Inferential Statistics and Probability
a Holistic Approach

Chapter 2
Descriptive Statistics

1

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

2

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

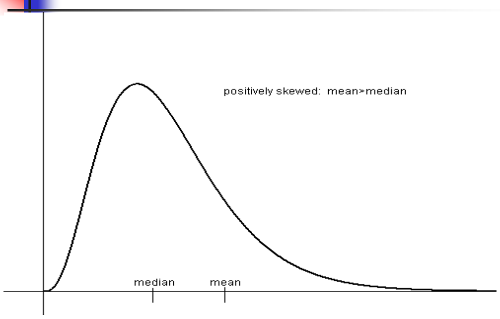
3

Example – 5 Recent Home Sales

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

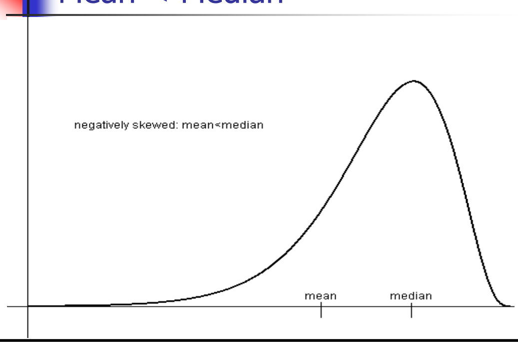
4

Positively Skewed Data Set Mean > Median



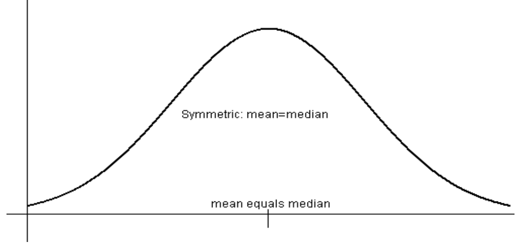
5

Negatively Skewed Data Set Mean < Median



6

Symmetric Data Set
Mean = Median



7

Measures of Variability

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

8

Range

Max(Xi) –Min(Xi)

125 – 67 = 58

9

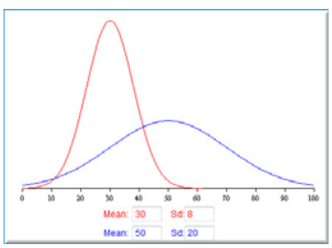
Sample Variance

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

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

10

Sample Standard Deviation



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

11

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**

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

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

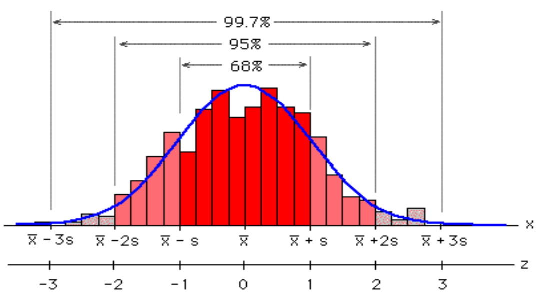
12

Interpreting the Standard Deviation

- Chebyshev's Rule
  - At least 100 x (1-(1/k)^2)% of any data set must be within k standard deviations of the mean.
- Empirical Rule (68-95-99 rule)
  - 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

13

Empirical Rule



14

Measures of Relative Standing

- Z-score
- Percentile
- Quartiles
- Box Plots

15

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

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

16

Percentile Rank

Formula for ungrouped data

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

17

Quartiles

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

18

IQR example

n+1=31

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

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

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

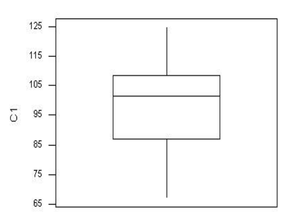
19

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.

20

Box Plot



21

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

22

Outliers have a dramatic effect on some statistics

- Example quarterly home sales for 10 realtors:

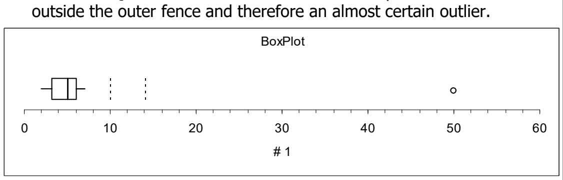
2 2 3 4 5 5 6 6 7 50

	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

23

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.



24

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.

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

25

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.

26

Bivariate Data

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

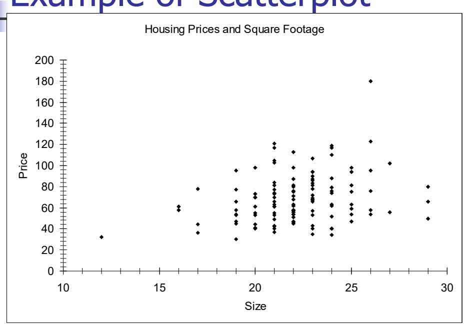
27

Example of Bivariate Data

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

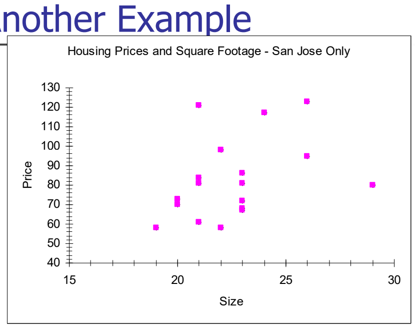
28

Example of Scatterplot



29

Another Example



30

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!)

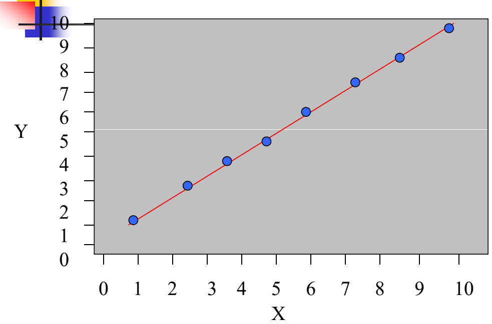
31

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.

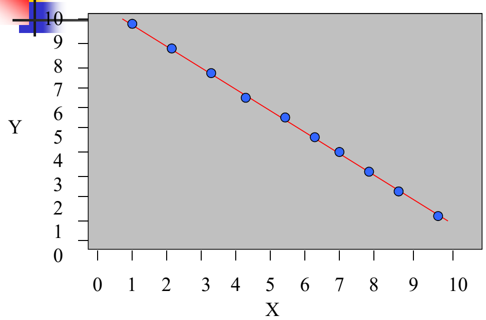
32

Perfect Positive Correlation



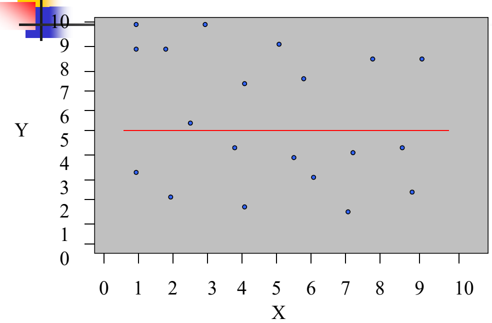
33

Perfect Negative Correlation



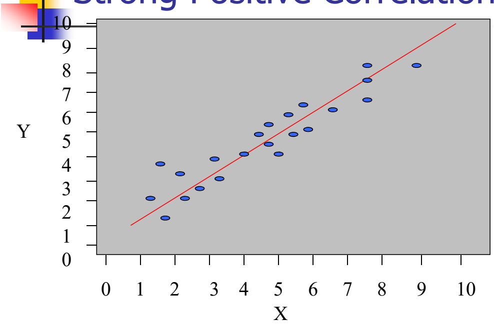
34

Zero Correlation



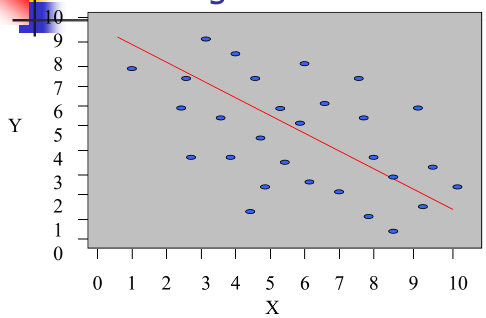
35

Strong Positive Correlation



36

Weak Negative Correlation



37

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.

38

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.

39

Formula for correlation coefficient r

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

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$
$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$
$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma Y)$$

40

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

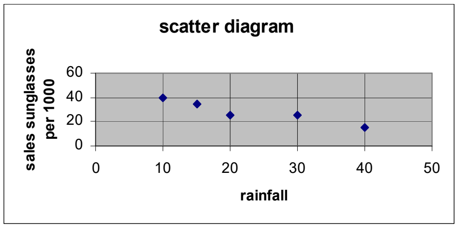
41

Example *continued*

- Make a Scatter Diagram
- Find the correlation coefficient

42

Example *continued*



43

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

44

Example *continued*

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

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

- Strong negative correlation

45