Chapter 3.1, 3.2
Numerical descriptive measures

Graphs provide a global/qualitative description of a sample, but they are imprecise for use in statistical inferences.

We use numerical measures which can be calculated for either a sample (these measures are called statistics) or a population (parameters).

- Measures of location
- Measures of variability

1

Measures of central tendency (ungrouped data)

- The **mode**: is the sample value that occurs most frequently.
- The **median**: is the value that falls in the middle position when the sample values are ordered from the smallest to the largest.
- The **mean**: is the average value, the balance point.
  - The mode can be computed for both qualitative and quantitative variables.
  - The median and the mean we compute for quantitative variables.

2

Mean

The ***mean for ungrouped data*** is obtained by dividing the sum of all values by the number of values in the data set. Thus,

Mean for population data: $\mu = \dfrac{\sum x}{N}$

Mean for sample data: $\bar{x} = \dfrac{\sum x}{n}$

3

Population Parameters and Sample Statistics

- A numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data set is called a population parameter, or simply a **parameter**.
- A summary measure calculated for a sample data set is called a sample statistic, or simply a **statistic**.

4

Example 1

Table 1 lists the total philanthropic givings (in million dollars) by six companies . Find the mean contributions of the six companies

Corporation	Money Given in 2007 (millions of dollars)
CVS	22.4
Best Buy	31.8
Staples	19.8
Walgreen	9.0
Lowe's	27.5
Wal-Mart	337.9

5

Median: Computations

- The median: is the value in the middle position when the sample values are ordered from smallest to largest.
  - Order the sample values from smallest to largest.
  - Identify the sample size n.
  - Find the value in the position
    - (n+1)/2 if n is odd;
    - Average the values in the position n/2 and n/2 +1 when n is even.
- **Exercise 1.** Compute the median and mean for the data from example 1 – with and without Wal -Mart

Mode

- The mode: is the sample value that occurs most frequently.
- From the frequency distribution, identify the value with largest frequency.

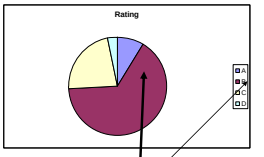
Example : Rating of quality of education, sample of 400 school administrators

Rating	Frequency	Relative F	Percent
A	35	0.09	9%
B	260	0.65	65%
C	93	0.23	23%
D	12	0.03	3%
Total	400	1	100%

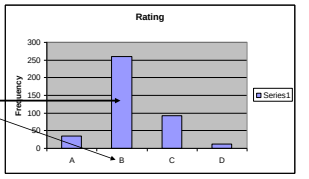
The mode is the category B

7

Read the Mode from Charts



Find the maximum Frequency, read the category label



8

Exercise 2

a)The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77 82 74 81 79 84 74 78

Find the mode.

b)Last year's incomes of five randomly selected families were

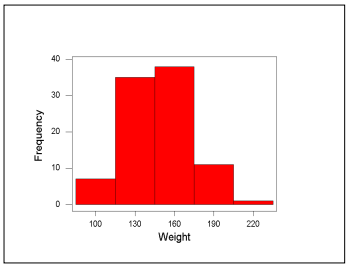
$76,150. $95,750, $124,985, $87,490, and $53,740.

Find the mode.

9

Exercise 3

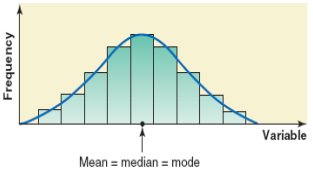
Read the median from an histogram



Hint: 6+34+38+12+2=92

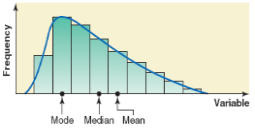
Relationships among the Mean, Median, and Mode

For a symmetric histogram and frequency curve with one peak , the values of the mean, median, and mode are identical, and they lie at the center of the distribution.

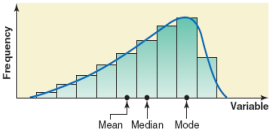


11

Mean, median, and mode for a histogram and frequency curve skewed to the right.



Mean, median, and mode for a histogram and frequency curve skewed to the left.



12

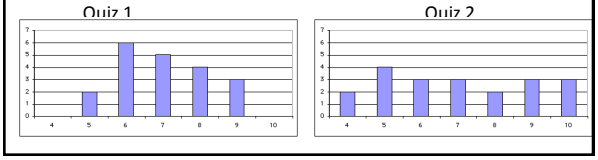
Properties

- When a distribution is symmetric, then the mode, the mean, and the median are the same.
- The mode is a meaningful measure of location when you are looking for the sample value with the largest frequency.
- The median gives an idea of the center of the distribution and, compared to the mean, it is less sensitive to unusually large or unusually small values (outliers).
- With very skewed distributions, the median is a better measure of location than the mean.

13

3.2 What is Variability?

- Variability refers to how "spread out" a group of scores is. These 2 graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different.
- The scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students was much greater on Quiz 2 than on Quiz 1.



- Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution.
- The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is.
- There are four frequently used measures of variability:
  - range:
  - interquartile range
  - variance, and standard deviation.

15

Range = Largest value – Smallest Value (easy to compute)

Exercise 4 Data - Flower petals: 5, 12, 6, 8, 14.
Calculate the range.

- Disadvantages:
  - The range, like the mean has the disadvantage of being influenced by outliers. Consequently, the range is not a good measure of dispersion to use for a data set that contains outliers.
  - Its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range.

16

- A deviation is the distance that a data value is from the mean.
  - Since adding all deviations together would total zero, we square each deviation and find an average of sorts for the deviations.
- The standard deviation is the most used measure of dispersion.
- The standard deviation is just the square root of the variance and is measured in the same units as the original data.
- The value of the standard deviation tells how closely the values of a data set are clustered around the mean.
- In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean.
- In contrast, a large value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively large range around the mean

17

- The **variance of a population** of *N* measurements is the average of the squared deviations of the measurements about their mean m.

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

The **variance of a sample** of *n* measurements is the sum of the squared deviations of the measurements about their mean, divided by $(n - 1)$.

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

Why divide by n –1?

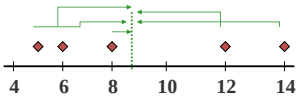
The sample standard deviation *s* is often used to estimate the population standard deviation s. Dividing by $n - 1$ gives us a better estimate of s.

18

Exercise 5

Data - Flower petals: 5, 12, 6, 8, 14.
Calculate the sample variance and standard deviation

$$\bar{x} = \frac{5+12+6+8+14}{5} = 9$$



Two Ways to Calculate the Sample Variance

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	5	5-9=-4	$(-4)^2$=16
	12	12-9=3	$(3)^2$=9
	6	6-9=-3	9
	8	8-9=-1	1
	14	14-9=5	25
Sum	45	0	60

Use the Definition Formula:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{60}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

Variance and Standard Deviation

Short-cut Formulas for the Variance and Standard Deviation for Ungrouped Data

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance.

Two Ways to Calculate the Sample Variance

	x_i	x_i^2
	5	25
	12	144
	6	36
	8	64
	14	196
Sum	45	465

Use the short –cut formula:

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{465 - \frac{45^2}{5}}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$