Basic statistics applied to DU research

Marike Cockeran

2015

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Outline

- Measurement scales
- Basic descriptive statistics
- Comparing groups of continuous data
- Comparing groups of categorical data
- Determining a linear relationship between two continuous variables

Measurement scales

Marike Cockeran

2015

Outline

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

Introduction

- Measurement scales are used to classify variables or types of data:
  - Nominal scale
  - Ordinal scale
  - Interval scale
  - Ratio scale

- It is important to know on what type of scale a variable is measured since certain statistical analysis is only applicable on variables measured on a certain measurement scale.

- Example:
  - It does not make sense to calculate the average gender of a sample.
  - However you can calculate the average weight or height of a sample.

Objectives

After completion of this study section you should be able to:

- Distinguish between the different measurement scales:
  - Nominal scale
  - Ordinal scale
  - Interval scale
  - Ratio scale

Nominal scale

- Values (numbers) are assigned to different categories of a variable.
- Example:
  - The variable gender has two categories: male and female.
  - The number one is assigned to the male category.
  - The number two is assigned to the female category.
- The sequence of the values is not important.
- The numbers serve as labels for the different categories.
- The categories do not overlap.
- If a variable is measured on a nominal scale and takes on one of two distinct values, the variable is called a dichotomous or binary variable.

Ordinal scale

- Values (numbers) are assigned to different categories of a variable, but categories now have an ordered relationship to one another.

- Example: Variable measuring physical activity
  - 1 – low physical activity
  - 2 – medium physical activity
  - 3 – high physical activity

- The categories do not overlap.

Interval scale

- Each value on the scale has a unique meaning.
- Values have an ordered relationship to one another.
- Scale units along the scale are equal to one another.
- The scale does not have a true zero point.
  - Zero does not represent the absolute lowest value.
  - It is a point on the scale with numbers both above and below it.
- Example: Temperature

Ratio scale

- Each value on the scale has a unique meaning.
- Values have an ordered relationship to one another.
- Scale units along the scale are equal to one another.
- The scale has a true zero point.
  - Zero presents the absolute lowest value.
- Example: Weight

Additional remarks

- **Discrete data**
  - Data are restricted to taking on only specified values – often integers.
  - Fractional values are not possible.
  - Example: The number of new cases of tuberculosis reported.

- **Continuous data**
  - Data are not restricted to taking on certain specified values.
  - Fractional values are possible.
  - Example: Serum cholesterol level of a patient.

Descriptive statistics

Marike Cockeran

2015

Outline

- Measures of location
  - Arithmetic mean
  - Median
  - Mode
- Measures of spread
  - Range
  - Interquartile range
  - Variance and standard deviation
  - Coefficient of variation
- Possible distributions of data values
- Histogram
- Boxplot

Measures of location

- Descriptive statistics are a means of organising and summarising observations.

- The most commonly investigated characteristic of a set of data is its center, or the point about which the observations tend to cluster.

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the center of the dataset.

Objectives

After completion of this study section you must be able to:

- Describe the properties of different measures of locations:
  - Arithmetic mean
  - Median
  - Mode
- Interpret the output of these measures of location.

Arithmetic mean

- The mean is calculated by summing all the observations in a set of data and dividing by the total number of measurements.

$$\bar{X} = \sum_{i=1}^{n} X_i$$

- Properties of the mean:
  - The mean takes into consideration the magnitude of every observation in a set of data.
  - This causes the mean to be extremely sensitive to unusual values.

Arithmetic mean: Example

- HDL cholesterol values of 7 women:

Dataset 1	Dataset 2
1.30	1.30
1.38	1.38
1.42	1.42
1.58	1.58
1.61	1.61
1.45	1.45
1.57	5.17

- Average value of Dataset 1: $\bar{X} = 1.47$
- Average value of Dataset 2: $\bar{X} = 1.99$

Median

- If a list of observations is ranked from smallest to largest, half the values are greater than or equal to the median, whereas the other half are less than or equal to it.

- The median is the middle value of an ordered dataset.

- Properties of the median:
  - The median takes into consideration only the ordering and relative magnitude of observations.
  - The median is less sensitive to unusual data points.

Median: Example

- HDL cholesterol values of 7 women:

Dataset 1	Dataset 2
1.30	1.30
1.38	1.38
1.42	1.42
1.45	1.45
1.57	1.58
1.58	1.61
1.61	5.17

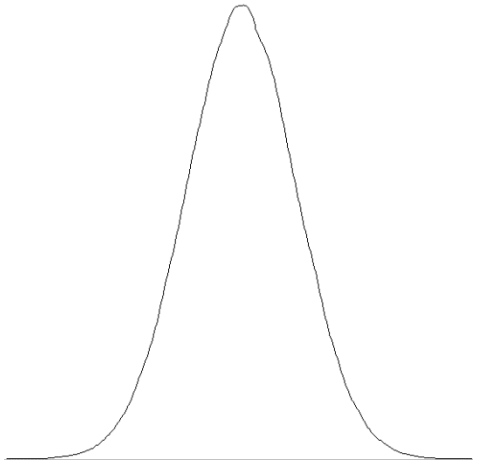
- Average value for Dataset 1: $\bar{X} = 1.47$
- Median value for Dataset 1: $m = 1.45$

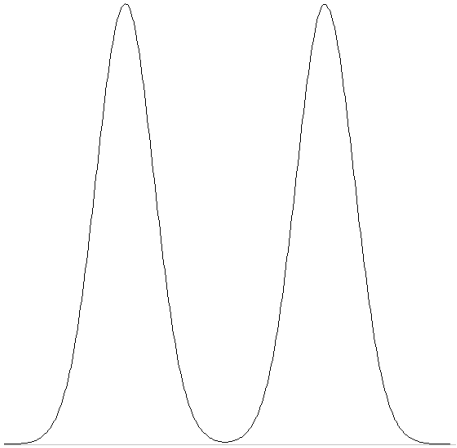
- Average value for Dataset 2: $\bar{X} = 1.99$
- Median value for Dataset 2: $m = 1.45$

Mode

- The mode of a set of values is the observation that occurs most frequently.

- Unimodal vs. Bimodal

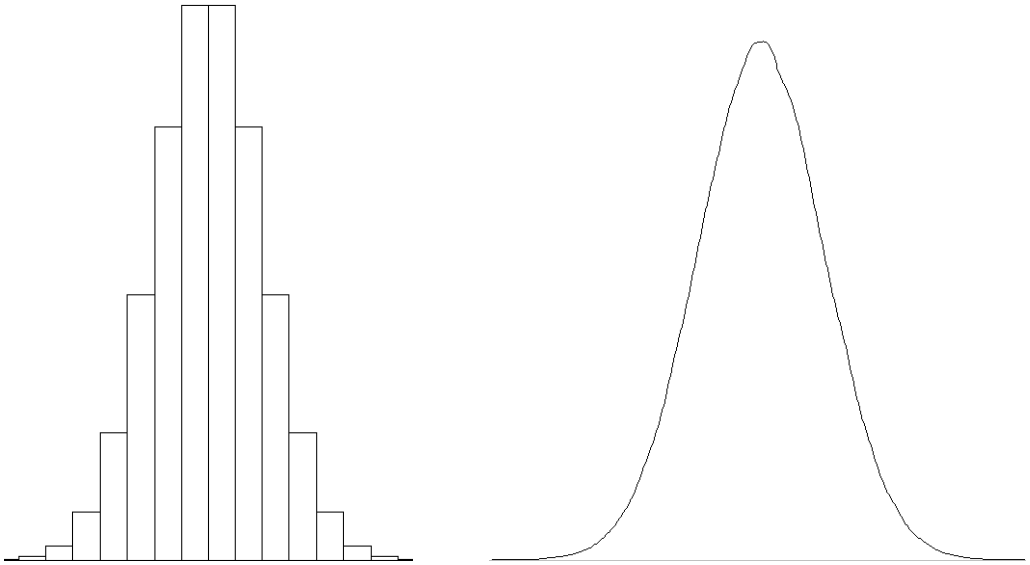




Symmetric and unimodel distributions

- The best measure of central tendency depends on the way in which the values are distributed.

- If they are symmetric and unimodel then the mean, median and the mode should all be roughly the same.

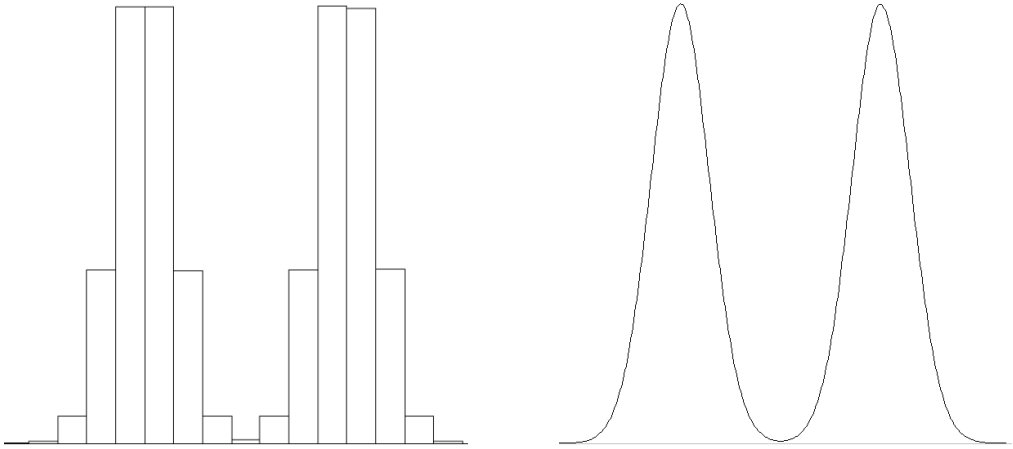


Symmetric and bimodal distributions

- If the distribution of values is symmetric but bimodal, then the mean and median should be approximately the same.

- However the value could lie between the two peaks and is extremely unlikely to occur.

- A bimodal distribution often indicates that the population from which the values are taken actually consists of two distinct subgroups.

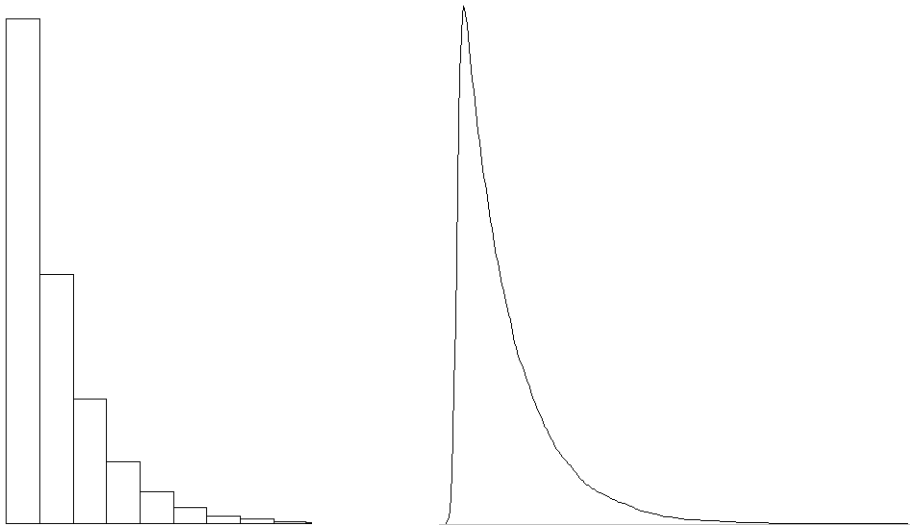


Right skewed distributions

- If the data are skewed to the right, the mean lies to the right of the median.

- When the data are not symmetric, the median is often the best measure of central tendency.

- Since the mean is sensitive to extreme observations, it is pulled in the direction of the outlying data.

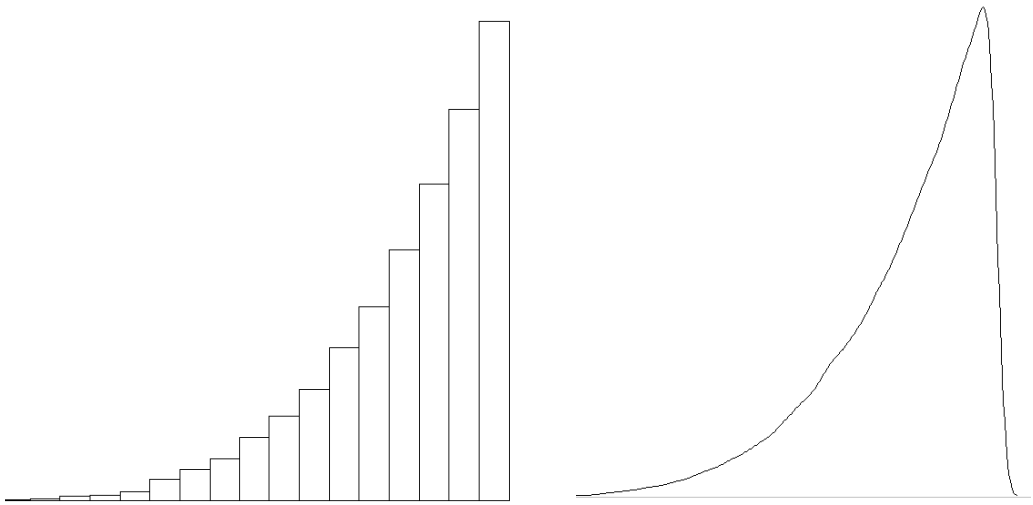


Left skewed distributions

- If the data are skewed to the left, the mean lies to the left of the median.

- When the data are not symmetric, the median is often the best measure of central tendency.

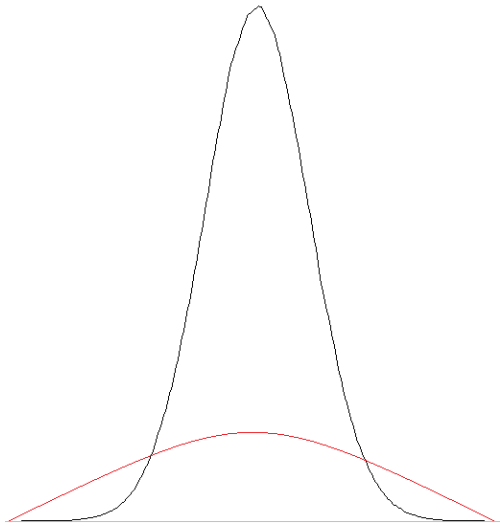
- Since the mean is sensitive to extreme observations, it is pulled in the direction of the outlying data.



Measures of spread

- To know how good our measure of central tendency is, we need to have some idea about the variation among the measurements.

- Do all the observations tend to be quite similar and therefore lie close to the center, or are they spread out across a broad range of values?



Objectives

After completion of this study section you must be able to:

- Describe the properties of different measures of spread:
  - Range
  - Interquartile range
  - Variance and standard deviation
  - Coefficient of variation

- Interpret the output of these measures of spread.

Range

- The range is defined as the difference between the largest observation and the smallest observation.

- The usefulness of the range as a measure of spread is limited since it considers only the extreme values rather than the majority of the observations.

Interquartile range

- The interquartile range is calculated by subtracting the 25^{th} percentile of the data from the 75^{th} percentile.

- The interquartile range includes the middle 50% of the observations.

Variance

- The variance quantifies the amount of variability, or spread, around the mean of the measurements.

- The variance of a set of observations is defined as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Standard deviation

- The standard deviation of a set of observations is the square root of the variance.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- The standard deviation has the same units of measurement as the mean, rather than squared units.
- In a comparison of two groups of data, the group with the smaller standard deviation has the more homogeneous observations and the group with the larger standard deviation exhibits a greater amount of variability.
- The magnitude of the standard deviation depends on the values in the dataset – what is large for one group of data may be small for another.
- Since the standard deviation has units of measurement, it is meaningless to compare standard deviations for two unrelated quantities.

Coefficient of variation

- The coefficient of variation relates the standard deviation of a set of values to its mean.
- The coefficient of variation is the ratio of s to $\bar{X}$ multiplied by 100

$$CV = \frac{s}{\bar{X}} \times 100\%$$

- It is therefore, a measure of relative variability.
- The coefficient of variation is most useful for comparing two or more sets of data.
- Since it is independent of measurement units, it can be used to evaluate the relative variation between any two sets of observations.

Measures of spread: Example

Patient	Heart rate (beats per minute)
1	167
2	150
3	125
4	120
5	150
6	150
7	40
8	136
9	120
10	150

With outlier	Without outlier
$\bar{X} = 130.8$	$\bar{X} = 140.89$
$m = 143$	$m = 150$
$sd = 35.47$	$sd = 16.44$

Grouped mean and grouped variance

Cholesterol level (mg/100 ml)	Number of men
80-119	13
120-159	150
160-199	442
200-239	299
240-279	115
280-319	34
320-359	9
360-399	5
Total	1067

$$\bar{X} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} \qquad s^2 = \frac{\sum_{i=1}^{k} (m_i - \bar{X})^2 f_i}{\left[\sum_{i=1}^{k} f_i\right] - 1}$$

Example: Grouped mean

Cholesterol level (mg/100 ml)	Number of men
80-119	13
120-159	150
160-199	442
200-239	299
240-279	115
280-319	34
320-359	9
360-399	5
Total	**1067**

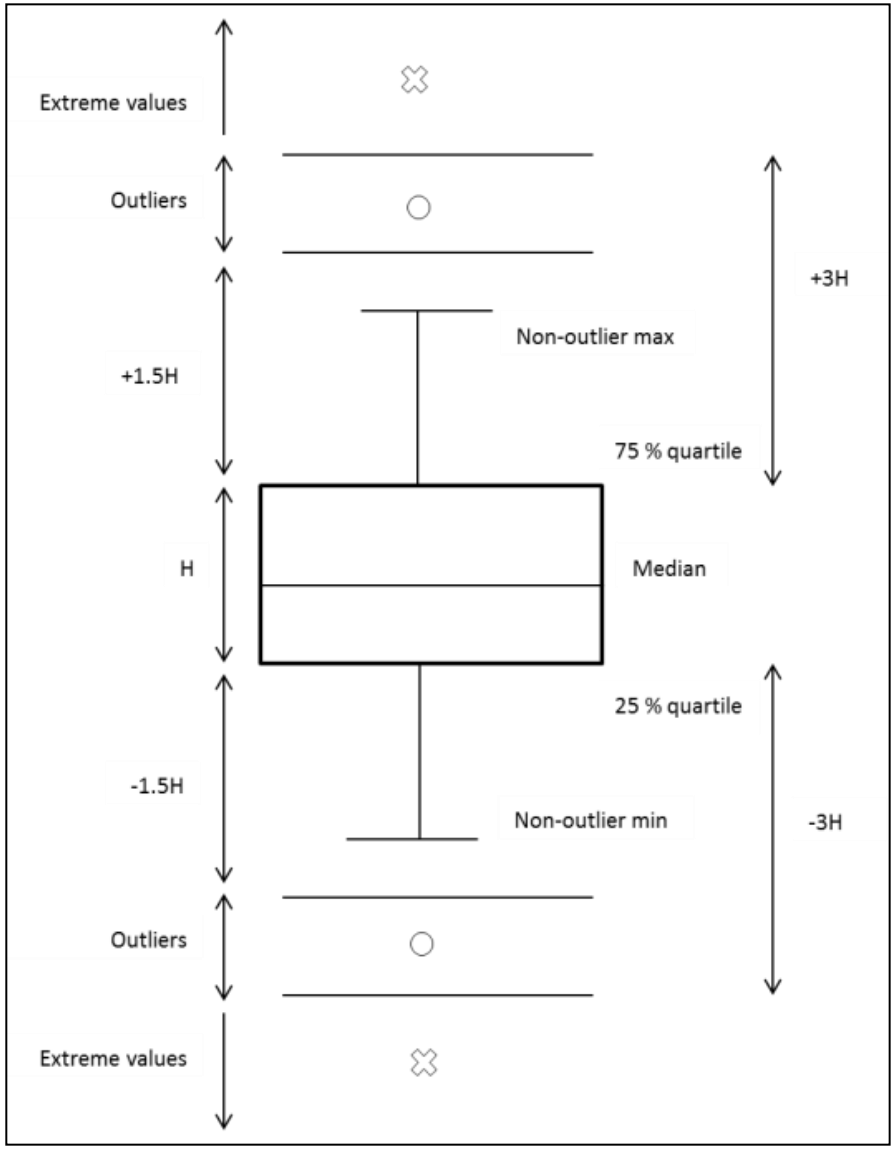
$$\frac{[99.5(13) + 139.5(15) + 179.5(442) + 219.5(299) + 259.5(115) + 299.5(34) + 339.5(9) + 379.5(5)]}{1067}$$

$$= 198.8$$

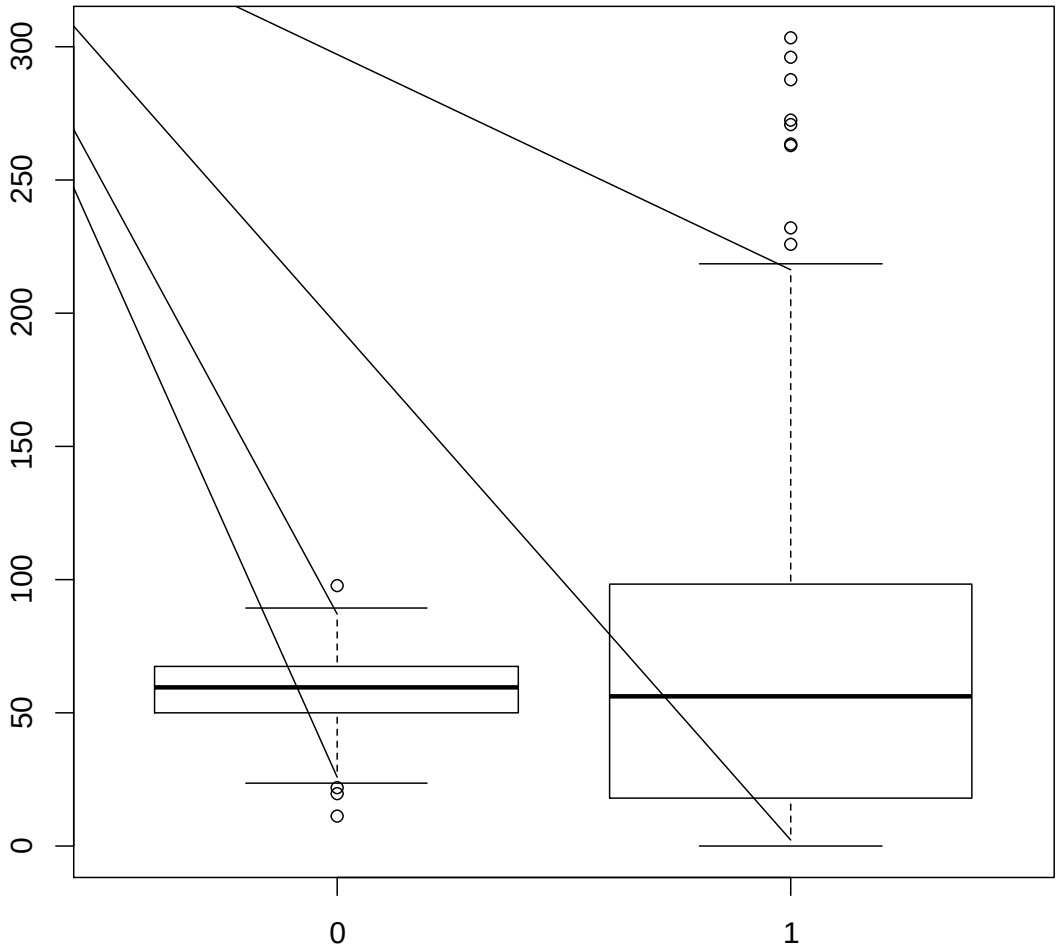
Box plot

- The box and whisker plot is a graphical representation of the numerical summary measures calculated in the previous section.
- The box plot shows the centre, spread and skewness of a dataset.
- The median (50^{th} percentile) is indicated by a vertical line, within the box.
- The lower and upper quartiles (25^{th} and 75^{th} percentiles) are indicated by the corresponding vertical ends of the box.
- The box thereby encloses the interquartile range, sometimes referred to as the *H* spread.
- The minimum and maximum non-outlier values are indicated by the whiskers.
- An outlier is beyond the whisker but less than three interquartile ranges from the box edge and finally an extreme value is more than three interquartile ranges from the box edge.
- It should be noted that the preceding explanation is merely one way to define a box and whisker plot.

Box plot


Extreme values
Outliers
+1.5H
+3H
Non-outlier max
75 % quartile
H
Median
25 % quartile
-1.5H
Non-outlier min
-3H
Outliers
Extreme values

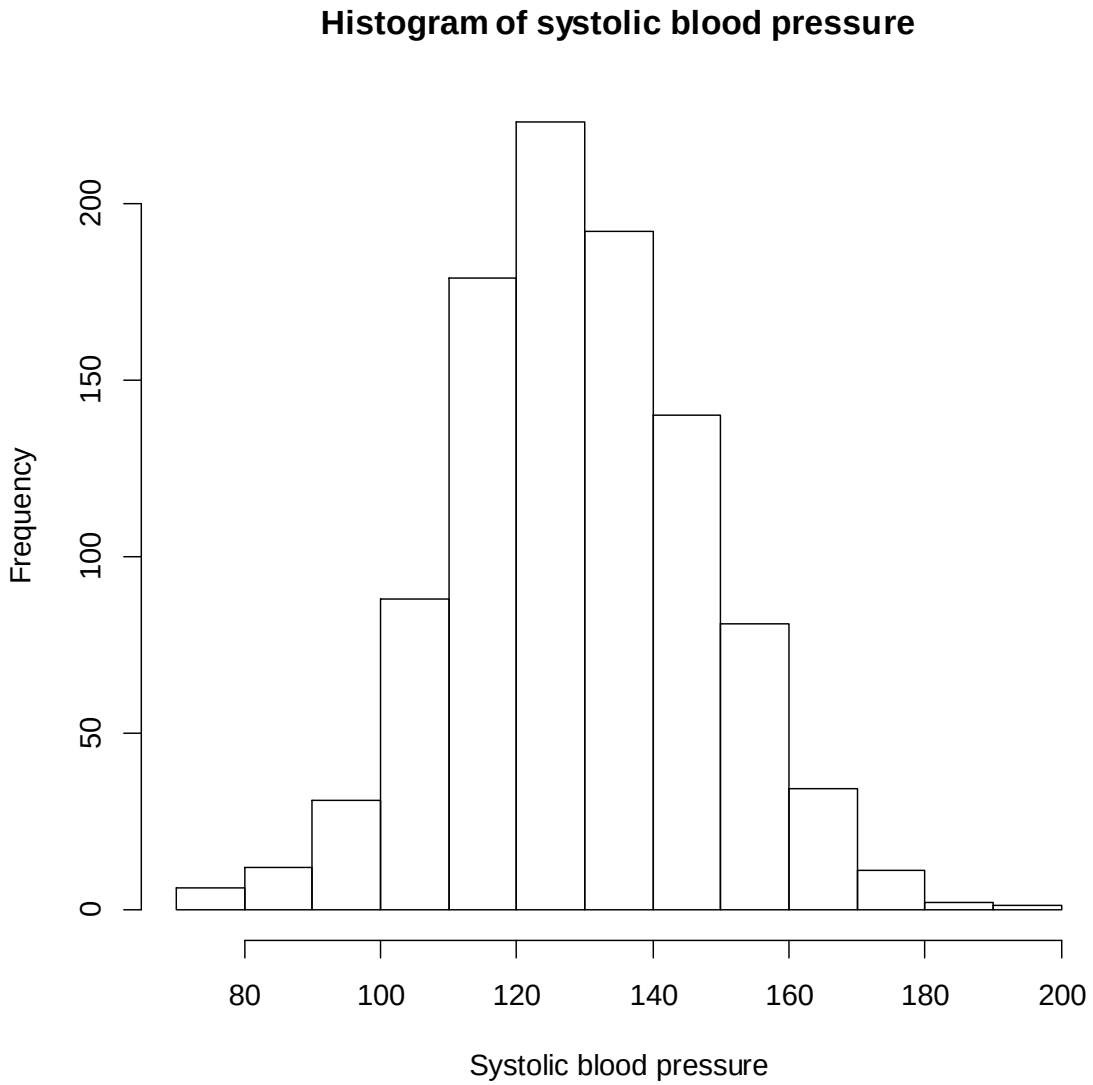
Box plot: Example



Histogram

- The horisontal axis shows possible intervals for the values of the variable.
- The vertical axis shows either the frequency or the relative frequency of observations within each interval.
- This is also known as a frequency distribution.

Histogram: Example



Sampling scheme effect

- Your sampling scheme will have an effect on the techniques used to calculate the mean and variance of your sample.
- The method used to calculate the mean and variance will influence all other statistical inference.
- All statistical techniques discussed in this course are based on a simple random sample.

Example: Stratified random sampling

Stratum 1	Stratum 2
100	25
95	32
94	31
91	27
89	21
92	
97	
88	
92	
96	

Stratum weights

$W_1 = 0.67$

$W_2 = 0.33$

Stratum mean

$\bar{X}_1 = 93.4$

$\bar{X}_2 = 27.2$

Estimate of the population mean

$$\begin{aligned} \bar{X} &= (W_1\bar{X}_1) + (W_2\bar{X}_2) \\ &= 71.55 \end{aligned}$$

Comparing groups of continuous data

Marike Cockeran

2015

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Outline

- Comparing the means of two independent groups
  - Independent t-test
- Comparing the means of two dependent groups
  - Dependent t-test
- Test questions
- Comparing the means of several independent groups
  - Analysis of Variance (ANOVA)
- Test questions

Independent t-test

- Assumptions of the independent t-test
- Null hypothesis and alternative hypothesis
- Test statistic
- Effect size: Cohen's d-value

Independent t-test: Objectives

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

After completion of this study section you should be able to:

- Identify research questions for which the independent t-test is the appropriate analysis.
- Write down the null hypothesis and the alternative hypothesis.
- Interpret the p-value of an independent t-test.
- Interpret Cohen's d-value of an independent t-test.

Independent t-test: Introduction

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

- The independent samples t-test compares the means between two unrelated groups on the same continuous, dependent variable.

- For example, you could use an independent t-test to compare blood pressure levels between smokers and non-smokers.

Independent t-test: Assumptions

- The dependent variable is measured on a continuous scale.
- The independent variable should consist of two categorical, independent groups.
- There are no significant outliers.
- The dependent variable is approximately normally distributed for each group of the independent variable.
- The variances in each group of the independent variable is roughly equal. This is called homogeneity of variance.

Independent t-test: Hypothesis

- Null hypothesis:

$$H_0: \mu_1 = \mu_2$$

- Alternative hypothesis:

$$H_A: \mu_1 \neq \mu_2$$

Independent t-test: Test statistic

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Independent t-test: Effect size

- A measure of practical significance (effect size) is Cohen's d-value:

$$d = \frac{|\bar{X}_1 - \bar{X}_2|}{\max(s_1, s_2)}$$

- Guideline values for interpreting Cohen's d-value:

$$|d| \approx 0.2 \; small \; effect$$

$$|d| \approx 0.5 \; medium \; effect$$

$$|d| \approx 0.8 \; large \; effect$$

Dependent t-test

- Assumptions of the dependent t-test
- Null hypothesis and alternative hypothesis
- Test statistic
- Effect size: Cohen's d-value

Dependent t-test: Objectives

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

After completion of this study section you should be able to:

- Identify research questions for which the dependent t-test is the appropriate analysis.
- Write down the null hypothesis and the alternative hypothesis.
- Interpret the p-value of a dependent t-test.
- Interpret Cohen's d-value of a dependent t-test.

Dependent t-test: Introduction

- The dependent t-test (paired t-test) compares the means between two related groups on the same continuous dependent variable.

- Measurements are taken on a single subject at two distinct points in time.

- The researcher matches the subjects in one group with those in a second group so that the members of a pair are as much alike as possible with respect to important characteristics such as age and gender.

Dependent t-test: Assumptions

- The dependent variable is measured on a continuous scale.
- The independent variable consists of two categorical related groups or matched pairs.
- There are no significant outliers in the differences between the two related groups.
- The distribution of the differences in the dependent variable between the two related groups is approximately normally distributed.

Dependent t-test: Hypothesis

- Denote the difference in population means by:

$$\delta = \mu_1 - \mu_2$$

- Null hypothesis:

$$H_0 : \delta = 0$$

- Alternative hypothesis:

$$H_A : \delta \neq 0$$

Dependent t-test: Test statistic

$$t = \frac{\bar{d} - \delta}{s_d / \sqrt{n}}$$

- Where $\bar{d}$ and s_d is defined as

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}}$$

Dependent t-test: Effect size

- A measure of practical significance (effect size) is Cohen's d-value:

$$d = \frac{|\bar{X}_1 - \bar{X}_2|}{s_1}$$

- Guideline values for interpreting Cohen's d-value:

$$|d| \approx 0.2 \; small \; effect$$

$$|d| \approx 0.5 \; medium \; effect$$

$$|d| \approx 0.8 \; large \; effect$$

Analysis of variance

- Assumptions of analysis of variance
- Null hypothesis and alternative hypothesis
- Bonferroni multiple comparison procedure
- Additional remarks

Analysis of variance: Introduction

- In the previous section we looked at techniques for determining whether a difference exists between the means of two independent populations.

- In some situations we would like to test for differences among three or more independent means rather than just two.

- The extension of the independent two-sample t-test to three or more samples is known as the analysis of variance.

ANOVA: Objectives

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

After completion of this study section you should be able to:

- Interpret the results of a one-way analysis of variance to test hypotheses concerning equality of population means.
- Explain why post hoc tests are necessary after analysis of variance has been done.
- Interpret the results of a post hoc test.

ANOVA: Assumptions

- The dependent variable is measured on a continuous scale.
- The independent variable consists of three or more, categorical, independent groups.
- There are no significant outliers.
- The dependent variable is approximately normally distributed for each group of the independent variable.
- The variances in each group of the independent variable is roughly equal. This is called homogeneity of variance.

ANOVA: Hypothesis

- Null hypothesis

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

- Alternative hypothesis
  - At least one of the population means differs from one of the others.

- The One-way ANOVA is an omnibus test and cannot tell you which specific groups were significantly different from each other.

- The omnibus test is referred to as the F-test.

- To determine which groups differ from each other you need to use post hoc tests.

Bonferroni multiple comparison test

- $H_0: \mu_1 = \mu_2 = \mu_3$
- F-test: A p-value of 0.03 is obtained. We can reject the null hypothesis. At least one of the population means differs from one of the others.
- $H_0: \mu_1 = \mu_2$ and $H_0: \mu_1 = \mu_3$ and $H_0: \mu_2 = \mu_3$
- You need to adjust the significance level used for each test to have an overall significance level of $\alpha = 0.05$.
- Bonferroni adjustment: $\frac{\alpha}{\#\ tests} = \frac{0.05}{3} = 0.0167$

Additional remarks

- Two-way ANOVA
  - Compare mean differences between groups that have been split on two independent variables (gender and smoking status).

- Analysis of covariance (ANCOVA)
  - You design involves a third variable (covariate) that you want to statistically control.

- Repeated measures ANOVA
  - Equivalent of One-way ANOVA, but used for related groups.

Comparing groups of categorical data

Marike Cockeran

2015

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Outline

- Pearson's chi-square test (independent groups)
- McNemar's test (dependent groups)
- Odds ratio
- Test questions

Objectives

After completion of this study section, you should be able to:

- State the null hypothesis and alternative hypothesis for Pearson's chi-square test.
- Interpret the output of Pearson's chi-square test.
- State the null hypothesis and alternative hypothesis for McNemar's test.
- Interpret the output of McNemar's test.
- Explain the difference between Pearson's chi-square test and McNemar's test.
- Interpret the odds ratio value.

Contingency tables

- When working with data that have been grouped into categories, we often arrange the counts in a tabular format known as a contingency table.

- The rows of the table represent the outcomes of one variable, and the columns represent the outcomes of the second variable.

- The entries in the table are the counts that correspond to a particular combination of categories.

	Variable 2		
Variable 1	Category 1	Category 2	Total
Category 1	a	b	a+b
Category 2	c	d	c +d
Total	a+c	b+d	n

Chi-square test: Example

- A study is conducted to determine the effectiveness of bicycle safety helmets in preventing head injury.

- The data consist of a random sample of 793 individuals who were involved in bicycle accidents during a specified one year period.

	Wearing helmet		
Head injury	Yes	No	Total
Yes	17	218	235
No	130	428	558
Total	147	646	793

Chi-square test: Example

- To examine the effectiveness of bicycle safety helmets, the researcher wants to determine whether there is an association between incidence of head injury and the use of helmets among individuals who have been involved in accidents.

- Null hypothesis: The proportion of persons suffering head injuries when wearing safety helmets at the time of the accident is equal to the proportion of persons sustaining head injuries among those not wearing helmets. In other words, there is no association between incidence of head injury and the use of helmets.

- Alternative hypothesis: The proportions of persons suffering head injuries are not identical in the two populations.

Chi-square test: Example

- Under the null hypothesis, the proportion of individuals experiencing head injuries among those wearing helmets and those not wearing helmets are equal.

- Therefore, one can ignore the two separate categories and treat all 793 individuals as a single homogeneous sample.

- The proportion of persons sustaining head injuries is:

$$\frac{235}{793} = 29.6\%$$

- The proportion of persons not sustaining head injuries is:

$$\frac{558}{793} = 70.4\%$$

Chi-square test: Example

- We would expect that of the 147 individuals wearing safety helmets 29.6% would suffer head injuries.

$$147 \times 29.6\% = 43.6$$

- We would expect that of the 147 individuals wearing safety helmets 70.4% will not suffer head injuries.

$$147 \times 70.4\% = 103.4$$

- We would expect that of the 646 individuals not wearing safety helmets 29.6% would suffer head injuries.

$$646 \times 29.6\% = 191.4$$

- We would expect that of the 646 individuals not wearing safety helmets 70.4% will not suffer head injuries.

$$646 \times 70.4\% = 454.6$$

Chi-square test: Example

- Observed frequencies:

Head injury	Wearing helmet		Total
	Yes	No	
Yes	17	218	235
No	130	428	558
Total	147	646	793

- Expected frequencies under the null hypothesis:

Head injury	Wearing helmet		Total
	Yes	No	
Yes	43.6	191.4	235.0
No	103.4	454.6	558.0
Total	147.0	646.0	793.0

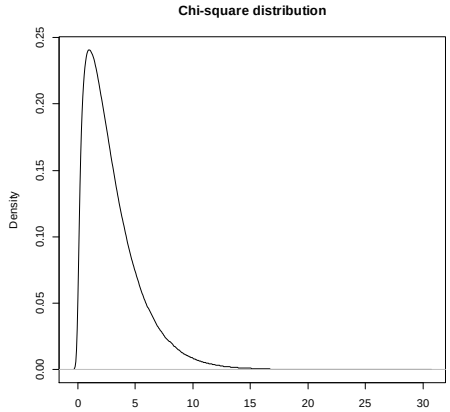
Pearson's chi-square test

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

- Pearson's Chi-square test is used to determine if there is a relationship between the two categorical variables.

- The chi-square test compares the observed frequencies in each category of the contingency table with the expected frequencies given that the null hypothesis is true.

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$



Assumptions of the chi-square test

- The two variables should be measured at a nominal or ordinal level (categorical data).
- Your two variables should consist of two or more categorical independent groups.
- No cell in the table should have an expected count less than 1.
- No more than 20% of the cells should have an expected count less than 5.

Additional remarks

- The Pearson chi-square test can be generalised to accommodate the comparison of three or more proportions.

- More complex contingency tables in which there are three or more variables cannot be analysed with the Pearson chi-square test and instead has to be analysed with a technique called loglinear analysis.

McNemar's test

- The McNemar test is used to determine if there are differences on a dichotomous dependent variable between two related groups.

- You need to have one categorical dependent variable with two categories and one categorical independent variable with two related groups.

- The groups of your dependent and independent variable must be mutually exclusive.

McNemar's test: Example

In 1980, a sample of 2110 adults over the age of 18 were asked to identify themselves as smokers or non-smokers. In 1982, the same 2110 individuals were again asked whether they were currently smokers or nonsmokers.

	1982		
1980	Smoker	Non-smoker	Total
Smoker	620	97	717
Non-smoker	76	1317	1393
Total	696	1414	2110

McNemar's test: Example

- Null hypothesis: Equal numbers switched from being smokers tot non-smokers and from being nonsmokers to smokers.

- Alternative hypothesis: There is an association, or there is a tendency for smoking status to change in one direction or the other.

- We have r=97 pairs in which smokers becomes nonsmokers and s=76 pairs in which a non-smoker becomes a smoker.

$$\chi^2 = \frac{|r-s|^2}{(r+s)}$$

Odds ratio

- If an event occurs with probability p, the odds in favour of the event are $p/(1-p)$ to 1.

- If we have two dichotomous random variables that represent a disease and an exposure, the odds ratio is defined as the odds in favour of disease among exposed individuals divided by the odds in favour of disease among the unexposed.

$$OR = \frac{P(disease|exposed)/[1 - P(disease|exposed)]}{P(disease|unexposed)/[1 - P(disease|unexposed)}$$

- OR=1 Exposure does not affect odds of outcome

- OR>1 Exposure associated with higher odds of outcome

Odds ratio

	Exposed	Unexposed	Total
Disease	a	b	$a+b$
No disease	c	d	$c+d$
Total	$a+c$	$b+d$	n

$$P(disease|exposed) = \frac{P(disease \cap exposed)}{P(exposed)} = \frac{a}{a+c}$$

$$P(disease|unexposed) = \frac{P(disease \cap unexposed)}{P(unexposed)} = \frac{b}{b+d}$$

Odds ratio

	Exposed	Unexposed	Total
Disease	a	b	$a+b$
No disease	c	d	$c+d$
Total	$a+c$	$b+d$	n

$$1 - P(disease|exposed) = 1 - \frac{a}{a+c} = \frac{c}{a+c}$$

$$1 - P(disease|unexposed) = 1 - \frac{b}{b+d} = \frac{d}{b+d}$$

$$OR = \frac{P(disease|exposed)/[1-P(disease|exposed)]}{P(disease|unexposed)/[1-P(disease|unexposed)} = \frac{\left[\frac{a}{a+c}\right]/\left[\frac{c}{a+c}\right]}{\left[\frac{b}{b+d}\right]/\left[\frac{d}{b+d}\right]} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Odds ratio: Example

- A study attempts to determine whether the use of electronic fetal monitoring during labor affects the frequency of caesarean section deliveries.

	EFM exposure		
Caesarean delivery	Yes	No	Total
Yes	358	229	587
No	2492	2745	5237
Total	2850	2974	5824

- (358)(2745)/(229)(2492)=1.72
- The odds of being delivered by caesarean section are 1.72 times higher for fetuses that are electronically monitored during labor than for fetuses that are not monitored.
- This does not imply that electronic monitoring causes a caesarean delivery. It is possible that the fetuses at higher risk are the ones that are monitored.

Correlation analysis

Marike Cockeran

2015

Outline

- Two-way scatter plot
- Pearson's correlation coefficient
- Spearman's rank correlation coefficient

Introduction

- The aim of this section is to investigate the relationships that can exist among continuous variables.

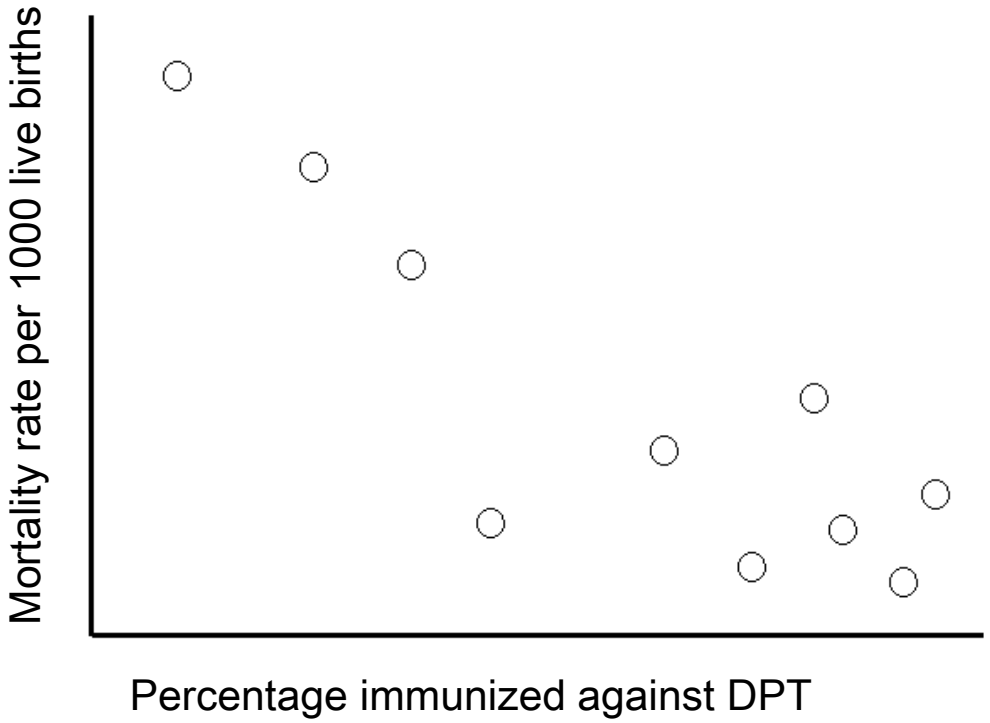
- One statistical technique often employed to measure such a relationship is known as correlation analysis.

- Correlation is defined as the quantification of the degree to which two random variable are related, provided that the relationship is linear.

Two-way scatter plot

- Place the outcomes of the X variable along the horizontal axis and the outcomes of the Y variable along the vertical axis.

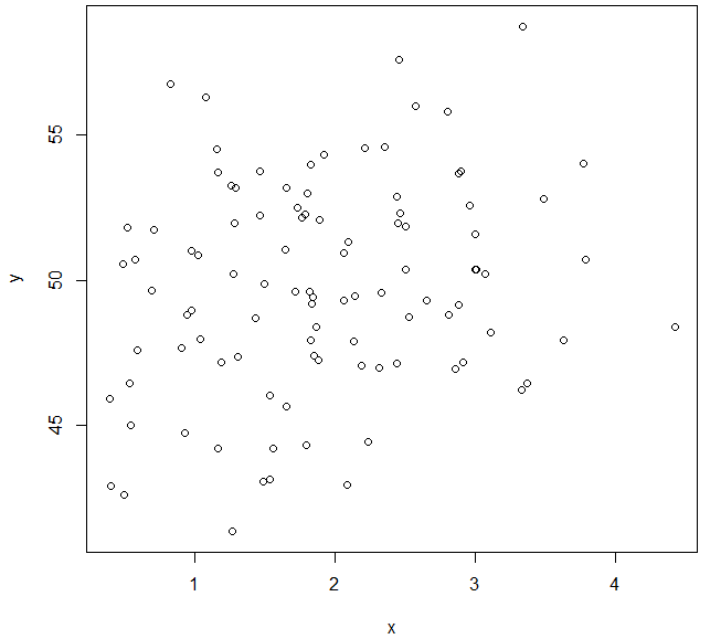
- Each point on the graph represents a combination of values (X_i, Y_i).



Correlation analysis

- Correlation is defined as the quantification of the degree to which two random variables are related, provided that the relationship is linear.

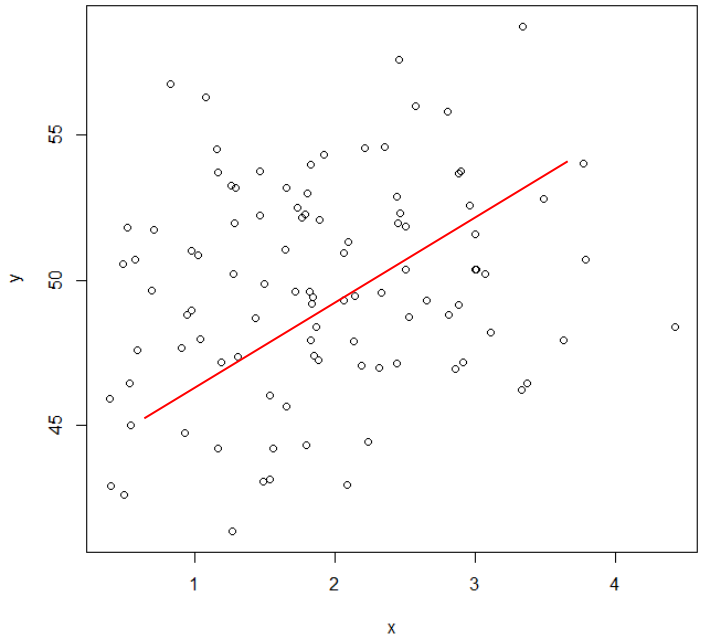
- Graphic presentation of the correlation between two variables is called a two-way scatter plot.



Correlation analysis

- Correlation is defined as the quantification of the degree to which two random variables are related, provided that the relationship is linear.

- Graphic presentation of the correlation between two variables is called a two-way scatter plot.



Pearson's correlation coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

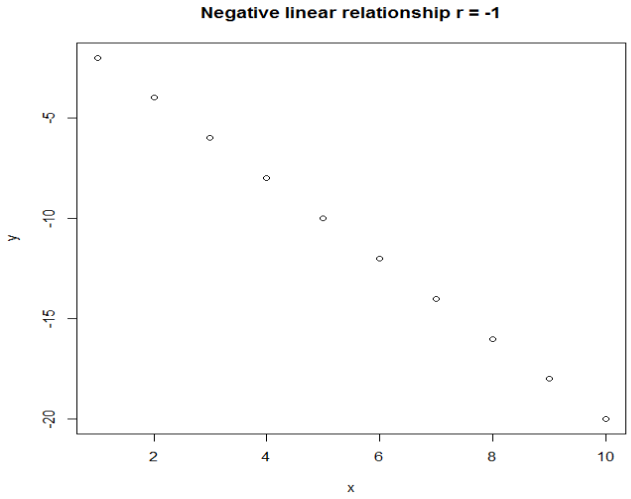
- $\bar{X}$ and $\bar{Y}$ are the sample means of the X and Y values.
- s_X and s_Y are the sample standard deviations of the the X and Y values.

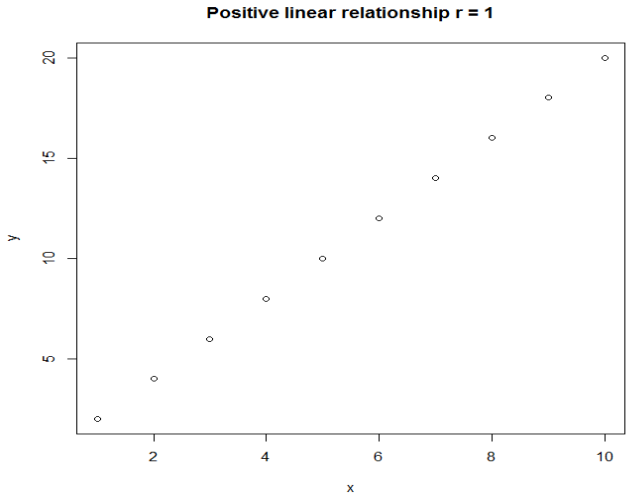
Pearson correlation coefficient

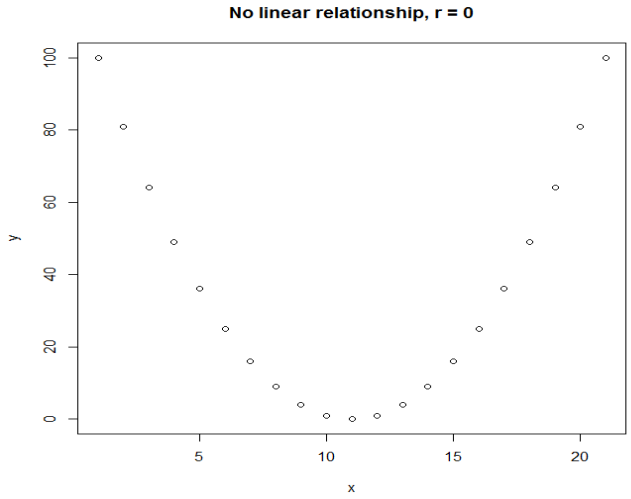
NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

- The parametric estimator of the correlation between two variables is known as Pearson's correlation coefficient (r).
- The maximum value of r is 1; the minimum value of r is -1.
- If $r = 1$ or $r = -1$ then an exact linear relationship exists between the two variables.
- If $r = 0$ there is no linear relationship between the two variables and the variables are uncorrelated.
- If the values of the first variable increase as the values of the second variable increase, then the two variables are positively correlated.
- If the values of the first variable decrease as the values of the

Pearson's correlation coefficient







Pearson's correlation: Hypothesis testing

- Null hypothesis:

$$H_0: \rho = 0$$

- Alternative hypothesis:

$$H_0: \rho \neq 0$$

Correlation coefficient as an effect size

- The effect size can be used to measure the strength of the relationship between the two continuous variables.

- $|r| = 0.1$ small effect

- $|r| = 0.3$ medium effect

- $|r| = 0.5$ large effect

Spearman's rank correlation coefficient

- Pearson's correlation coefficient is very sensitive to outlying values.
- We may be interested in calculating a measure of association that is more robust.
- One approach is to rank the two sets of outcomes X and Y separately and calculate a coefficient of rank correlation.
- This procedure is known as Spearman's rank correlation coefficient.

$$r_s = \frac{\sum_{i=1}^{n}(X_{ri} - \bar{X}_r)(Y_{ri} - \bar{Y}_r)}{\sqrt{\left[\sum_{i=1}^{n}(X_{ri} - \bar{X}_r)^2\right]\left[\sum_{i=1}^{n}(Y_{ri} - \bar{Y}_r)^2\right]}}$$

The effect of outliers

- Pearson's correlation coefficient = 0.43
- Spearman's correlation coefficient = 0.56