13

Range

•**L.O. 1.2.**
•Range
•IQR
•Variance
•St. Deviation
•Coefficient of variation

- The range of a data set is the difference between the largest and smallest data values.
- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

- Range of the sample:
  53, 55, 70, 58, 64, 57, 53, 69, 57, 68, 53

  = 70 – 53 = 17

Slide 14

Interquartile Range (IQR)

- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

- IQR of the sample:
  53, 55, 70, 58, 64, 57, 53, 69, 57, 68, 53

  = 68 – 53 = 15

Variance

The variance is a measure of variability that utilizes all the data.

The variance is the average of the squared differences between each data value and the mean.

The variance is computed as follows:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

for a sample

$$s^2 = \frac{\sum (x_i - m)^2}{N}$$

for a population

Standard Deviation

•L.O. 1.2.
•Range
•IQR
•Variance
•St. Deviation
•Coefficient of variation

- The <u>standard deviation</u> of a data set is the positive square root of the variance.
- It is measured in the <u>same units as the data</u>, making it more easily interpreted than the variance.
- The standard deviation is computed as follows:

$$s = \sqrt{s^2}$$
for a sample

$$s = \sqrt{s^2}$$
for a population

Slide 17

Coefficient of Variation

•L.O. 1.2.
•Range
•IQR
•Variance
•St. Deviation
•Coefficient of variation

- The <u>coefficient of variation</u> indicates how large the standard deviation is in relation to the mean.
- The coefficient of variation is computed as follows:

$$\left(\frac{s}{\bar{x}} \times 100\right)\%$$
for a sample

$$\left(\frac{s}{m} \times 100\right)\%$$
for a population

Slide 18

Example: Variance, Standard Deviation, And Coefficient of Variation

•L.O. 1.2.
•Range
•IQR
•Variance
•St. Deviation
•Coefficient of variation

Consider the same data set:
53, 55, 70, 58, 64, 57, 53, 69, 57, 68, 53

- Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 45.418$$

- Standard Deviation

$$s = \sqrt{s^2} = \sqrt{33.82} = 6.74$$

the standard deviation is about 11% of of the mean

- Coefficient of Variation

$$\left(\frac{s}{\bar{x}} \times 100\right)\% = \left(\frac{6.74}{59.73} \times 100\right)\% = 11.28\%$$

Slide 19

L.O. 2. Numerical measure – Part II

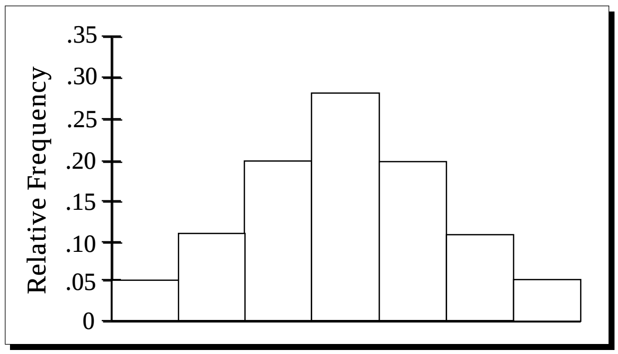
- Measures of Distribution Shape
- Detecting Outliers
  - z-score, empirical rule
- Exploratory Data Analysis
- The Weighted Mean and Working with Grouped Data

Slide 20

Distribution Shape

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- Symmetric (not skewed)
  - Skewness is zero.
  - Mean and median are equal.

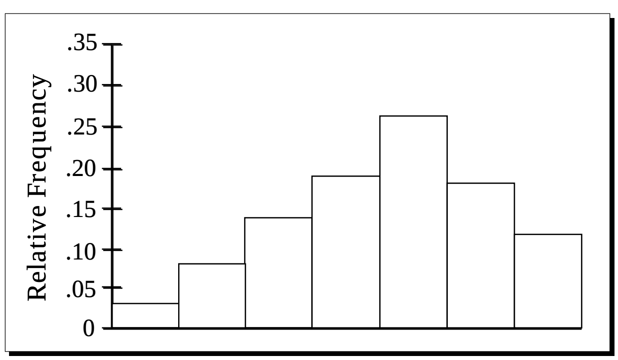


Slide 21

Distribution Shape

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

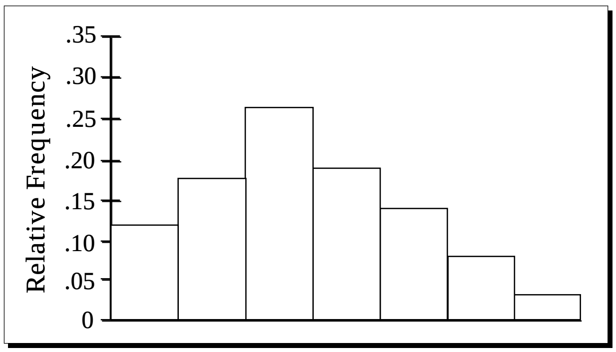
- Moderately Skewed Left
  - Skewness is negative.
  - Mean will usually be less than the median.



Slide 22

Distribution Shape

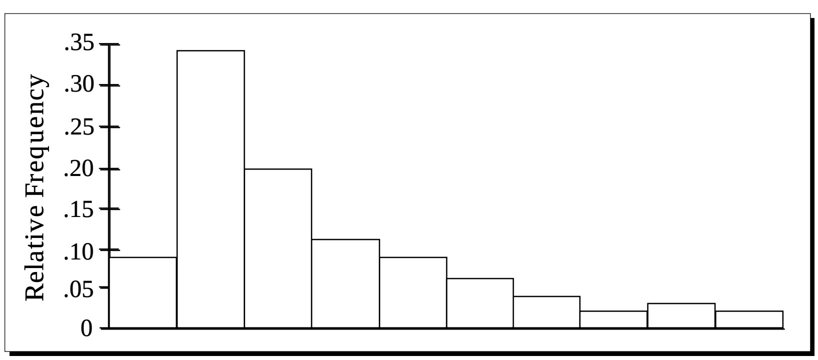
- Moderately Skewed Right
  - Skewness is positive.
  - Mean will usually be more than the median.



•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

Distribution Shape

- Highly Skewed Right
  - Skewness is positive (often above 1.0).
  - Mean will usually be more than the median.



z-Scores

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- The z-score is often called the standardized value.
- It denotes the number of standard deviations a data value x_i is from the mean.

$$z_i = \frac{x_i - \overline{x}}{s}$$

Slide 25

z-Scores

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- An observation's z-score is a measure of the relative location of the observation in a data set.
- A data value less than the sample mean will have a z-score less than zero.
- A data value greater than the sample mean will have a z-score greater than zero.
- A data value equal to the sample mean will have a z-score of zero.

Slide 26

Empirical Rule

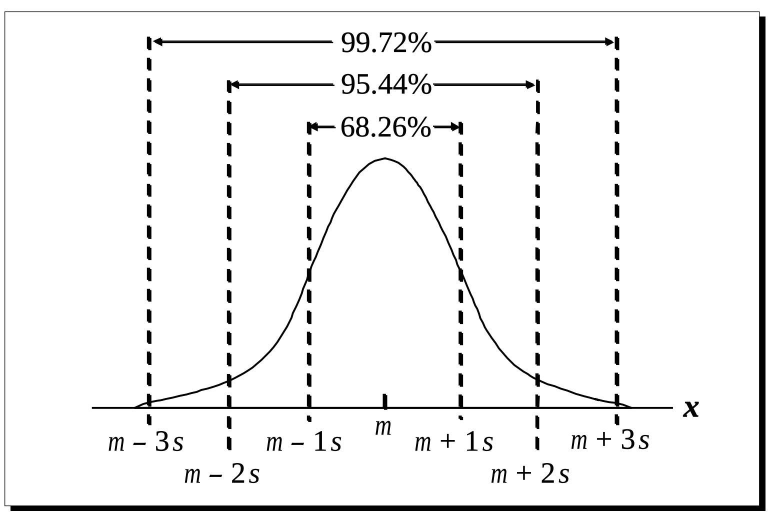
•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

For data having a bell-shaped distribution:

- 68.26% of the values of a normal random variable are within +/- 1 standard deviation of its mean.
- 95.44% of the values of a normal random variable are within +/- 2 standard deviations of its mean.
- 99.72% of the values of a normal random variable are within +/- 3 standard deviations of its mean.

Empirical Rule

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data



Detecting Outliers

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- An outlier is an unusually small or unusually large value in a data set.
- A data value with a z-score less than -3 or greater than +3 might be considered an outlier.
- It might be:
  - an incorrectly recorded data value
  - a data value that was incorrectly included in the data set
  - a correctly recorded data value that belongs in the data set

Slide 29

Exploratory Data Analysis

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- The techniques of exploratory data analysis consist of simple arithmetic and easy-to-draw pictures that can be used to summarize data quickly.

  - Five-Number Summary
  - Box Plot

Slide 30

Five-Number Summary

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

Sample: 53, 55, 70, 58, 64, 57, 53, 69, 57, 68, 53

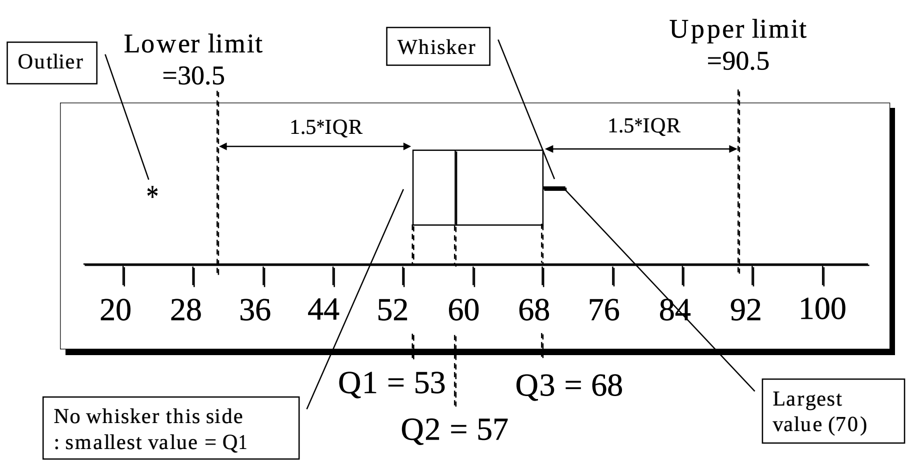
1	Smallest Value	53
2	First Quartile	53
3	Median	57
4	Third Quartile	68
5	Largest Value	70

Slide 31

Box Plot

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- A box plot is based on a five-number summary.



Slide 32

The Weighted Mean and Working with Grouped Data

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- Weighted Mean
- Mean for Grouped Data
- Variance for Grouped Data
- Standard Deviation for Grouped Data

Slide 33

Weighted Mean

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- When the mean is computed by giving each data value a weight that reflects its importance, it is referred to as a weighted mean.
- Class grade is usually computed by weighted mean.

In class midterm exam	Descriptive statistics and distributions	40%
Final group project	Statistical inference	30%
Group project presentation		10%
Homework		10%
Participation		10%

weight

- When data values vary in importance, the analyst must choose the weight that best reflects the importance of each value.

Slide 34

Weighted Mean

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

where:

x_i = value of observation i

w_i = weight for observation i

Grouped Data

- The weighted mean computation can be used to obtain approximations of the mean, variance, and standard deviation for the grouped data.
- To compute the weighted mean, we treat the midpoint of each class as though it were the mean of all items in the class.
- We compute a weighted mean of the class midpoints using the class frequencies as weights.
- Similarly, in computing the variance and standard deviation, the class frequencies are used as weights.

Mean for Grouped Data

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- Sample Data

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

- Population Data

$$\mu = \frac{\sum f_i M_i}{N}$$

where:

f_i = frequency of class i
M_i = midpoint of class i

Slide 37

Sample Mean for Grouped Data

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

Given below is the previous sample of monthly rents for 70 efficiency apartments, presented here as grouped data in the form of a frequency distribution.

Rent ($)	Frequency
420-439	8
440-459	17
460-479	12
480-499	8
500-519	7
520-539	4
540-559	2
560-579	4
580-599	2
600-619	6

Slide 38

Sample Mean for Grouped Data

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

Rent ($)	f_i	M_i	$f_i M_i$
420-439	8	429.5	3436.0
440-459	17	449.5	7641.5
460-479	12	469.5	5634.0
480-499	8	489.5	3916.0
500-519	7	509.5	3566.5
520-539	4	529.5	2118.0
540-559	2	549.5	1099.0
560-579	4	569.5	2278.0
580-599	2	589.5	1179.0
600-619	6	609.5	3657.0
Total	70		34525.0

$$\overline{x} = \frac{34,525}{70} = 493.21$$

This approximation differs by $2.41 from the actual sample mean of $490.80.

Slide 39

Variance for Grouped Data

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- For sample data

$$s^2 = \frac{\sum f_i (M_i - \overline{x})^2}{n-1}$$

- For population data

$$s^2 = \frac{\sum f_i (M_i - m)^2}{N}$$

Slide 40

Sample Variance for Grouped Data

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

Rent ($)	f_i	M_i	$M_i - \bar{x}$	$(M_i - \bar{x})^2$	$f_i(M_i - \bar{x})^2$
420-439	8	429.5	-63.7	4058.96	32471.71
440-459	17	449.5	-43.7	1910.56	32479.59
460-479	12	469.5	-23.7	562.16	6745.97
480-499	8	489.5	-3.7	13.76	110.11
500-519	7	509.5	16.3	265.36	1857.55
520-539	4	529.5	36.3	1316.96	5267.86
540-559	2	549.5	56.3	3168.56	6337.13
560-579	4	569.5	76.3	5820.16	23280.66
580-599	2	589.5	96.3	9271.76	18543.53
600-619	6	609.5	116.3	13523.36	81140.18
Total	70				208234.29

continued →

Sample Variance for Grouped Data

•L.O. 2.
•Shape
•z-score
•Empirical Rule
•Exploratory
•Weighted mean
•Grouped data

- Sample Variance

$$s^2 = 208{,}234.29 / (70 - 1) = 3{,}017.89$$

- Sample Standard Deviation

$$s = \sqrt{3{,}017.89} = 54.94$$

This approximation differs by only $.20
from the actual standard deviation of $54.74.

L.O. 3. Measures of Association Between Two Variables

- Covariance
- Correlation Coefficient

Slide 43

•L.O. 3.
•Covariance
•Correlation

Covariance

- The covariance is a measure of the linear association between two variables.
- Positive values indicate a positive relationship.
- Negative values indicate a negative relationship.

Slide 44

Covariance

•L.O. 3.
•Covariance
•Correlation

- The correlation coefficient is computed as follows:
  - $$s_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$ for samples
  - $$s_{xy} = \frac{\sum (x_i - m_x)(y_i - m_y)}{N}$$ for populations

Slide 45

Correlation Coefficient

•L.O. 3.
•Covariance
•Correlation

- The coefficient can take on values between -1 and +1.
- Values near -1 indicate a strong negative linear relationship.
- Values near +1 indicate a strong positive linear relationship.

Slide 46

•L.O. 3.
•Covariance
•Correlation

Correlation Coefficient

- The correlation coefficient is computed as follows:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

for samples

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

for populations

Slide 47

•L.O. 3.
•Covariance
•Correlation

Correlation Coefficient

- Correlation is a measure of linear association and not necessarily causation.
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other.

Slide 48

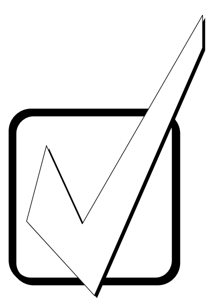
•L.O. 3.
•Covariance
•Correlation

In class Exercise

- Q45 (p. 112)
- Q46 (p. 112)

Slide 49

End of Chapter 3



Slide 50