Workshop 2
How to interpret and unitize elementary statistics

Dr. Naser Alziyadat

Murdoch
UNIVERSITY

Two types of statistics

1. Descriptive statistics.

2. Inferential statistics.

What are the differences?

What do I mean when I ask someone to describe himself/herself?

DESCRIPTIVE STATISTICS - ORGANISING THE DATA

Murdoch UNIVERSITY

- They often involve calculating as well as producing graphical displays of (inter alia)
  - frequencies of different occurrences,
  - central location of a distribution,
  - spread of a distribution,
  - shape of a distribution

Four Common Descriptive 'Summary' Measures:

- *The distribution of* ***frequencies***
- *Measures of the* ***central tendency*** of the data
- *Measures of* ***dispersion, spread or variability***
- ***Skewness*** *or* ***normality of spread***

Measures of Central Tendency

'Measures of central tendency' are also referred to as 'averages'.

The purpose of a measure of central tendency is to provide a single value which summarises a variable.

MEASURES OF CENTRAL TENDENCY

- ***The Mode***

  The most frequently occurring score value

- ***The Median***

  The middle value when scores are placed in rank order

- ***The Mean***

  the sum of all the scores divided by the number of scores i.e. the arithmetical average

Measures of Central Tendency: Mode

Murdoch UNIVERSITY

Mode: the most commonly occurring observation/value/case.

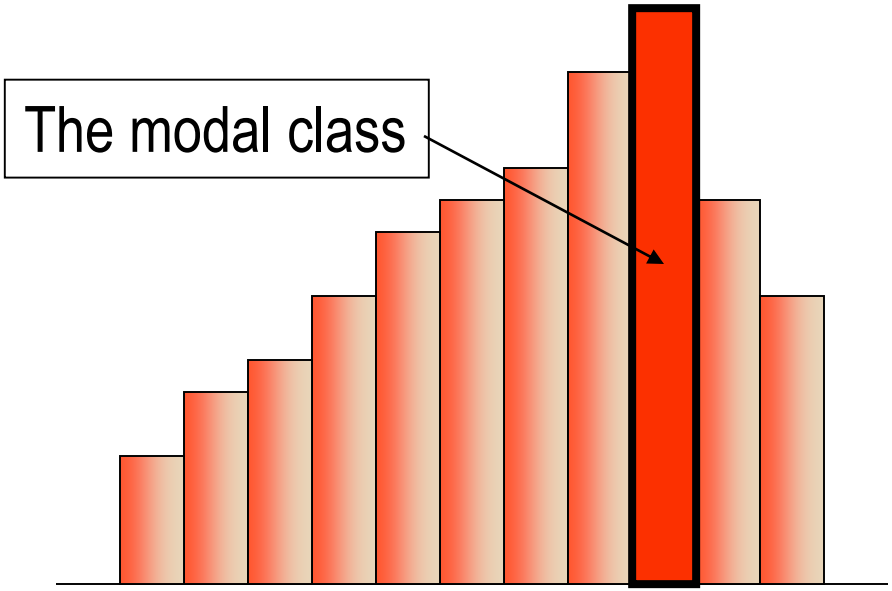
The mode (or modal class) can be calculated for **nominal**, **ordinal**, **interval**, or **ratio** scale data.

e.g. 1, 1, 1, 4, 5, 6, 1000. Mode = 1

With a variable measured on interval or ratio scales, there may be no two values which are identical. In this case a '**modal class**' may be calculated by first grouping your cases into ranges (e.g. 0-18 years, 19-29 years, 30-39 years etc) and determining the category with the largest number of cases (the 'modal class').

Measures of Central Tendency: Mode

- The *mode* of a set of measurements is the value that occurs most frequently.
- A set of data may have one mode (or modal class), or two or more modes.



Task 1

- Access the data file ICSR.SAV.
- Describe the variables age and level of education.
- - Procedure: Analyze-Descriptive Statistics-Frequencies- Select the variables and move them to the right window- Click OK.

Interpreting the results

Murdoch UNIVERSITY

Check your variables

Check the number of observations.

Statistics

		What is your age at the time of this interview?	What is your level of education?
N	Valid	27	27
	Missing	0	0

The portion of that group out of the total

Make sure that the categories are correct

What is your age at the time of this interview?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	24-29	3	11.1	11.1	11.1
	30-35	8	29.6	29.6	40.7
	36-41	7	25.9	25.9	66.7
	42and over	9	33.3	33.3	100.0
	Total	27	100.0	100.0	

Sometimes SPSS disregards some data. The valid percept tells you that

The percept of this group + the percent of the previous groups.

What is this?

How do we measure the mode in SPSS?

Measures of Central Tendency: Median

Murdoch UNIVERSITY

Median: the middle value in an ordered array.

The median can be calculated for **ordinal**, **interval**, or **ratio** scale data.
Used instead of mean for skewed distributions like that below.

e.g. 1, 1, 1, 4, 5, 6, 1000.

Median = 4

The Median

Example

Seven employee salaries were recorded (in 1000s) : 28, 109, 26, 32, 30, 26, 29. Find the median salary.

Odd number of observations

26,26,28,29,30,32,109

Example

Suppose one employee's salary of $31 000 was added to the group recorded before. Find the median salary.

Even number of observations

There are two middle values!

26,26,28,29,30,31,32,109

26,26,28,29,29.5,30,31,32,109

Measures of Central Tendency: Arithmetic Mean

Mean: measure of the central data point (the sum of the measures in the set divided by the number of scores in the set). Common name is 'average'

The arithmetic mean can be calculated for **interval** or **ratio** scale data.

e.g. 1, 1, 1, 4, 5, 6, 1000.

Mean = 145.4

The SAMPLE MEAN

If our sample comprises the scores: 1, 3, 4, 5, 7, 10

Then

$\Sigma x_i = 30$

$n = 6$ $\quad$ M = 30/6 = 5

The Mean

- *Characteristics:*

  determined by the value of every score

  amenable to arithmetic and algebraic manipulations

- *Problems with the mean:*

  when the distribution is very skewed it provides an inaccurate picture of where the central values are

  when the data are qualitative in character it cannot be calculated

Central Tendency and Levels of Measurement

Murdoch UNIVERSITY

- **nominal scale**:

  the mode is the only legitimate statistic to use.

- **ordinal scale**:

  median preferred over the mean which could be distorted by an extreme score

- **interval and ratio scales (grouped together as Scale level in SPSS)**:

  the mean is the recommended measure of central tendency, median & mode may also be reported for these types of scales

MEASURES OF VARIABILITY

- **Range**

  range between the lowest and highest scores

  is considerably influenced by extreme scores

- **Variance**

  incorporates all scores in the distribution

  $$V = \frac{\Sigma (X - M)^2}{N}$$

- **Standard Deviation**

  reflects the amount of spread that the scores exhibit around the mean. It is the square root of the Variance but using N-1 as the denominator

Measures of Variability: Range

Murdoch UNIVERSITY

The *range* of a set of measurements is the difference between the largest and smallest measurements.

Its major advantage is the ease with which it can be computed, and???!

Its major shortcoming is its failure to provide information on the dispersion of the values between the two end points. It is influenced totally by the values of the smallest and largest scores

Measures of Variability: Variance

This measure of variability (or dispersion) reflects the values of ***all* the measurements**.

The variance of a sample of n measurements $x_1, x_2, \ldots, x_n$ having a mean is defined as $\bar{x}$

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

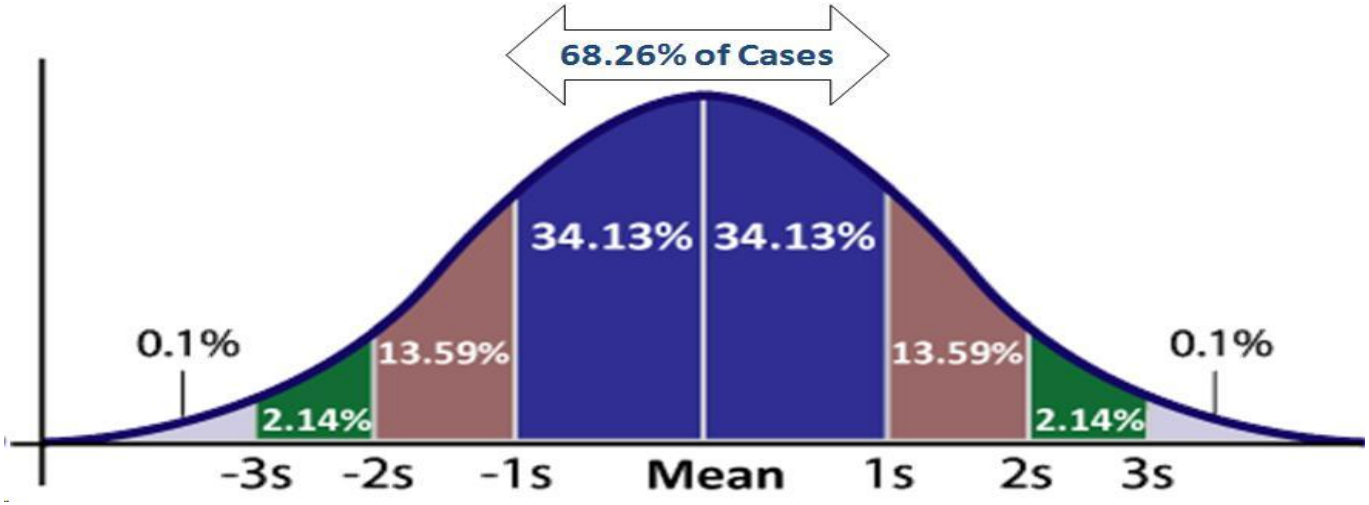
Standard Deviation

The *standard deviation* of a set of values is the **square root of the variance** of the measurements.

i.e the square root of the sum of the squared differences between each score and the mean divided by N - 1.

- The standard deviation is more commonly reported than the variance.
- It is calculated from all the values in a data set, representing the dispersal round the mean outward in each direction.

Interpreting the SD



- 68% of the values occur within (-1s to 1s).
- 95% of the values occur within (-2s to 2s).
- 99% of the values occur within (-3s to 3s).

Task 2

- Access data file ICSR from the LMS.
- We want to have the descriptive statistics of the variable ICSR2.

  First: Decide the type of measurement. Is it ordinal, nominal or scale? In our case it is scale.

  Second: go to analyze in the menu bar>> Descriptive Statistics>>> Explore>>> Move the variable ICSR2 to the dependent list>>> Click on Statistics, and ensure Descriptive is chosen>>> Click on Plots, untick steam-and-leaf, then tick Normality plots with tests>>>Click continue>> then OK.

Interpreting the results

Number of observations

Number of observations regarded by SPSS

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ICSR2	27	100.0%	0	0.0%	27	100.0%

Descriptives

			Statistic	Std. Error
ICSR2	Mean		23.5556	1.07990
	95% Confidence Interval for Mean	Lower Bound	21.3358	
		Upper Bound	25.7753	
	5% Trimmed Mean		23.5473	
	Median		24.0000	
	Variance		31.487	
	Std. Deviation		5.61134	
	Minimum		13.00	
	Maximum		34.00	
	Range		21.00	
	Interquartile Range		9.00	
	Skewness		-.064	.448
	Kurtosis		-.784	.872

What is this?

Where is the mode?

The estimated difference between the mean of the population and the mean of the sample

We are 95% confident that the mean is between those two bounds

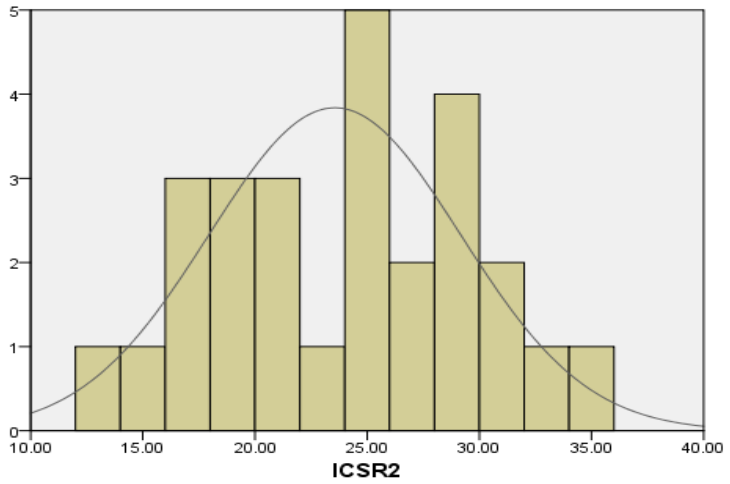
The mean without 5% of the observations in the upper and lower ends of the data

Interpreting the results + Histogram

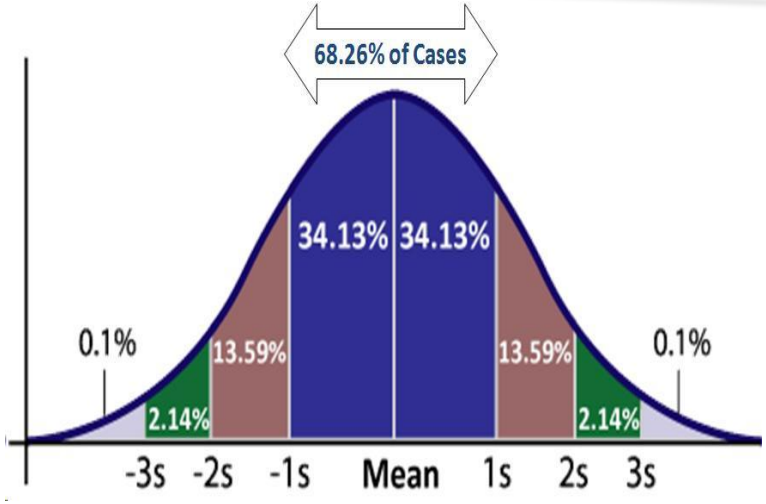
Go to Graphs in the menu>>> select Graphboard Template Chooser>>> Select the same variable of the previous example>>>Make sure the Basic tab is selected>>>Select Histogram with normal distribution>>>OK.

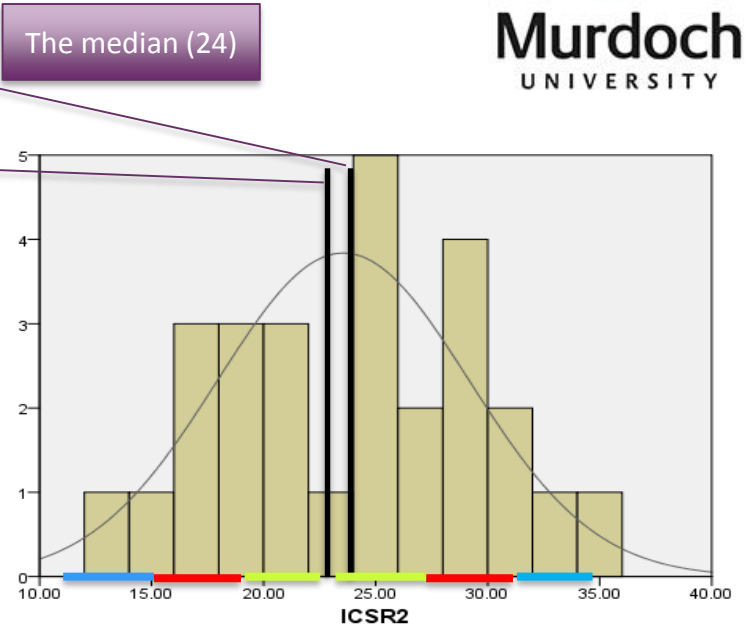
Descriptives

			Statistic	Std. Error
ICSR2	Mean		23.5556	1.07990
	95% Confidence Interval for Mean	Lower Bound	21.3358	
		Upper Bound	25.7753	
	5% Trimmed Mean		23.5473	
	Median		24.0000	
	Variance		31.487	
	Std. Deviation		5.61134	
	Minimum		13.00	
	Maximum		34.00	
	Range		21.00	
	Interquartile Range		9.00	
	Skewness		-.064	.448
	Kurtosis		-.784	.872



Interpreting the SD



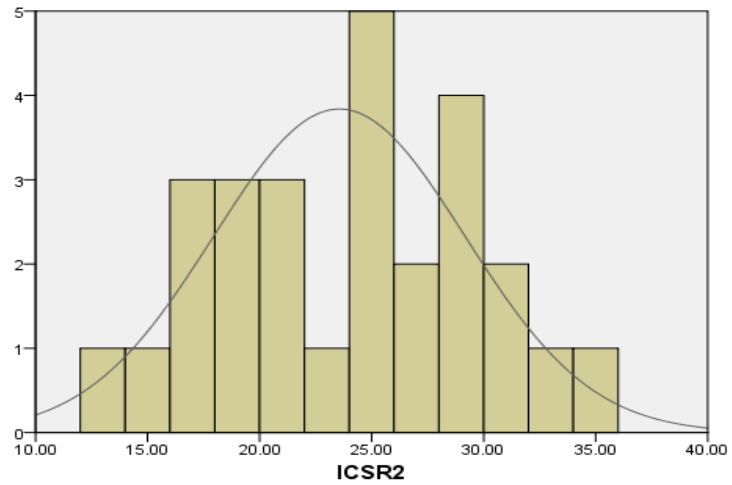


- 68% of the values occur within (-1s to 1s).
- 95% of the values occur within (-2s to 2s).
- 99% of the values occur within (-3s to 3s).

How to report it?

Descriptives

			Statistic	Std. Error
ICSR2	Mean		23.5556	1.07990
	95% Confidence Interval for Mean	Lower Bound	21.3358	
		Upper Bound	25.7753	
	5% Trimmed Mean		23.5473	
	Median		24.0000	
	Variance		31.487	
	Std. Deviation		5.61134	
	Minimum		13.00	
	Maximum		34.00	
	Range		21.00	
	Interquartile Range		9.00	
	Skewness		-.064	.448
	Kurtosis		-.784	.872



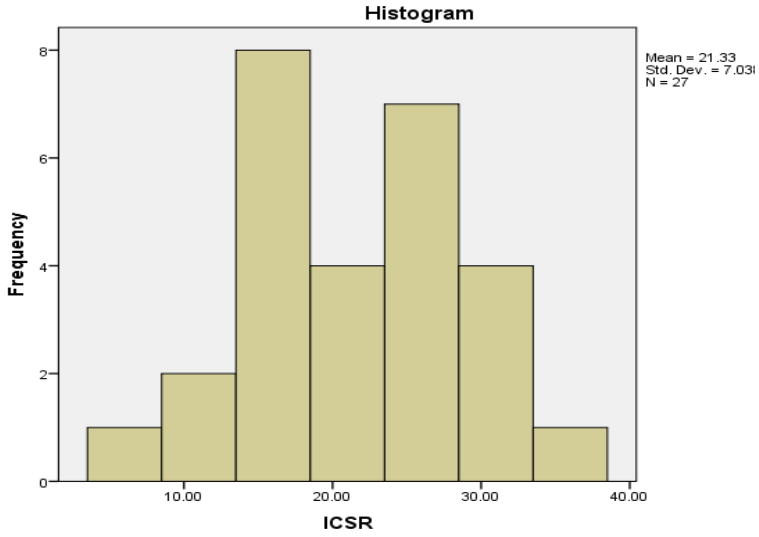
Descriptive statistics indicates that the scores of ICSR2 have a mean of 23.6 with 1.08 standard error from the mean of the population. A confidence interval of 95% of the mean makes it located between the upper bound of 25.8 and the lower bound of 21.3. The trimmed mean value is very closed to the mean value explaining the low influence of the extreme values. With a value of 24 that is close to the mean, the distribution of the scores is skewed slightly to the upper end of the scores with a value of -.06 as shown in the figure bellow. The standard deviation value is 5.6 can give more insights about the shape of the distribution when taking into consideration the minimum and maximum scores of 13 and 34 respectively.

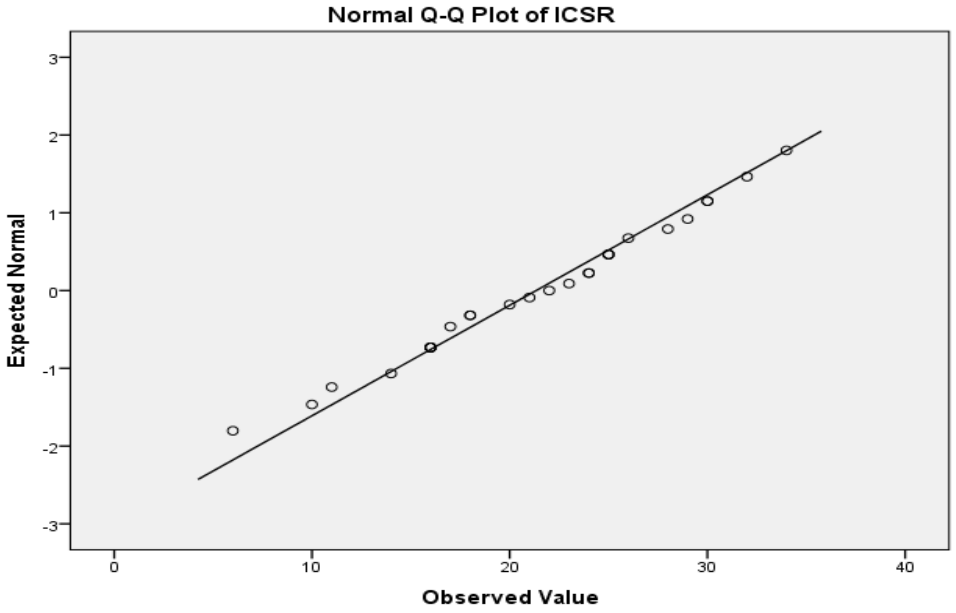
More about normality

- Assessing normality
  - Many of the statistical techniques assumes that the data are normally distributed (normality)
  - Normality concerns the shape of the distribution of the data (bell shape).
  - Procedure- Analyze-Descriptive statistics-explore-select ICSR2 and move it to the independent list-Statistics section>> select descriptive>>> Plot section>> Select histograms, untick Stem and Leaf, and select normality plots with tests>>>Options section, tick exclude case pairwise>>> continue and OK

Normality assessment

- Assessing normality
  - Practice: provide normality assessment for the variable ICSR.

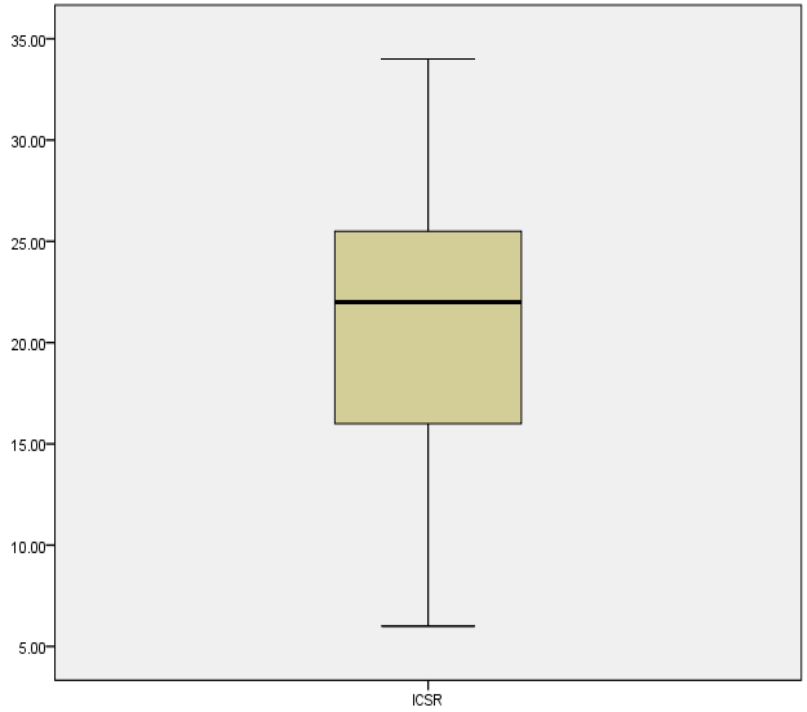




Histogram : shows that scores have been close to normal distribution.
Normal Q-Q Plots: The observed value of each score is plotted against the expected value from the normal distribution. A reasonably straight line suggests a normal distribution.

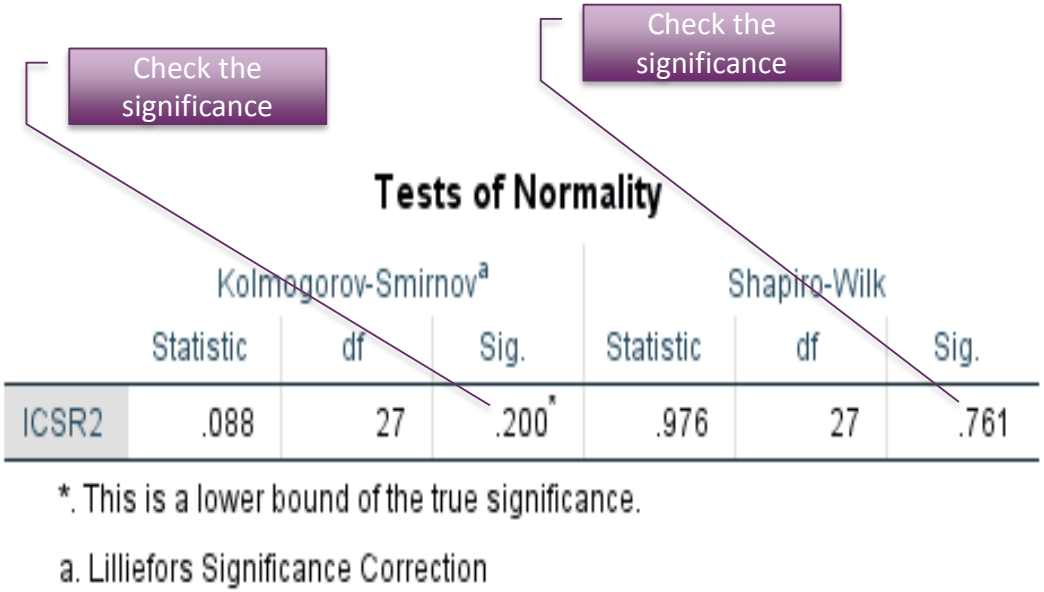
Normality assessment

- Assessing normality
  - Practice: provide normality assessment for the variable ICSR.



Boxplot: the rectangle represents 50% of the cases, the line going up to the largest and the line going down to the smallest.
This test is good for outliers. In our case we do not have outliers, but when they are present, SPSS will show them as small circles above the line of the largest cases and bellow the line of the smallest cases.

Normality assessment



In addition to all that we has explained about normality assessment. The most important tests are Kolmogrov-Samirnov and Shapiro-Wilk tests.

In this regard, we need to check the significance of the two tests. If they are less than 0.05 then the data is NOT normally distributed. Otherwise, the data is normally distributed.

Descriptive statistics indicates that the scores of ICSR2 have a mean of 23.6 with 1.08 standard error from the mean of the population. A confidence interval of 95% of the mean makes it located between the upper bound of 25.8 and the lower bound of 21.3. The trimmed mean value is very closed to the mean value explaining the low influence of the extreme values. With a value of 24 that is close to the mean, the distribution of the scores is skewed slightly to the upper end of the scores with a value of -.06 as shown in the figure bellow. The standard deviation value is 5.6 can give more insights about the shape of the distribution when taking into consideration the minimum and maximum scores of 13 and 34 respectively. With non-significant results of Kolmogrov-Samirnov and Shapiro-Wilk tests, the data is almost normally distributed.

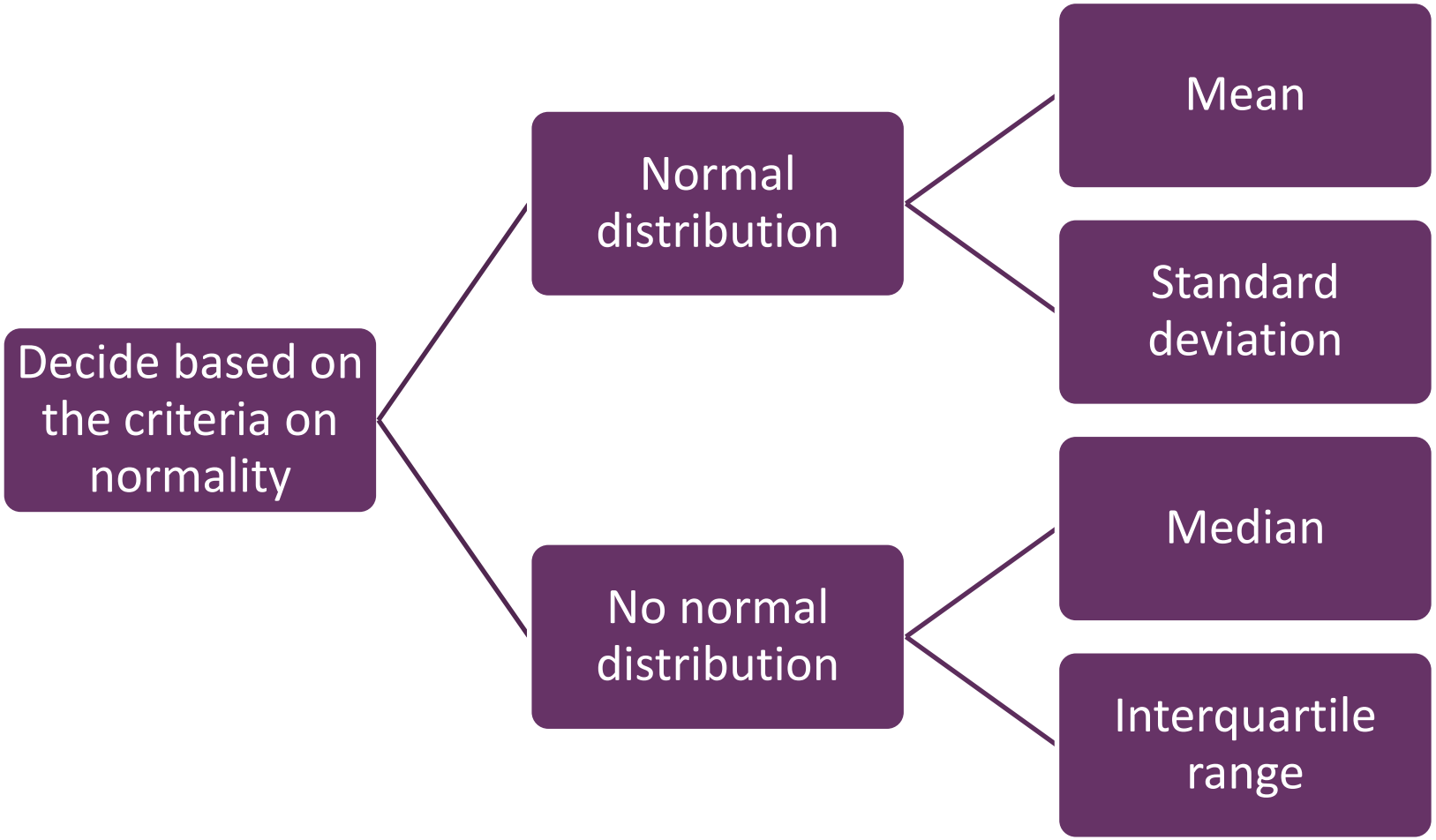
Criteria to decide on normality

- Histogram: the distribution of the data according to the normal distribution curve.
- Q-Q Plot: the distribution of the data according to the line.
- Skewness value: There is no cut point; however, some statisticians may consider the value of ±0.8 as a factor to consider when compromising all the criteria mentioned in this slide.
- The difference between the mean and the median. This should be considered along with the minimum and maximum scores of the variabvle to decide on how large is that difference between the mean and the median.
- Kolmogrov-Samirnov and Shapiro-Wilk tests are not always reliable. Thus, their results must be compromised with the above criteria.
- ***Deciding on normality is a bit subjective***

What to report

- Nominal variables: mode – frequencies – Pie or bar chart - a paragraph.
- Ordinal variables: mode – median – frequencies – Bar chart- a paragraph.
- Scale variables: Descriptive Table> Mean, lower and upper bounds of the mean, trimmed mean, median, skewness, kurtoses, histogram>> Q_Q Plot, Kolmogrov-Samirnov and Shapiro-Wilk tests>>>standard deviation>>Interquartile range.

How to describe a 'scale' variable



Murdoch
UNIVERSITY

Thank you
Please refer to the LMS and complete the tasks after the workshop