Editing file

Feedback

Description of data (Using summary & variability measures)

Objectives:

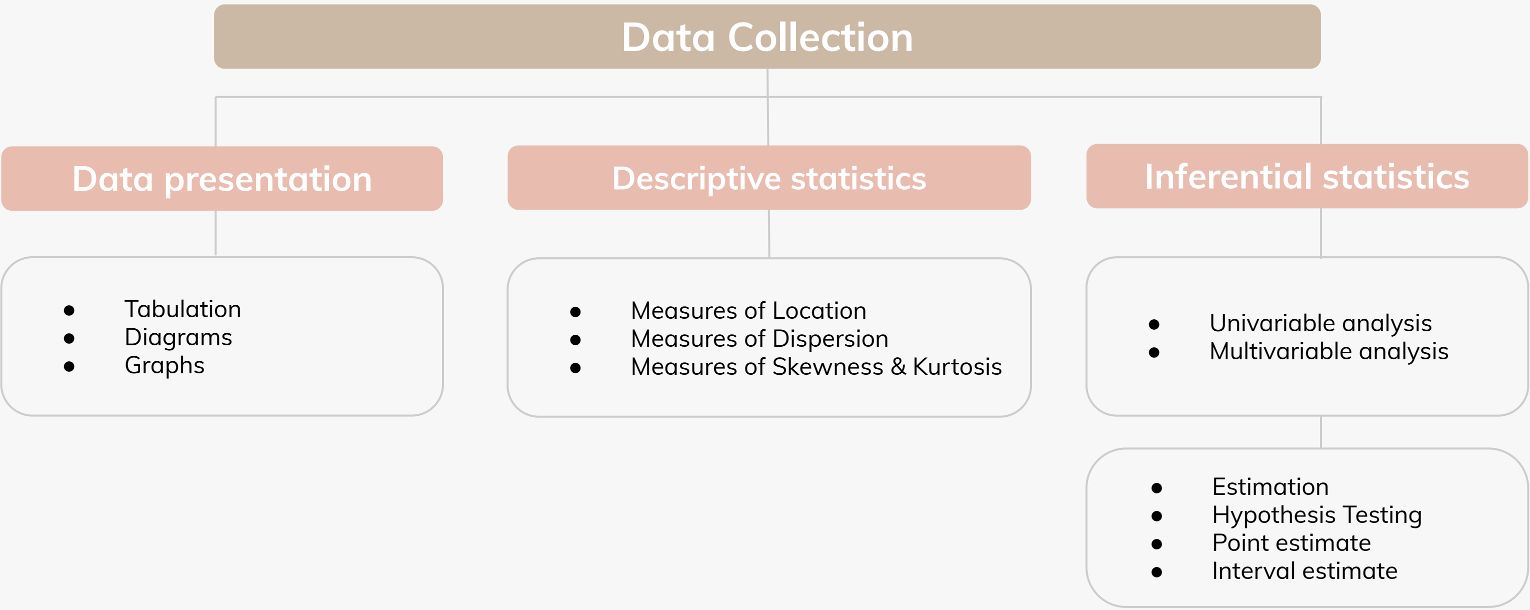
1. To understand how to summarize the data.
2. To understand how to measure the variability of the data.
3. To use and interpret appropriately the different summary and variability measures.

23rd lecture

Color Index:

- 🟦 Boys' Slides
- 🟪 Girls' Slides
- 🟩 Doctors Notes
- 🟨 Golden Notes
- 🟥 Important
- ⬜ Extra

Investigation



Data Collection

Data presentation

- Tabulation
- Diagrams
- Graphs

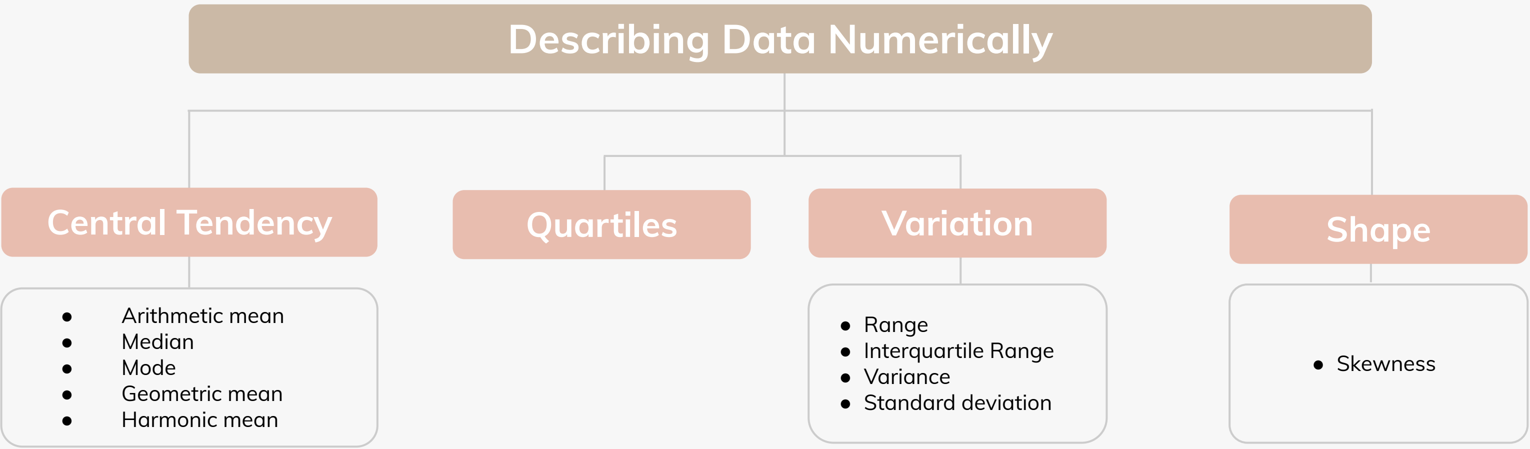
Descriptive statistics

- Measures of Location
- Measures of Dispersion
- Measures of Skewness & Kurtosis

Inferential statistics

- Univariable analysis
- Multivariable analysis

- Estimation
- Hypothesis Testing
- Point estimate
- Interval estimate



Describing Data Numerically

Central Tendency

- Arithmetic mean
- Median
- Mode
- Geometric mean
- Harmonic mean

Quartiles

Variation

- Range
- Interquartile Range
- Variance
- Standard deviation

Shape

- Skewness

Measures of Central Tendency

A statistical measure that identifies a single score as representative for an entire distribution. The goal of central tendency is to find the single score that is most typical or most representative of the entire group , **There are 3 common measures of central tendency:**

1 The Mean

Sum/total (Average)

2 The Median

Middle number of all data

3 The Mode

Most frequent value

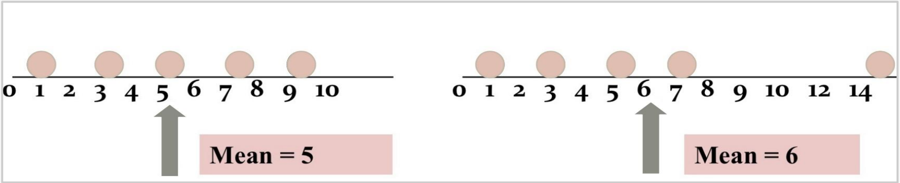
Mean (Arithmetic mean)

The most common measure of central tendency , Affected by extreme values (outliers)

(Suppose in your class there are 10 student who are brilliant , the class average will go up)

Calculate the mean of the following data: **1 5 4 3 2**

- **Sum the scores** (∑X): 1 + 5 + 4 + 3 + 2 = 15
- **Divide the sum** (∑X = 15) by the number of scores (N = 5): 15 / 5 = 3
- **Mean** = $\overline{X}$ = 3



Summary & variability measures

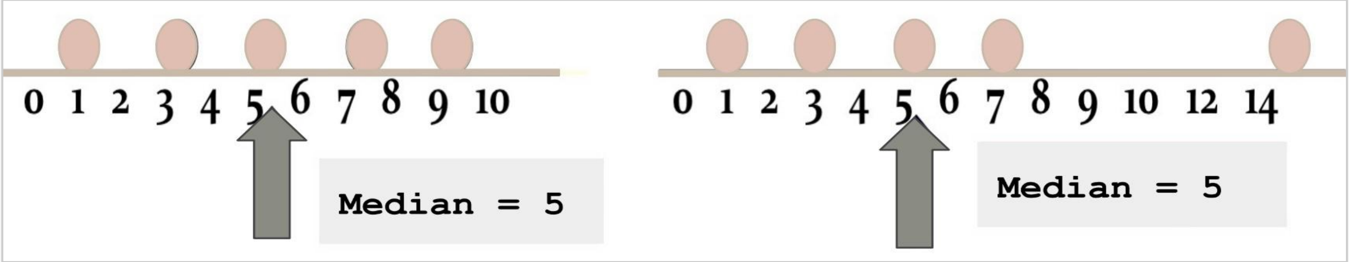
The Median

- The **median** is simply another name for the **50th percentile** (the central part of any data)
- It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median.
- Not affected by extreme values (because we are taking the center)

In an ordered array, the median is the "middle" number

❑ If n or N is odd, the median is the middle number.

❑ If n or N is even, the median is the average of the two middle numbers.



How To Calculate the Median:

Conceptually, it is easy to calculate the median:

1. Sort the data from highest to lowest
2. Find the score in the middle
   - Middle = (N + 1) / 2
   - If N, the number of scores is even, the median is the average of the middle two scores.

Example

What is the median of the following scores: 24 18 19 42 16 12

1. Sort the scores: 42 24 19 18 16 12
2. Determine the middle score: middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5
3. Median = average of 3rd and 4th scores: (19 + 18) / 2 = 18.5

What is the median of the following scores: 10 8 14 15 7 3 3 8 12 10 9

1. Sort the scores: 15 14 12 10 10 9 8 8 7 3 3
2. Determine the middle score: middle = (N + 1) / 2 = (11 + 1) / 2 = 6
3. Middle score = median = 9

Summary & variability measures

Measures of Central tendency

Mean ... the **most frequently used** but is **sensitive to extreme scores.**

Example

○ 1 2 3 4 5 6 7 8 9 10 → **Mean =** 5.5 **(median =** 5.5 **)**

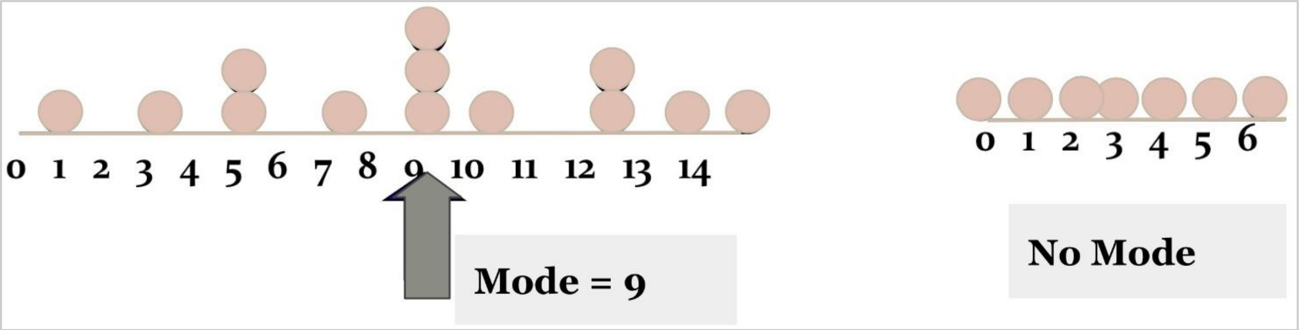
○ 1 2 3 4 5 6 7 8 9 20 → **Mean =** 6.5 **(median =** 5.5 **)**

○ 1 2 3 4 5 6 7 8 9 100 → **Mean =** 14.5 **(median =** 5.5 **)**

The Mode

Value that occurs most often , Not affected by extreme values.

- ➔ Used for either numerical or categorical (nominal) data.
- ➔ There may be no mode or there may be several modes.
  (Ex: 20 student got 90 out of 100)



The Shape of Distributions

Distributions can be either **symmetrical** **or** **skewed** (becoming narrow on one side), depending on whether there are more frequencies at one end of the distribution than the other.

Symmetrical distribution	A distribution is symmetrical if the frequencies at the right & left tails of the distribution are identical, so that if it is divided into two halves, each will be the mirror image of the other. (both sides will have equal area) ★ In a symmetrical distribution the mean, median, and mode are identical.

Summary & variability measures

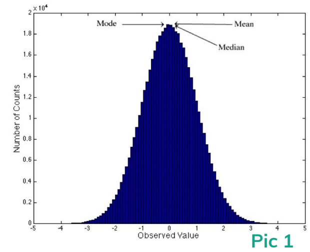
Distributions

To decide which is the statistical test appropriate to use , we must know what is the distribution is ? Symmetrical or skewed

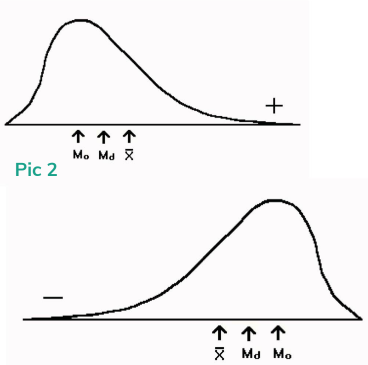
- **Bell-Shaped** (also known as symmetric" or "normal")
- **Skewed:**
  - Positively (skewed to the **right**) – it tails off toward larger values .

    Pic1: for example the grades of medicine only 5% of the student will get A+ but the majority will be less than that , the PIC reflet the majority of student in the left (less number) and the students who gets high grades are on the right
  - Negatively (skewed to the **left**)– it tails off toward smaller values (not negative values).

    Pic2: for example the the majority of the students grades in the research course are A or A+ so they will be on the right of the chart.

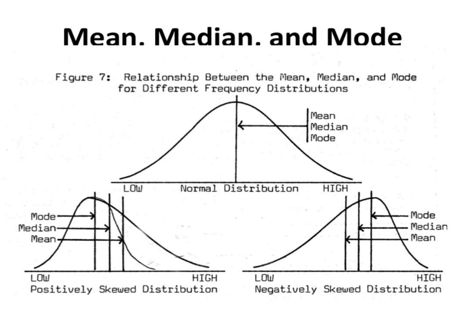


Pic 1



Pic 2

Skewed distribution	Few extreme values on one side of the distribution or on the other. - **Positively** skewed distributions: distributions which have few extremely high values **(Mean > Median)**. - **Negatively** skewed distributions: distributions which have few extremely low values **(Mean < Median)**.



Mean. Median. and Mode

Figure 7: Relationship Between the Mean, Median, and Mode for Different Frequency Distributions

Choosing a Measure of Central tendency

Prof: it's important

❏ IF variable is Nominal.. —> Mode

❏ IF variable is Ordinal... —> Mode or Median (or both)

❏ IF variable is Interval-Ratio and distribution is Symmetrical... —> Mode, Median or Mean

❏ IF variable is Interval-Ratio and distribution is Skewed... —> Mode or Median

Example

○ 7,8,9,10,11 n=5, ∑x=45, X̄=45/5=9

○ 3,4,9,12,15 n=5, ∑x=45, X̄=45/5=9

○ 1,5,9,13,17 n=5, ∑x=45, X̄=45/5=9

S.D. : (1) 1.58 (2) 4.74 (3) 6.32

1.58 —> Less variability
4.74 —> There is variability
6.32 —> High variability

Variability will be low when:
1-Accurate measurements
2-Good sample size

Measures of Dispersion / Variability

Measures of Dispersion (Variability) (Heterogeneity)

Measures of dispersion summarize differences in the data, how the numbers differ from one another.

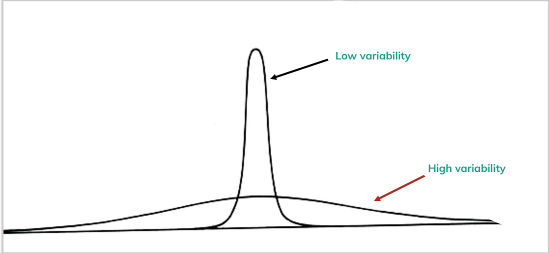
Example

Presence of variability:

- Series I : 70 70 70 70 70 70 70 70 70 70. **(no variation / dispersion)**
  In real life you can't get series I
- Series II : 66 67 68 69 70 70 71 72 73 74. **(low variation / dispersion)**
- Series III :1 19 50 60 70 80 90 100 110 120. **(high variation / dispersion)**

A single summary figure that describes the spread of observations within a distribution

In this figure, both curves have the same median, but curve 2 (red arrow) has greater variance.

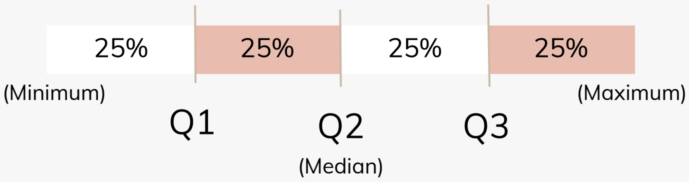


Range	Difference between the smallest and largest observations. Not a very good measure , because it only take 2 values & ignore the others. **Ex: marks of student** ⊙ 52, 76, 100, 36, 86, 96, 20, 15, 57, 64, 64, 80, 82, 83, 30, 31, 31, 31, 32, 37, 38, 38, 40, 40, 41, 42, 47, 48, 63, 63, 72, 79, 70, 71, 89 **Range: 100-15 = 85** take the difference It doesn't consider the other data
Interquartile range	Range of the middle half of scores.
Variance	Mean of all squared deviations from the mean.
Standard deviation	Rough measure of the average amount by which observations deviate from the mean. The square root of the variance.

Measures of Dispersion / variability

Quartiles

Divides ranked scores 4 four equal parts.



25% 25% 25% 25%

(Minimum) Q1 Q2 (Median) Q3 (Maximum)

Calculating Quartiles

Q1

$$Q_1 = \frac{n+1}{4} \text{th}$$

Q2

$$Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2} \text{th}$$

Q3

$$Q_3 = \frac{3(n+1)}{4} \text{th}$$

#438's team:
Example, Calculate the quartiles from this score (data): 6, 3, 1,7,4, 9, 4. :

- First rank the score (data):
  1, 3, 4, 4, 6, 7, 9.
- n is the number of observation.. x1, x2 … xn, in this case it equals 7.
- $Q1 = \frac{(7)+(1)}{4} = 2$
- $Q2 = \frac{[2][(7)+(1)]}{4} = 4$
- $Q3 = \frac{[3][(7)+(1)]}{4} = 6$

Q1 = (second observation) = 3
Q2 = (fourth observation) = 4
Q3 = (sixth observation) = 7

1, **3**, 4, **4**, 6, **7**, 9.

Inter Quartile Range (IQR)

IQR

$$\text{IQR} = Q_3 - Q_1$$

- ❏ The inter quartile range is Q3-Q1. (Not Q1-Q3)
- ❏ 50% of the observations in the distribution are in the inter quartile range.
- ❏ The following figure shows the interaction between the quartiles, the median and the inter quartile range.
- ❏ Interquartile range is still not a good measure , why? Because in uses only 50% of the data (the peripheral two 25% are not used)

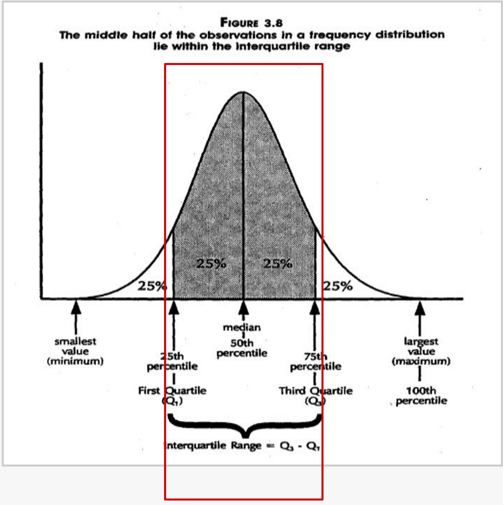


Figure 3.8
The middle half of the observations in a frequency distribution lie within the interquartile range

Percentile & Quartiles

- **Maximum is 100th percentile:** 100% of values lie at or below the maximum.
- **Median is 50th percentile:** 50% of values lie at or below the median.
- Any percentile can be calculated. But the most common are **25th (1st Quartile)** and **75th (3rd Quartile).**

Measures of dispersion / variability

Locating Percentiles in a Frequency Distribution

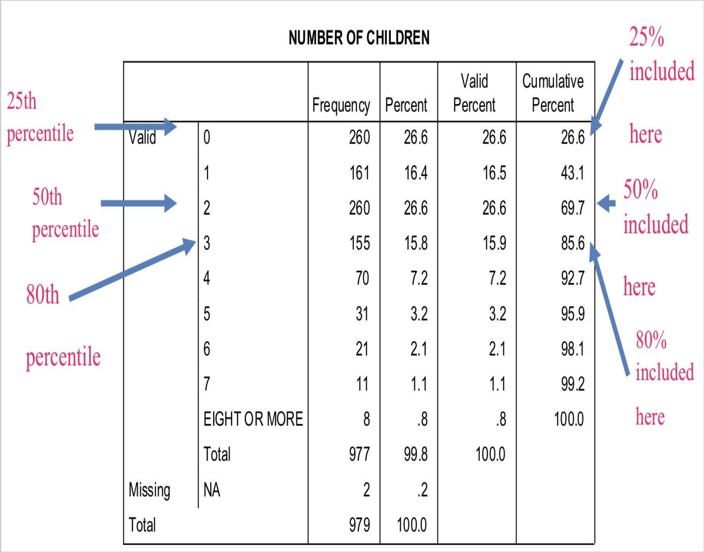
- A percentile is a score below which a specific percentage of the distribution falls (the median is the 50th percentile).
- The 75th percentile is a score below which 75% of the cases fall.
- The median is the 50th percentile: 50% of the cases fall below it.
- Another type of percentile :The quartile lower quartile is 25th percentile and the upper quartile is the 75th percentile.

979 is the total number of families in the this research

Notice NA - 2 (means missing data 2 families)

Notice that Valid percent exclude the missing data

- **What is the corresponding number of children for 25th percentile ?**
  0 children
- **What is the corresponding number of children for 50th percentile ?**
  2 children & less



NUMBER OF CHILDREN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	260	26.6	26.6	26.6
	1	161	16.4	16.5	43.1
	2	260	26.6	26.6	69.7
	3	155	15.8	15.9	85.6
	4	70	7.2	7.2	92.7
	5	31	3.2	3.2	95.9
	6	21	2.1	2.1	98.1
	7	11	1.1	1.1	99.2
	EIGHT OR MORE	8	.8	.8	100.0
	Total	977	99.8	100.0	
Missing	NA	2	.2		
Total		979	100.0		

25th percentile

50th percentile

80th percentile

25% included here

50% included here

80% included here

- **Frequency:** Number of families with the specific number of children, E.g. 161 families have one child.
- **Valid percent:** Percent when missing data are excluded from calculation. So in this schedule the valid percent is calculated after excluding 2 missing data.
- **Cumulative percent:** the valid percent + previous valid percents

Variance	Standard Deviation
Deviations of each observation from the mean, then averaging the sum of squares of these deviations.	**“ ROOT- MEANS-SQUARE-DEVIATIONS”.** - To “undo” the squaring of difference scores, take the square root of the variance. - Return to original units rather than squared units. - Measures the variation of a variable in the sample. - Technically: $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$

Variance and Standard deviation

Calculation of Variance & Standard deviation

Example

- Data: X = {6, 10, 5, 4, 9, 8};
- N = 6

Mean : $\bar{X} = \frac{\sum X}{N} = \frac{42}{6} = 7$

Variance: $s^2 = \frac{\sum(\bar{X} - X)^2}{N} = \frac{28}{6} = 4.67$

Standard Deviation: $s = \sqrt{s^2} = \sqrt{4.67} = 2.16$

	X	$X - \bar{X}$	$(X - \bar{X})^2$
	6	-1	1
	10	3	9
	5	-2	4
	4	-3	9
	9	2	4
	8	1	1
Total	42	0	28

Interpretation: All 6 values on average are deviating by 2.16. On average each student is different from other by 2.16.

Why we don't use variance?
The problem with variance is that the unit of value will be squared which will not make sense and it will be hard to use)

Using the deviation & computational method to calculate the variance and standard deviation

- 3,4,4,4,6,7,7,8,8,9 ; Given n=10; Sum= 60; Mean= 6

$$S = \sqrt{\frac{\sum(X-\bar{X})^2}{n}}$$

$$S = \sqrt{\frac{(3-6)^2+(4-6)^2+(4-6)^2+(4-6)^2+(6-6)^2+(7-6)^2+(7-6)^2+(8-6)^2+(8-6)^2+(9-6)^2}{10}}$$

$$S = \sqrt{\frac{40}{10}} = 2.0; variance = 4$$

$$S = \sqrt{\frac{n\sum X^2 - (\sum X)^2}{n^2}}$$

$$S = \sqrt{\frac{10(400) - (60)^2}{10^2}}$$

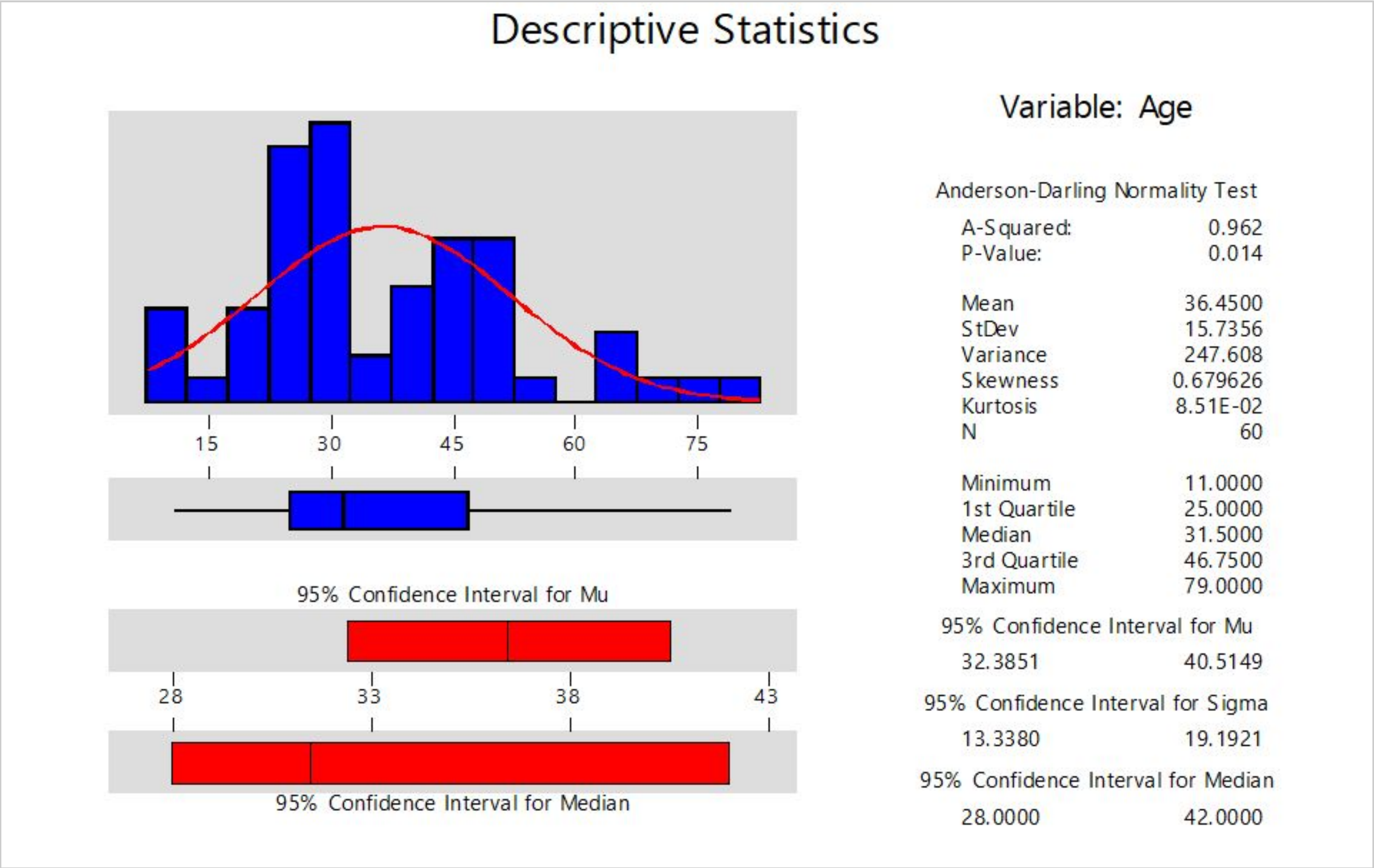
$$S = \sqrt{\frac{4000 - 3600}{100}}$$

$$S = \sqrt{4.0}$$

$$S = 2.0, \text{var } iance = 4$$

X	X^2
3	9
4	16
2	16
4	16
6	36
7	49
7	49
8	64
8	64
9	81
Sum:60	Sum:400

Measures of Dispersion / Variability



Descriptive Statistics

Variable: Age

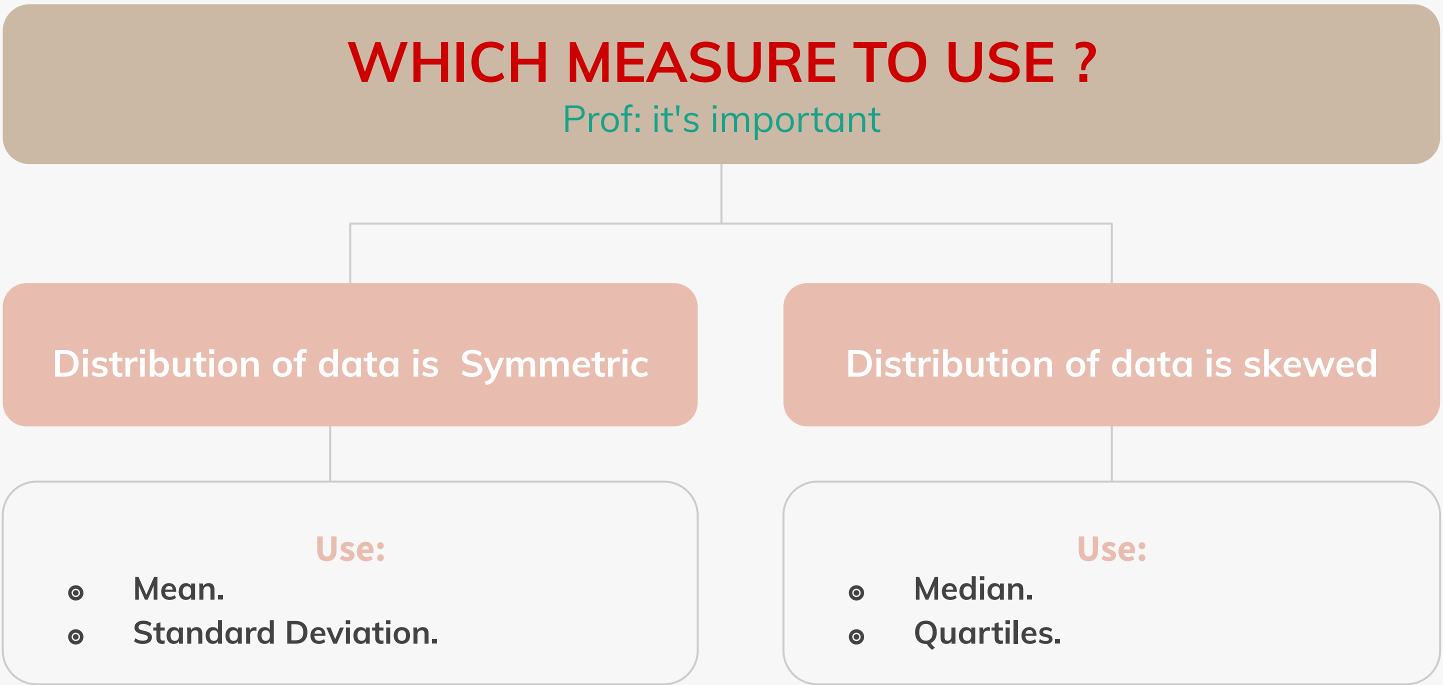
```
Anderson-Darling Normality Test
  A-Squared:                0.962
  P-Value:                  0.014

  Mean                    36.4500
  StDev                   15.7356
  Variance                247.608
  Skewness               0.679626
  Kurtosis               8.51E-02
  N                            60

  Minimum                 11.0000
  1st Quartile            25.0000
  Median                  31.5000
  3rd Quartile            46.7500
  Maximum                 79.0000
 95% Confidence Interval for Mu
  32.3851                 40.5149
95% Confidence Interval for Sigma
  13.3380                 19.1921
95% Confidence Interval for Median
  28.0000                 42.0000
```

95% Confidence Interval for Mu

95% Confidence Interval for Median



WHICH MEASURE TO USE ?

Prof: it's important

Distribution of data is Symmetric

Use:

- Mean.
- Standard Deviation.

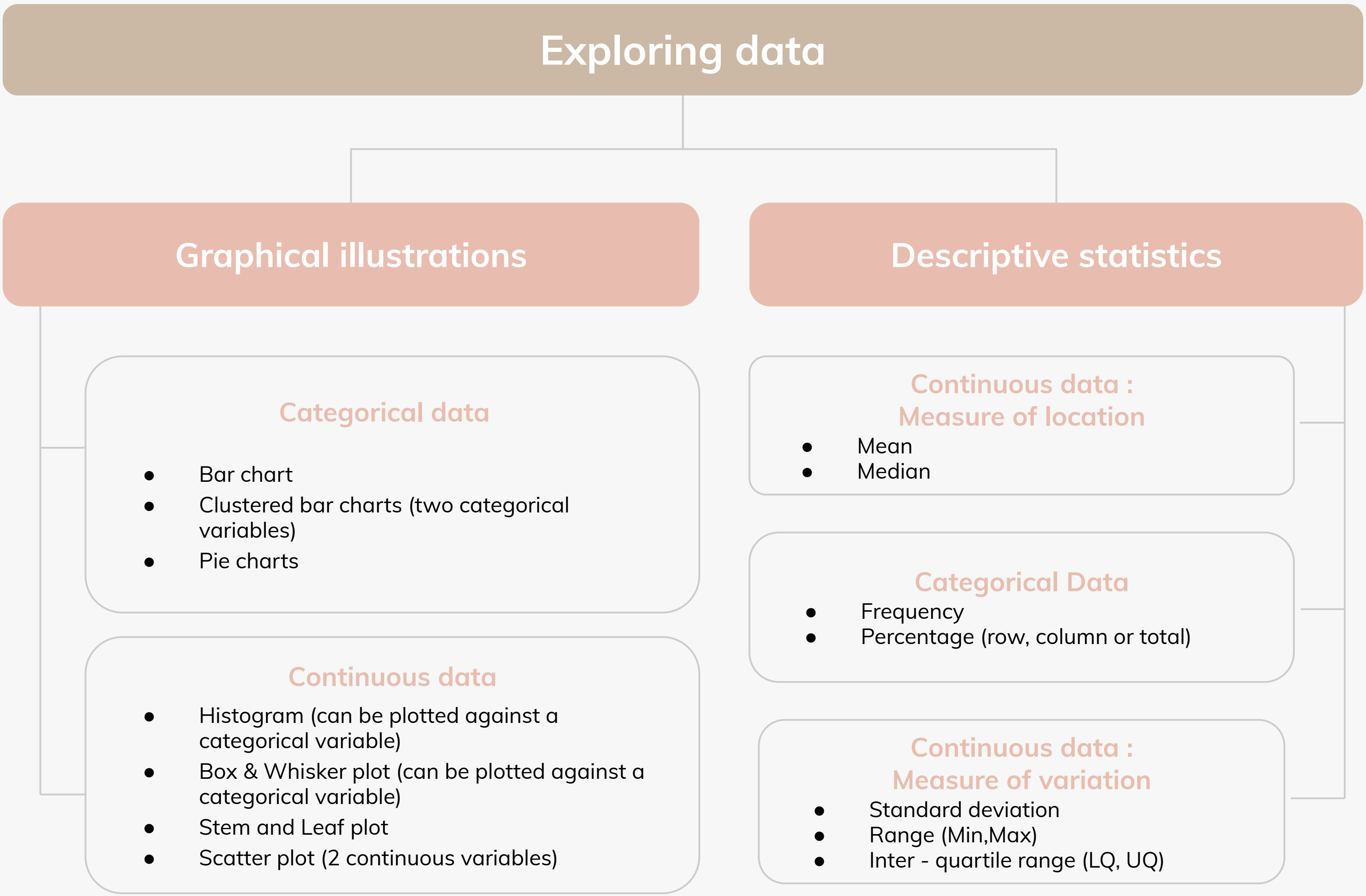
Distribution of data is skewed

Use:

- Median.
- Quartiles.

Measures of dispersion / variability

Flow chart of commonly used descriptive statistics and graphical illustrations



Exploring data

Graphical illustrations

Categorical data

- Bar chart
- Clustered bar charts (two categorical variables)
- Pie charts

Continuous data

- Histogram (can be plotted against a categorical variable)
- Box & Whisker plot (can be plotted against a categorical variable)
- Stem and Leaf plot
- Scatter plot (2 continuous variables)

Descriptive statistics

Continuous data : Measure of location

- Mean
- Median

Categorical Data

- Frequency
- Percentage (row, column or total)

Continuous data : Measure of variation

- Standard deviation
- Range (Min,Max)
- Inter - quartile range (LQ, UQ)

Dr’s Notes

- Another name of central tendency is measure of location and measure of average.
- Harmonic- Geometric mean, will not be used.
- Harmonics analyzer is used to provide a detailed analysis of the suspect source. Ex; When you take the average speed of a trip with different models of transportation.
- The geometric mean is most useful when numbers in the series are not independent of each other or if numbers tend to make large fluctuations. Ex: Measuring antibody titration.
- If you divide the data into 4 parts we call it Quartile and if we divide them into 100 parts we call it Percentile.
- Extreme value: means that it is away from the normal subjects.
- Another name for the median is second quartile (Q2).
- The mean is affected by extreme value, where the median does not get affected by extreme values.
- Skewed means becoming narrow on one side (left or right).
- Positively skewed: the mean is higher than mode.
- Negatively skewed: the mode is higher than mean.
- Measure of variability, dispersion, Heterogeneity all are same.
- Range: it’s not good mathematically because it depend only on 2 observations.
- Interquartile range used for skewed data.
- SD: in average how much each value differ from the mean value in a data set.
- We are using only the SD, variance is not used because it is not justifying data.

Research Team

RESEARCH TEAM
438
Keep it simple, Keep it focused

Thank you for checking our work!

Leaders:

Shuaa Khdary   Sarah AlQuwayz

Abdulrhman Alsuhaibany

Member:

Sara Alharbi

Note Taker:

Lama Alahmadi   Abdulaziz Alderaywsh

Contact us:

Research4390@gmail.com