Inferential Statistics and Probability a Holistic Approach

Chapter 2
Descriptive Statistics

1

1

Example

Anthony's Pizza, a Detroit based company, offers pizza delivery to its customers. A driver for Anthony's Pizza will often make several deliveries on a single delivery run. A sample of 5 delivery runs by a driver showed that the total number of pizzas delivered on each run

2 2 5 9 12

What is the Average?

a) 2
b) 5
c) 6

2

2

Measures of Central Tendency

- Mean
  - Arithmetic Average $\overline{X} = \frac{\sum X_i}{n}$
- Median
  - "Middle" Value after ranking data
  - Not affected by "outliers"
- Mode
  - Most Occurring Value
  - Useful for non-numeric data

3

3

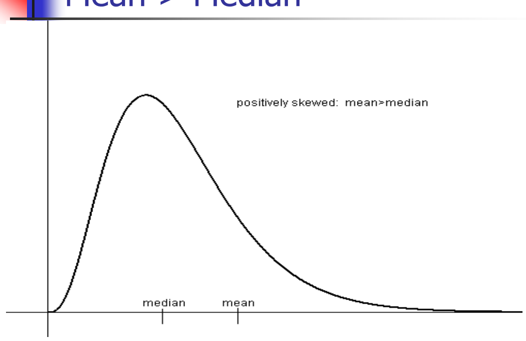
Example – 5 Recent Home Sales

- $500,000
- $600,000
- $600,000
- $700,000
- $2,600,000

4

4

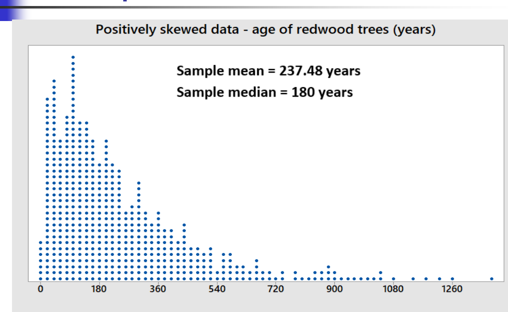
Positively Skewed Data Set Mean > Median



5

5

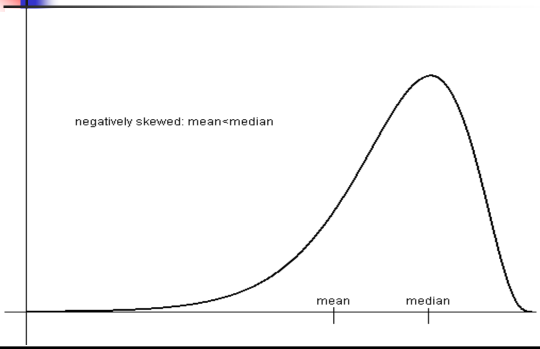
Example – Skewed Positive



6

6

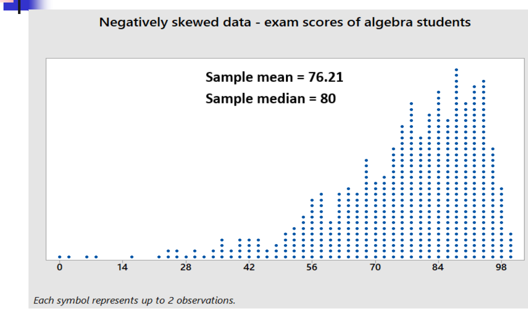
Negatively Skewed Data Set Mean < Median



7

7

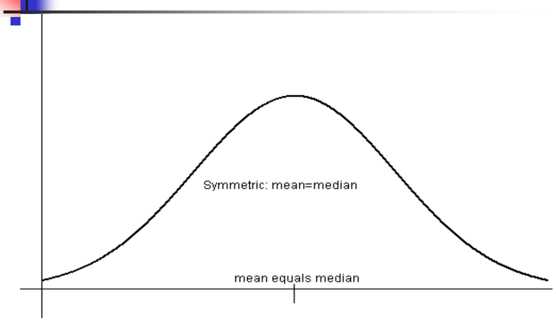
Example – Skewed Negative



8

8

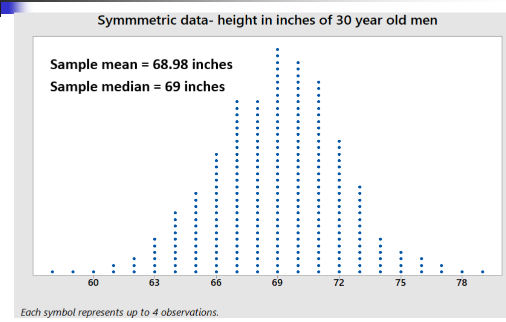
Symmetric Data Set Mean = Median



9

9

Example - Symmetric



10

10

Measures of Variability

- Range
- Variance
- Standard Deviation
- Interquartile Range (percentiles)

11

11

Range

Range = Max(Xi) –Min(Xi) (high – low)

Example – Pizza Delivery
Max = 12 pizzas
Min = 2 pizzas
Range =12 – 2 =10 pizzas

12

12

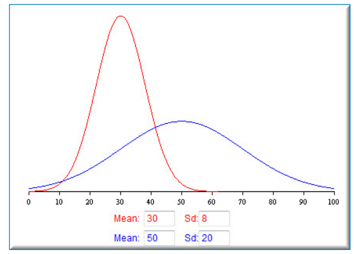
Sample Variance

$$s^2 = \frac{\textit{Sum of Squared Deviations}}{n-1}$$

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

13

Sample Standard Deviation



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

14

Variance and Standard Deviation

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2		
2		
5		
9		
12		
30		

15

Variance and Standard Deviation

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	-4	16
2	-4	16
5	-1	1
9	3	9
12	6	36
30	**0**	**78**

$$s^2 = \frac{78}{4} = 19.5$$

$$s = \sqrt{19.5} \approx 4.42$$

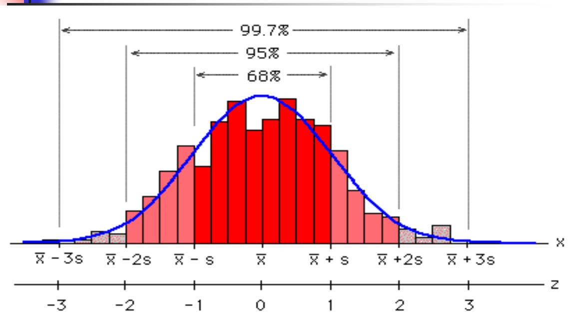
16

Interpreting the Standard Deviation

- Empirical Rule (68-95-99 rule)
  - For bell shaped data
  - 68% within 1 standard deviation of mean
  - 95% within 2 standard deviations of mean
  - 99.7% within 3 standard deviations of mean

17

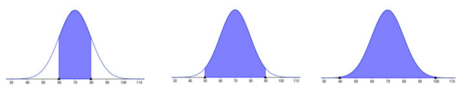
Empirical Rule



18

Example

- An exam has a mean score of 70 and a standard deviation of 10

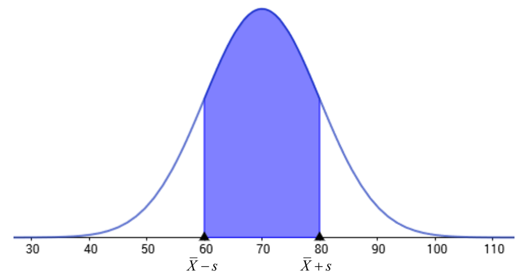


- 68% of scores are between 60 and 80
- 95% of scores are between 50 and 90
- 99.7% of scores are between 40 and 100

19

19

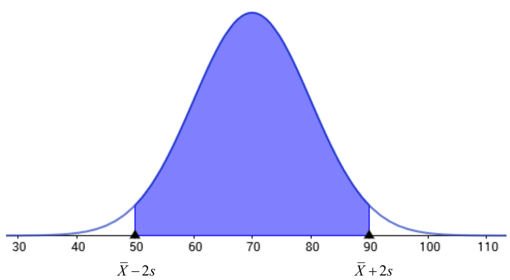
68% of Data



20

20

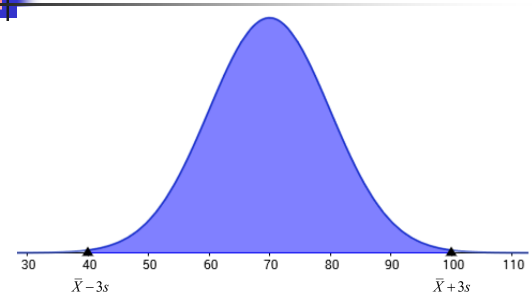
95% of Data



21

21

99.7%% of Data



22

22

Measures of Relative Standing

- Z-score
- Percentile
- Quartiles
- Box Plots

23

23

Z-score

- The number of Standard Deviations from the Mean
- Z>0, X_i is greater than mean
- Z<0, X_i is less than mean

$$Z = \frac{X_i - \bar{X}}{s}$$

24

24

Percentile Rank

Formula for ungrouped data

- The location is (n+1)p (interpolated or rounded)
- n= sample size
- p = percentile

25

25

Quartiles

- 25th percentile is 1st quartile
- 50th percentile is median
- 75th percentile is 3rd quartile
- 75th percentile – 25th percentile is called the Interquartile Range which represents the "middle 50%"

26

26

Alternate method to find Quartiles

- First find median of data. This splits the data into two groups, the lower half and the upper half.
- The median of the lower half of the data is the first quartile.
- The median of the upper half of the data is the third quartile.

27

27

Daily Minutes upload/download on the Internet - 30 students

102	104	85	67	101
71	116	107	99	82
103	97	105	103	95
105	99	86	87	100
109	108	118	87	125
124	112	122	78	92

28

28

Stem and Leaf Graph

```
 6 7
 7 18
 8 25677
 9 25799
10 01233455789
11 268
12 245
```

29

29

IQR Time on Internet data

n+1=31

.25 x 31 = 7.75   location 8 = **87** ← 1st Quartile

.75 x 31 = 23.25   location 23 = **108** ← 3rd Quartile

Interquartile Range (IQR) =108 – 87 = **21**

30

30

Alternate method to find Quartiles

- The median of the data is 101.5
- Q1: The median of the 15 values below 101.5 is 87.
- Q3: The median of the 15 values above 101.5 is 108.
- IQR = 108 - 87 = 21

31

31

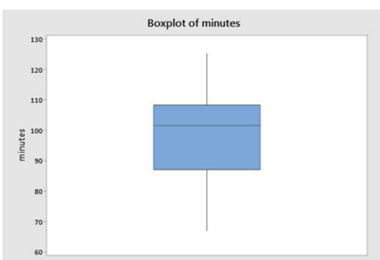
Box Plots

- A box plot is a graphical display, based on quartiles, that helps to picture a set of data.
- Five pieces of data are needed to construct a box plot:
  - Minimum Value
  - First Quartile
  - Median
  - Third Quartile
  - Maximum Value.

32

32

Boxplot



33

33

Outliers

- An outlier is data point that is far removed from the other entries in the data set.
- Outliers could be
  - Mistakes made in recording data
  - Data that don't belong in population
  - True rare events

34

34

Outliers have a dramatic effect on some statistics

- Example quarterly home sales for 10 realtors:

2 2 3 4 5 5 6 6 7 50

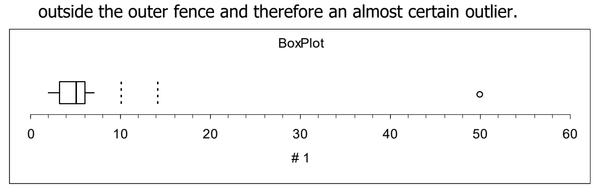
	with outlier	without outlier
Mean	9.00	4.44
Median	5.00	5.00
Std Dev	14.51	1.81
IQR	3.00	3.50

35

35

Using Box Plot to find outliers

- The "box" is the region between the 1st and 3rd quartiles.
- Possible outliers are more than 1.5 IQR's from the box (inner fence)
- Probable outliers are more than 3 IQR's from the box (outer fence)
- In the box plot below, the dotted lines represent the "fences" that are 1.5 and 3 IQR's from the box. See how the data point 50 is well outside the outer fence and therefore an almost certain outlier.



36

36

Using Z-score to detect outliers

- Calculate the mean and standard deviation without the suspected outlier.
- Calculate the Z-score of the suspected outlier.
- If the Z-score is more than 3 or less than -3, that data point is a probable outlier.

$$Z = \frac{50 - 4.4}{1.81} = 25.2$$

37

37

Outliers – what to do

- Remove or not remove, there is no clear answer.
- For some populations, outliers don't dramatically change the overall statistical analysis. Example: the tallest person in the world will not dramatically change the mean height of 10000 people.
- However, for some populations, a single outlier will have a dramatic effect on statistical analysis (called "**Black Swan**" by Nicholas Taleb) and inferential statistics may be invalid in analyzing these populations. Example: the richest person in the world will dramatically change the mean wealth of 10000 people.

38

38

Bivariate Data

- Ordered numeric pairs (X,Y)
- Both values are numeric
- Paired by a common characteristic
- Graph as Scatterplot

39

39

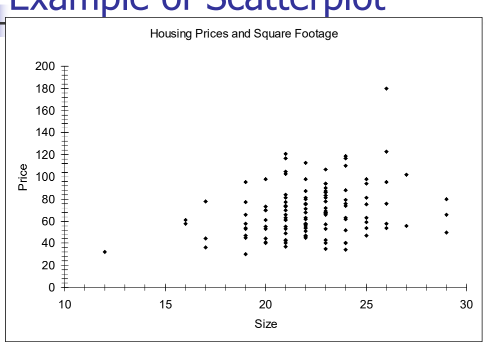
Example of Bivariate Data

- Housing Data
  - X = Square Footage
  - Y = Price

40

40

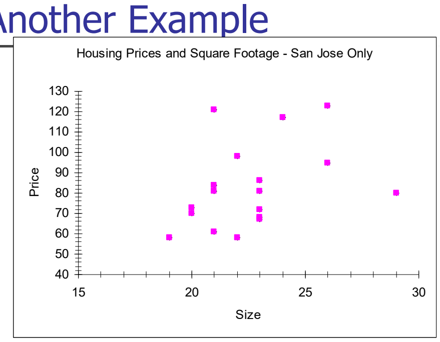
Example of Scatterplot



41

41

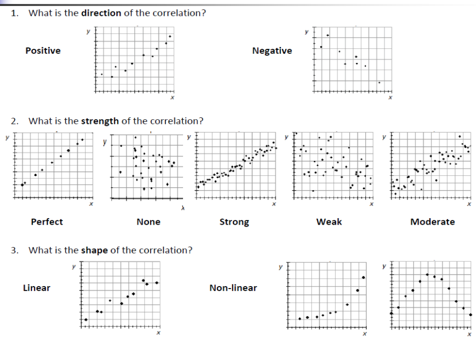
Another Example



42

42

Types of Correlation



43

43

Correlation Analysis

- Correlation Analysis: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- Scatter Diagram: A chart that portrays the relationship between the two variables of interest.
- Dependent Variable: The variable that is being predicted or estimated. "Effect"
- Independent Variable: The variable that provides the basis for estimation. It is the predictor variable. "Cause?" (Maybe!)

44

44

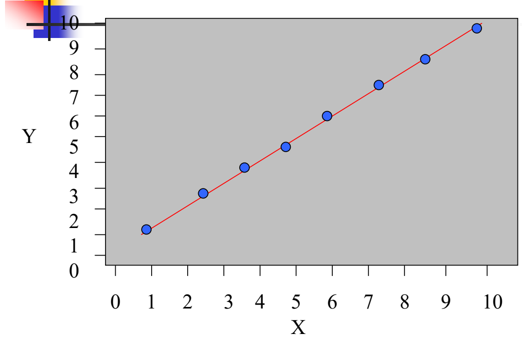
The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the **strength** of the relationship between two variables.
  - It requires interval or ratio-scaled data (variables).
  - It can range from -1 to 1.
  - Values of -1 or 1 indicate perfect and strong correlation.
  - Values close to 0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

45

45

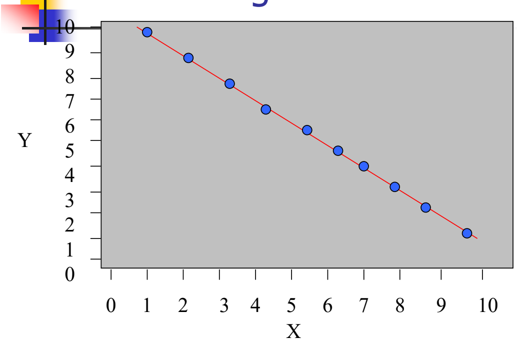
Perfect Positive Correlation



46

46

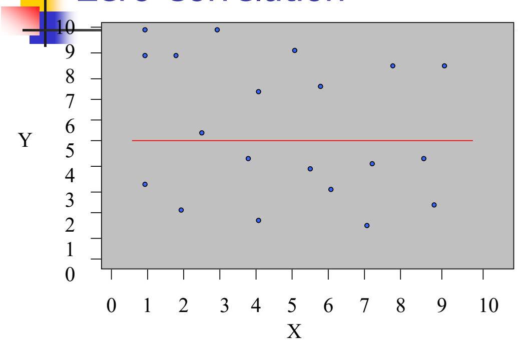
Perfect Negative Correlation



47

47

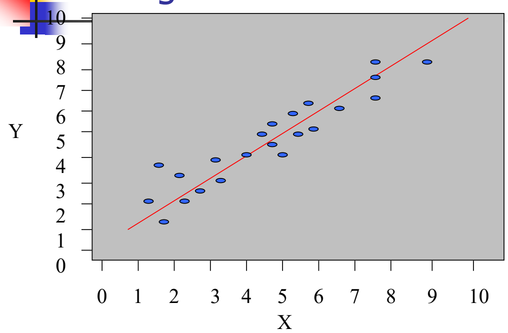
Zero Correlation



48

48

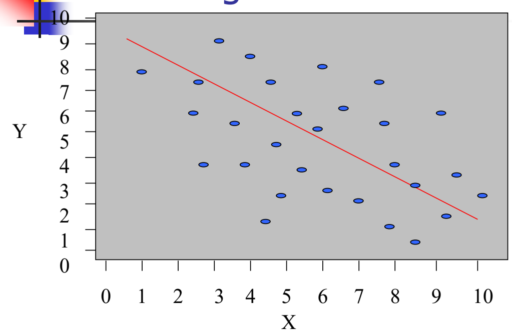
Strong Positive Correlation



49

49

Weak Negative Correlation



50

50

Causation

- Correlation does not necessarily imply causation.
- There are 4 possibilities if X and Y are correlated:
  1. X causes Y
  2. Y causes X
  3. X and Y are caused by something else.
  4. Confounding - The effect of X and Y are hopelessly mixed up with other variables.

51

51

Causation - Examples

- City with more police per capita have more crime per capita.
- As Ice cream sales go up, shark attacks go up.
- People with a cold who take a cough medicine feel better after some rest.

52

52

Formula for correlation coefficient r

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$

$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$

$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma Y)$$

53

53

Example

- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000
- Make a Scatter Diagram
- Find the correlation coefficient

X	10	15	20	30	40
Y	40	35	25	25	15

54

54

Example *continued*

- Make a Scatter Diagram
- Find the correlation coefficient

55

Example *continued*

X	Y	X^2	Y^2	XY
10	40	100	1600	400
15	35	225	1225	525
20	25	400	625	500
30	25	900	625	750
40	15	1600	225	600
115	140	3225	4300	2775

- SSX = $3225 - 115^2/5$ = 580
- SSY = $4300 - 140^2/5$ = 380
- SSXY= 2775 - (115)(140)/5 = -445

56

Example *continued*

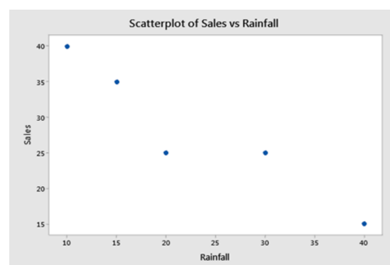
$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

$$r = \frac{-445}{\sqrt{580 \cdot 330}} = -0.9479$$

- Strong negative correlation

57

Minitab: Graphs>Scatterplot



58

Minitab - Correlation

Stat>
Basic Statistics>
Correlation

Correlations

	Rainfall
Sales	-0.948

59