Marmara University, **2021**

Probability and Statistics

Subject 2

Describing Data with Numerical Measures

Mujdat Soyturk, Ph.D.
Associate Professor

Contents

Most parts of the slides are derived from the textbook: "Mendenhall, Beaver, Beaver, *Introduction to Probability and Statistics*, 14th Ed., Brooks/Cole, Cengage Learning, 2013"

Describing Data with Numerical Measures

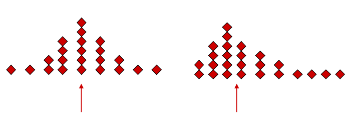
- Graphical methods may not always be sufficient for describing data.
  - Doesn't present the "degree of differences".
- **Numerical measures (*mental pictures*)** can be created for both **populations** and **samples.**
  - A **parameter** is a numerical descriptive measure calculated for a population.
  - A **statistic** is a numerical descriptive measure calculated for a sample.

Describing Data with Numerical Measures

- These measures are
  - Measures of Center
  - Measures of Variability (Dispersion)
  - Measures of Relative Standing

Measures of Center

A measure along the horizontal axis of the data distribution that locates the center of the distribution.



Arithmetic Mean or Average

The **mean or average** of a set of measurements is the sum of the measurements divided by the total number of measurements.

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

where n = number of measurements

$\sum x_i$ = sum of all the measurements

Example

•The set: 2, 9, 11, 5, 6

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2+9+11+5+6}{5} = \frac{33}{5} = 6.6$$

If we were able to enumerate the whole population, the **population mean** would be called μ (the Greek letter "mu").

Median

- The **median** of a set of measurements is the middle measurement when the measurements are ranked from smallest to largest.
- The **position of the median** is

$$\boxed{0.5(n + 1)}$$

once the measurements have been ***ordered***.

Example

- The set: 2, 4, 9, 8, 6, 5, 3 $n = 7$
- Sort : 2, 3, 4, 5, 6, 8, 9
- Position: $0.5(n + 1) = 0.5(7 + 1) = 4^{th}$

Median = 4^{th} largest measurement

- The set: 2, 4, 9, 8, 6, 5 $n = 6$
- Sort: 2, 4, 5, 6, 8, 9
- Position: $.5(n + 1) = .5(6 + 1) = 3.5^{th}$

Median = (5 + 6)/2 = 5.5 — average of the 3^{rd} and 4^{th} measurements

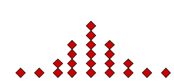
Mean vs. Median

- Although both the **mean** and the **median** are good measures of the center of a distribution, the median is less sensitive to extreme values or *outliers*.
- When a data stream has extremely small or extremely large observations, the sample mean is drawn toward the direction of extreme measurements.

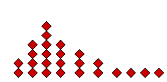
Mean vs. Median

- If a distribution *skewed to the right*, the **mean** *shifts to the right*.
- If a distribution *skewed to the left*, the **mean** *shifts to the left*.
- The **median** is not affected by these extreme values.
- When a distribution is **symmetric**, the mean and the median are equal.

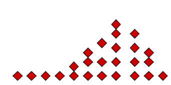
Extreme Values



Symmetric: Mean = Median



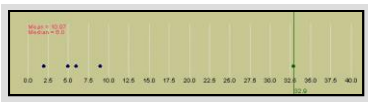
Skewed right: Mean > Median



Skewed left: Mean < Median

Extreme Values

- The mean is more easily affected by extremely large or small values than the median.



- The median is often used as a measure of center when the distribution is skewed.

Mode

- The **mode** is the measurement which occurs most frequently.
- The set: 2, 4, 9, 8, 8, 5, 3
  - The mode is **8**, which occurs twice
- The set: 2, 2, 9, 8, 8, 5, 3
  - There are two modes — **8** and **2** (bimodal)
- The set: 2, 4, 9, 8, 5, 3
  - There is no mode (each value is unique).

Example

The number of quarts (U.S. liquid unit equal to 1.136 liters) of milk purchased by 25 households:

0 0 1 1 1 1 1 2 2 2 2 2 2 2 2
3 3 3 3 3 4 4 4 5

- **Mean?**

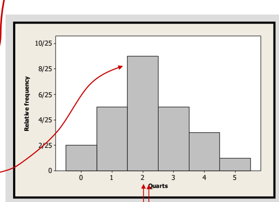
$$\boxed{\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2}$$

- **Median?**

$$\boxed{m = 2}$$

- **Mode? (Highest peak)**

$$\boxed{\text{mode} = 2}$$



Measures of Dispersion

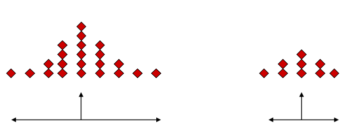
Consider the following two sets of integers:

S = {5, 5, 5, 5, 5, 5} and R = {0, 0, 0, 10, 10, 10}

- If we calculated the mean for both S and R, we will get the number 5 for both.
- However, these are two vastly different data sets. Therefore we need another descriptive statistic besides a measure of central tendency, which we shall call a measure of dispersion or measure of variability.
- We shall measure the dispersion or scatter of the values of our data set about the mean of the data set. If the values tend to be concentrated near the mean, then this measure shall be small, while if the values of the data set tend to be distributed far from the mean, then the measure will be large.
- The two measures of dispersions that are usually used are called the **variance** and **standard deviation**.

Measures of Variability

A measure along the horizontal axis of the data distribution that describes the **spread** of the distribution from the center.



The Range

- The **range, R,** of a set of *n* measurements is the difference between the largest and smallest measurements.
- **Example:** A botanist records the number of petals on 5 flowers:

  5, 12, 6, 8, 14

- The range is

  R = 14 – 5 = 9.

- Quick and easy, but only uses 2 of the 5 measurements.
- Doesn't show the relative variability of two sets.

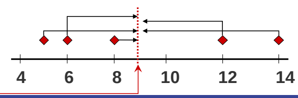
The Variance

The **variance** is measure of variability that uses all the measurements.

It measures the average deviation of the measurements about their mean.

Flower petals: 5, 12, 6, 8, 14

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



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 $(n - 1)$.

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

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.

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

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

Two Ways to Calculate the Sample Variance

Use the Definition Formula:

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

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

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

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

Two Ways to Calculate the Sample Variance

Use the Calculational Formula:

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

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

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

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

Some Notes

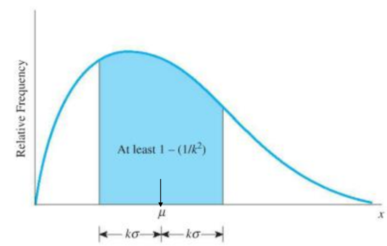
- The value of s is **ALWAYS** positive.
- The larger the value of s^2 or s, the larger the variability of the data set.
- **Why divide by n –1?**
  - The sample standard deviation s is often used to estimate the population standard deviation σ. Dividing by n –1 gives us a better estimate of σ.

Using Measures of Center and Spread: Tchebysheff's Theorem

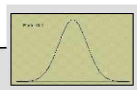
Given a number k greater than or equal to 1 and a set of n measurements, **at least** 1-(1/k^2) of the measurement will lie within k standard deviations of the mean.

- Can be used for either samples ($\bar{x}$ and s) or for a population (μ and σ).
- **Important results:**
  - If k = 2, at least $1 - 1/2^2 = 3/4$ of the measurements are within 2 standard deviations of the mean.
  - If k = 3, at least $1 - 1/3^2 = 8/9$ of the measurements are within 3 standard deviations of the mean.

Using Measures of Center and Spread: Tchebysheff's Theorem



Using Measures of Center and Spread: The Empirical Rule



Given a distribution of measurements that is approximately mound-shaped:

✓The interval $\mu \pm \sigma$ contains approximately 68% of the measurements.

✓The interval $\mu \pm 2\sigma$ contains approximately 95% of the measurements.

✓The interval $\mu \pm 3\sigma$ contains approximately 99.7% of the measurements.

Example

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

Example

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.

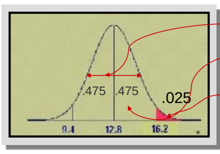
Using Tchebysheff's Theorem and the Empirical Rule

- **Tchebysheff's Theorem**
  - Applies to any set of measurements; *large* or *small*, *mound-shaped* or *skewed*.
  - Always satisfied.
  - Gives lower bounds for the given intervals.
- **The Empirical Rule**
  - Is applied only when the data tend to be roughly mound-shaped.
  - Rule of thumb
  - More accurate estimation

Example

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, Question: 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

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

Number of Measurements	Expected Ratio of Range to s
5	2.5
10	3
25	4

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

$$\boxed{\begin{array}{l} s \approx R/4 \\ \text{or } s \approx R/6 \text{ for a large data set.} \end{array}}$$

Approximating s

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						

R = 70 – 26 = 44

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

Actual s = 10.73

$\boldsymbol{s \approx R/4}$

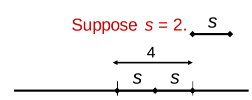
gives only an ***approximate*** value for s.

Measures of Relative Standing

- Sometimes we need to know the position of one observation relative to the others in a set of data.
- Where does one particular measurement stand in relation to the other measurements in the data set?
- How many standard deviations away from the mean does the measurement lie? This is measured by the ***z*-score.**
- ***z*-score** measures the distance between an observation and the mean, measured in units of standard deviation.

$$z\text{-score} = \frac{x - \bar{x}}{s}$$

Suppose s = 2.



$\bar{x} = 5$ $\quad x = 9$

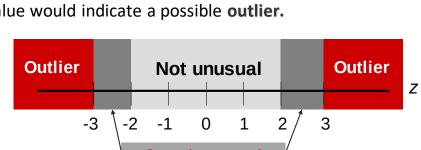
x = 9 lies z =2 std dev from the mean.

Measures of Relative Standing

- The ***z-score*** is a valuable tool for determining whether a particular observation is likely to occur quite frequently or whether it is unlikely and might be considered an ***outlier***.

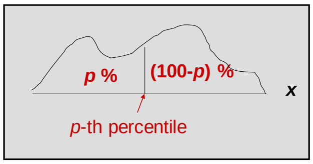
z-Scores

- From Tchebysheff's Theorem and the Empirical Rule
  - At least 3/4 and more likely 95% of measurements lie within 2 standard deviations of the mean.
  - At least 8/9 and more likely 99.7% of measurements lie within 3 standard deviations of the mean.
- *z*-scores between –2 and 2 are not unusual. *z*-scores should not be more than 3 in absolute value. *z*-scores larger than 3 in absolute value would indicate a possible **outlier.**

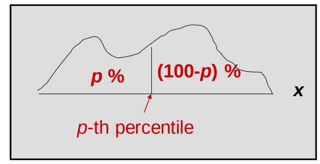


Measures of Relative Standing

- A percentile is another measure of relative standing and is most often used for large data sets.
- How many measurements lie below the measurement of interest? This is measured by the p^{th} **percentile.**

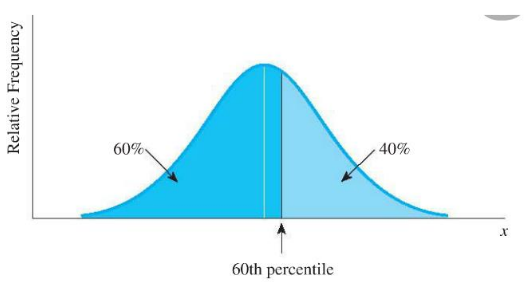


Measures of Relative Standing



- The **p^{th} percentile** is the value of ***x*** that is greater than *p%* of the measurements and is less than the remaining *(100-p)%.*

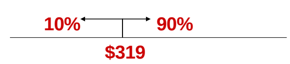
Measures of Relative Standing



Examples

- 90% of all men (16 and older) earn more than $319 per week.

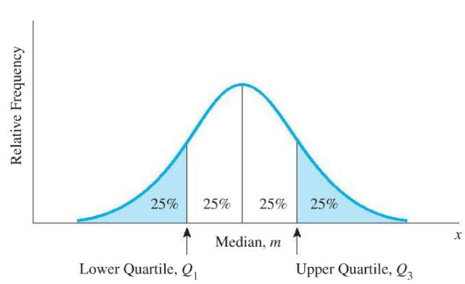
BUREAU OF LABOR STATISTICS



$319 is the 10th percentile.

50th Percentile	≡ Median
25th Percentile	≡ Lower Quartile (Q_1)
75th Percentile	≡ Upper Quartile (Q_3)

Quartiles and the IQR



Quartiles and the IQR

- The **lower quartile** **(Q_1)** is the value of *x* which is larger than 25% and less than 75% of the ordered measurements.
- The **upper quartile** **(Q_3)** is the value of *x* which is larger than 75% and less than 25% of the ordered measurements.
- The range of the “middle 50%” of the measurements is the **interquartile range,**

$$IQR = Q_3 - Q_1$$

Calculating Sample Quartiles

- The **lower and upper quartiles (Q_1 and Q_3),** can be calculated as follows:
- The **position of Q_1** is

.25(*n* + 1)

- The **position of Q_3** is **.75(*n* + 1)**

once the measurements have been ordered.
If the positions are not integers, find the quartiles by interpolation.

Calculating Sample Quartiles

interpolation for calculating the quartiles
(if the quartile position is not an integer):

- The lower quartile is taken to be the value of ¾ of the distance between the adjacent measurements at the quartile position.
- The upper quartile is taken to be the value of ¼ of the distance between the adjacent measurements at the quartile position.
- Example 2.13

Example

The prices ($) of 18 brands of walking shoes:
40 60 65 65 65 68 68 70 70
70 70 70 70 74 75 75 90 95

Position of Q_1 = .25(18 + 1) = 4.75

Position of Q_3 = .75(18 + 1) = 14.25

✓Q_1 is 3/4 of the way between the 4th and 5th ordered measurements, or
$Q_1 = 65 + .75(65 - 65) = 65.$

Example

The prices ($) of 18 brands of walking shoes:
40 60 65 65 65 68 68 70 70
70 70 70 70 74 75 75 90 95

Position of Q_1 = .25(18 + 1) = 4.75

Position of Q_3 = .75(18 + 1) = 14.25

✓Q_3 is 1/4 of the way between the 14th and 15th ordered measurements, or
$Q_3 = 74 + .25(75 - 74) = 74.25$
✓and
$IQR = Q_3 – Q_1 = 74.25 - 65 = 9.25$

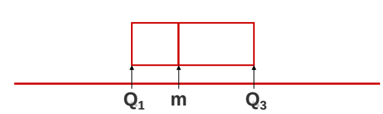
Using Measures of Center and Spread: The Box Plot

The Five-Number Summary:				
Min	Q_1	Median	Q_3	Max

- ***Five-Number Summary*** divides the data into 4 sets containing an equal number of measurements. (1/4 measurements)
- A quick summary of the data distribution.
- Use to form a **box plot** to describe the **shape** of the distribution and to detect **outliers.**

Constructing a Box Plot

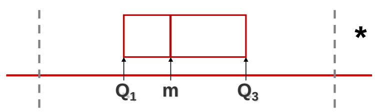
i. Calculate Q_1, the median, Q_3 and IQR.
ii. Draw a horizontal line to represent the scale of measurement.
iii. Draw a box using Q_1, the median, Q_3.



Constructing a Box Plot

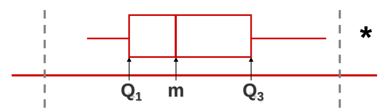
iv. Isolate outliers by calculating
- Lower fence: Q_1 - 1.5 IQR
- Upper fence: Q_3 + 1.5 IQR

v. Measurements beyond the upper or lower fence is are **outliers** and are marked (*).

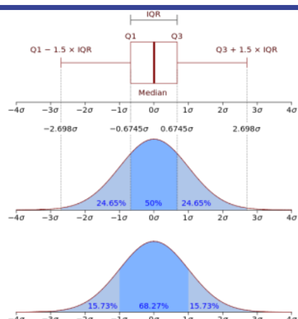


Constructing a Box Plot

vi. Draw "whiskers" connecting the largest and smallest measurements that are NOT outliers to the box.



Box Plot

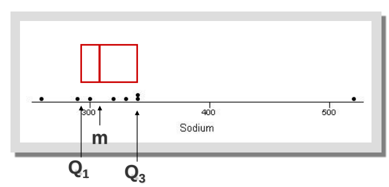


Example

Amount of sodium in 8 brands of cheese:

260 290 300 320 330 340 340 520

Q_1 = 297.5   m = 325 Q_3 = 340



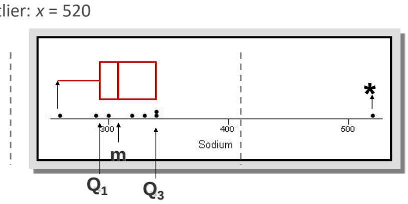
Example

IQR = 340-297.5 = 42.5

Lower fence = 297.5-1.5(42.5) = 233.75

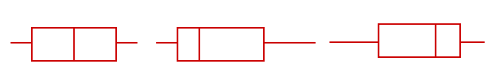
Upper fence = 340 + 1.5(42.5) = 403.75

Outlier: x = 520



Interpreting Box Plots

- Median line in center of box and whiskers of equal length — symmetric distribution
- Median line left of center and long right whisker — skewed right
- Median line right of center and long left whisker — skewed left



Key Concepts

I. Measures of Center

1. Arithmetic mean (mean) or average

   a. Population: μ

   b. Sample of size n: $\bar{x} = \dfrac{\sum x_i}{n}$

2. Median: **position** of the median = $.5(n+1)$
3. Mode
4. The median may preferred to the mean if the data are highly skewed.

II. Measures of Variability

1. Range: R = largest − smallest

Key Concepts

2. Variance

   a. Population of N measurements: $\sigma^2 = \dfrac{\Sigma(x_i - \mu)^2}{N}$

   b. Sample of n measurements:

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

3. Standard deviation

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

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

4. A rough approximation for s can be calculated as $s \approx R/4$. The divisor can be adjusted depending on the sample size.

Key Concepts

III. Tchebysheff's Theorem and the Empirical Rule

1. Use Tchebysheff's Theorem for any data set, regardless of its shape or size.

   a. At least 1-($1/k^2$) of the measurements lie within k standard deviation of the mean.

   b. This is only a lower bound; there may be more measurements in the interval.

2. The Empirical Rule can be used only for relatively mound-shaped data sets.

   – Approximately 68%, 95%, and 99.7% of the measurements are within one, two, and three standard deviations of the mean, respectively.

Key Concepts

IV. Measures of Relative Standing

1. Sample z-score:
2. pth percentile; p% of the measurements are smaller, and $(100 - p)$% are larger.
3. Lower quartile, Q_1; **position** of $Q_1 = .25(n+1)$
4. Upper quartile, Q_3; **position** of $Q_3 = .75(n+1)$
5. Interquartile range: IQR = $Q_3 - Q_1$

V. Box Plots

1. Box plots are used for detecting outliers and shapes of distributions.
2. Q_1 and Q_3 form the ends of the box. The median line is in the interior of the box.

Key Concepts

3. Upper and lower fences are used to find outliers.

   a. **Lower fence:** $Q_1 - 1.5$(IQR)

   b. **Upper fence:** $Q_3 + 1.5$(IQR)

4. **Whiskers** are connected to the smallest and largest measurements that are not outliers.
5. Skewed distributions usually have a long whisker <u>in the direction</u> of the skewness, and the median line is drawn <u>away from the direction</u> of the skewness.