Handout 3.
Numerical descriptive measures (chapter 3)

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

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.

Mode

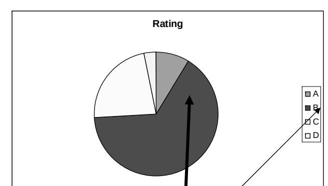
- The mode: is the sample value that occurs most frequently.
- From the frequency distribution, identify the value with largest frequency.

Example 1: Rating of quality of education, sample of 400 school administrators

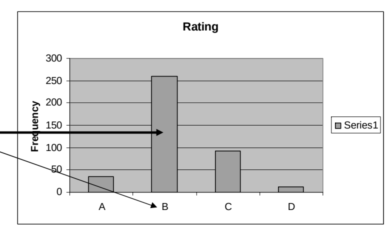
The mode is the category B

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%

Read the Mode from Charts



Find the maximum Frequency, read the category label



Mode

- A major shortcoming of the mode is that a data set may have none or may have more than one mode, whereas it will have only one mean and only one median.
  - Unimodal: A data set with only one mode.
  - Bimodal: A data set with two modes.
  - Multimodal: A data set with more than two modes.
  When all categories occur with the same frequency, the mode is not defined.

Median: Computations

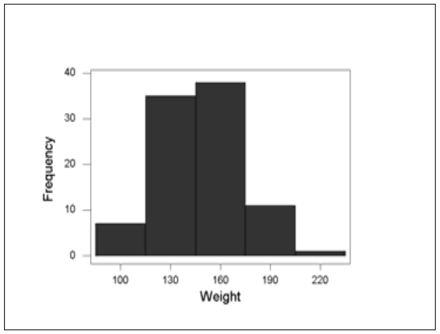
- The median: is the value in the middle position when the sample values are ordered from smallest to largest.
  - Order the sample values from smallest to largest.
  - Identify the sample size n.
  - Find the value in the position
    - (n+1)/2 if n is odd;
    - Average the values in the position n/2 and n/2 +1 when n is even.
- **Exercise 1.** Compute the median for the data sets **:**

Data 1: 2 9 11 5 6 27

Data 2: 7 10 34 6 8

Exercise 2

Read the median from an histogram



Hint: n=6+34+38+12+2=92 np = 46

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:

$$\mu = \frac{\sum x}{N}$$

Mean for sample data:

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

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**.

Exercise 3

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

$$\textbf{Mean} = \frac{22.4 + 31.8 + 19.8 + 9.0 + 27.5 + 337.9}{6} = \$74.73 \textbf{ million}$$

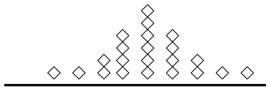
Notice that the charitable contributions made by Wal-Mart are very large compared to those of other companies. Hence, it is an outlier.

If we do not include the charitable givings of Wal-Mart (the outlier), the mean of the charitable contributions of the five companies is :

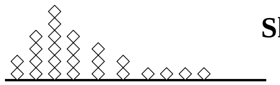
$$\textbf{Mean} = \frac{22.4 + 31.8 + 19.8 + 9.0 + 27.5}{5} = \$22.1 \textbf{ million}$$

Properties

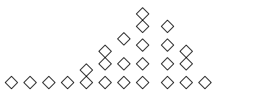
- When a distribution is symmetric, then the mode, the mean, and the median are the same.
- The mode is a meaningful measure of location when you are looking for the sample value with the largest frequency.
- The median gives an idea of the center of the distribution and, compared to the mean, it is 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.



Symmetric: Mean = Median



Skewed right: Mean > Median



Skewed left: Mean < Median

MEASURES OF DISPERSION FOR UNGROUPED DATA

- Range
- Variance and Standard Deviation

Range = Largest value – Smallest Value

- 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.

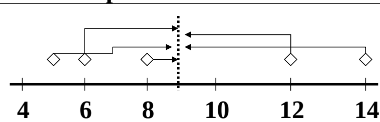
Its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range.

- **The variance** measures spread,how distant are the sample points from the mean.
- Deviation is the distance from a sample point to the mean).
- It is easy to show,that the average deviation is always equal to 0.This is why we compute the sample variance as the average squared deviation

Data - Flower petals: 5, 12, 6, 8, 14.

Exercise 4:Calculate the sample variance and standard deviation

$$\bar{x} = \frac{45}{5} = 9$$



The Variance

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

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

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

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.

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

The Standard Deviation

- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.
- To return this measure of variability to the original units of measure, we calculate the **standard deviation**, the positive square root of the variance.

Population standard deviation : $\sigma = \sqrt{\sigma^2}$

Sample standard deviation : $s = \sqrt{s^2}$

Two Ways to Calculate the Sample Variance

	x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	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:

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

$$= \frac{60}{4} = 15$$

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

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 Calculational Formula:

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

$$= \frac{465 - \frac{45^2}{5}}{4} = 15$$

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

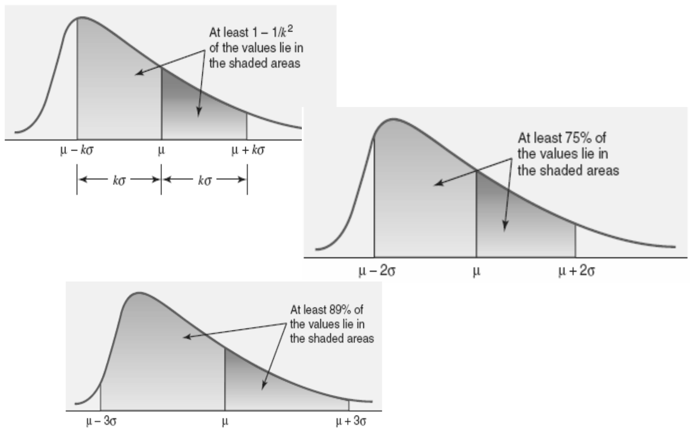
USE OF STANDARD DEVIATION

- Chebyshev's Theorem
- Empirical Rule

Chebyshev's Theorem:

For any number k greater than 1, at least $(1 - 1/k^2)$ of the data values lie within k standard deviations of the mean.

Chebyshev's theorem.



✓**Important results:**

✓If $k = 2$, at least $1 – 1/2^2 = ¾$ (75%) of the measurements are within 2 standard deviations of the mean.

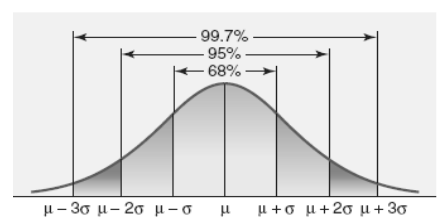
✓If $k = 3$, at least $1 – 1/3^2 = 8/9$ (89%) of the measurements are within 3 standard deviations of the mean.

Empirical Rule

For a bell shaped distribution approximately

1. 68% of the observations lie within one standard deviation of the mean
2. 95% of the observations lie within two standard deviations of the mean
3. 99.7% of the observations lie within three standard deviations of the mean

Illustration of the empirical rule.



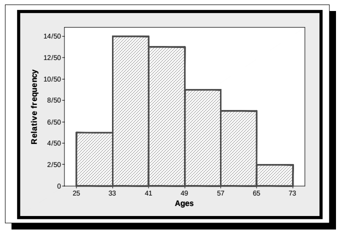
Exercise 5

The ages of 50 tenured faculty at a state university.

- 34 48 **70** 63 52 52 35 50 37 43 53 43 52 44 42 31 36 48 43 **26** 58 62 49 34 48 53 39 45
- 34 59 34 66 40 59 36 41 35 36 62 34 38 28 43 50 30 43 32 44 58 53

$$\bar{x} = 44.9$$
$$s = 10.73$$

Shape? Skewed right

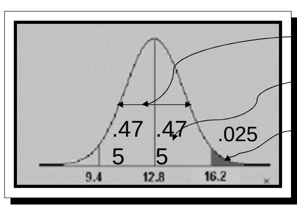


k	$\bar{x} \pm ks$	Interval	Proportion in Interval	Tchebysheff	Empirical Rule
1	**44.9 ±10.73**	**34.17 to 55.63**	**31/50 (.62)**	**At least 0**	**≈ .68**
2	**44.9 ±21.46**	**23.44 to 66.36**	**49/50 (.98)**	**At least .75**	**≈ .95**
3	**44.9 ±32.19**	**12.71 to 77.09**	**50/50 (1.00)**	**At least .89**	**≈ .997**

- Do the actual proportions in the three intervals agree with those given by Tchebysheff's Theorem?
  - **Yes. Tchebysheff's Theorem must be true for any data set.**
- Do they agree with the Empirical Rule?
  - **No. Not very well.**
- Why or why not?
  - **The data distribution is not very mound-shaped, but skewed right.**

Exercise 6

The length of time for a worker to complete a specified operation averages 12.8 minutes with a standard deviation of 1.7 minutes. If the distribution of times is approximately mound-shaped, what proportion of workers will take longer than 16.2 minutes to complete the task?



95% between 9.4 and 16.2

47.5% between 12.8 and 16.2

(50-47.5)% = 2.5% above 16.2

Approximating *s*

- From Tchebysheff's Theorem and the Empirical Rule, we know that

$$R \approx 4\text{-}6\, s$$

- To approximate the standard deviation of a set of measurements, we can use:

$$s \approx R/4$$
$$\text{or } s \approx R/6 \text{ for a large data set.}$$

R = 70 – 26 = 44

$$s \approx R/4 = 44/4 = 11$$