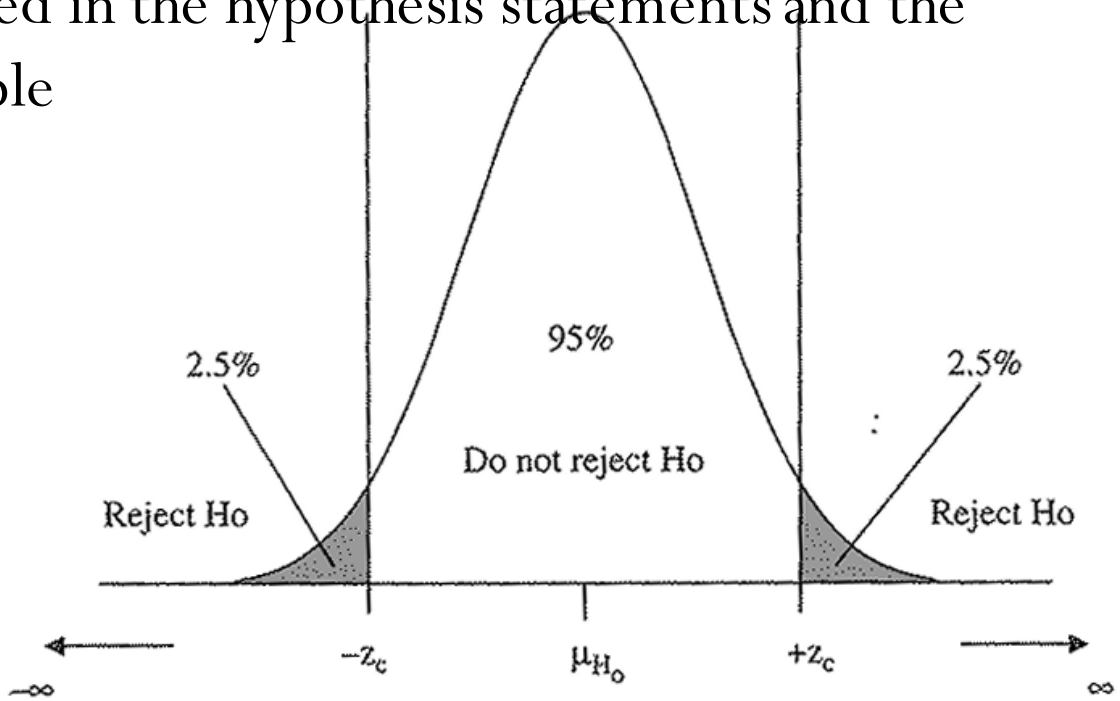
Data Exploration – Statistics 3

Dr. Markus Hofmann

Revision

Hypothesis Assessment

- First we calculate the statistic of interest from the sample
- The hypothesis test will then look at the difference between the value claimed in the hypothesis statements and the calculated sample

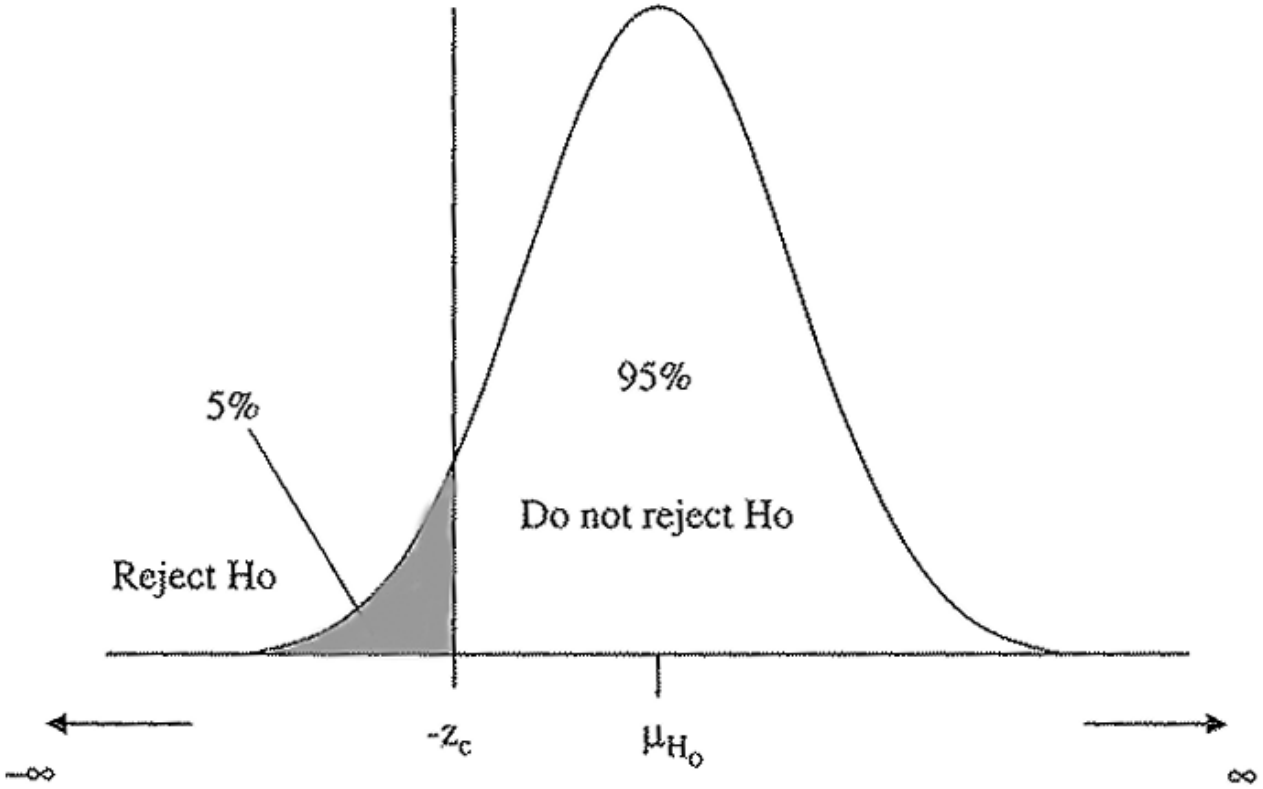


Revision

Hypothesis Assessment

- The graph on the previous slide indicated a two tailed test meaning that we reject if the value is less than or greater than.
- If the alternative hypothesis states only one of these conditions (e.g. Ha: $\mu < \mu_{H_0}$ and $\alpha = 0.05$) then we would reject the null hypothesis if the hypothesis test result has a z-score to the left of the critical value of z.
  - Since this is a one-tailed test the single area shaded should be equal to 5%

Revision



Revision - p-values

- A hypothesis test is usually converted into a p-value.
- The p-value is the probability of getting the recorded value or a more extreme value
- It is measured of the likelihood of the result given the null hypothesis is true or the statistical significance of the claim
- Where the alternative hypothesis is not equal then the area under the curve value is doubled
- p-values range from 0 to 1
- Where the p-value is less than α the null hypothesis is rejected. If the p-value is greater than α the null hypothesis is not rejected

Hypothesis Test: Single Group, Continuous Data

- H_0: $\mu = 12$
- Ha: $\mu \neq 12$
- Where μ is the claimed average of days to process a passport
- n= 45, x bar = 12.1 and standard deviation = 0.23
- α was set to 0.05
- To calculate the hypothesis test the following formula can be used:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{12.1 - 12}{\frac{0.23}{\sqrt{45}}} = 2.9$$

Hypothesis Test: Single Group, Continuous Data

- For =0.05 the critical value of z would be 1.96. This is where the are under each extreme would be equal 2.5%.
- Since the z-score of 2.9 is greater than 1.96 we reject the null hypothesis and make a statement that the average number of days to process a passport is not 12 days.
- To calculate the p-value we look-up the calculated hypothesis score of 2.9 in the normal distribution table which holds a value of 0.0019. Since the hypothesis is two sided we have to double this value to obtain a p-value of 0.0038
- As the p value is less than 0.05 the null hypothesis needs to be rejected.

Hypothesis Test: Single Group, Categorical Data

- Let's test the claim that more than eight of ten data miners prefer a certain Data Mining tool (tool x):
  - H0: $\pi = 0.8$
  - Ha: $\pi > 0.8$
  - Where π is the claimed proportion of data miners preferring tool x
  - 40 random data miners were questioned of which 33 preferred tool x ($p = 0.825$). α was set to 0.05.
  - We can calculate the hypothesis test by taken the difference between the value stated in the null hypothesis and the recorded sample divided by the standard error of proportions.

Hypothesis Test: Single Group, Categorical Data

$$z = \frac{\rho - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.825 - 0.8}{\sqrt{\frac{0.8(1-0.8)}{40}}} = 0.395$$

- The critical z-score when $\alpha = 0.05$ is 1.65
  - Note: this is a one-tailed test therefore the shaded area is 5% on the right side of the distribution graph
- The p-value is 0.3446
- Since the p-value is greater than α, we do not reject the hypothesis and therefore cannot make the claim that more than 80% of data miners prefer tool x

Hypothesis Test: Two Groups, Continuous Data

- Let's look at the following scenario:
  - The average fuel efficiency for 4-cylinder vehicles is greater than the average fuel efficiency for 6-cylinder vehicles
    - H_0: $\mu_1 = \mu_2$
    - H_a: $\mu_1 > \mu_2$
  - where μ_1 is the average fuel efficiency for 4-cylinder vehicles and μ_2 is the average fuel efficiency for 6-cylinder vehicles
  - α=0.05
  - Two groups of cars were randomly selected. The following results were collected:

Group	n	X bar (mpg)	Variance
4 cylinder	24	25.85	50.43
6 cylinder	27	23.15	48.71

Hypothesis Test: Two Groups, Continuous Data

- Since the group sizes are less than 30 we will work with the Student's t distribution
- The following formulae are used:
  - Where s_p is the pooled standard deviation and can be calculated using the pooled variance

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  - $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Hypothesis Test: Two Groups, Continuous Data

- Pooled Variance

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{(24-1)50.43+(27-1)48.71}{(24-1)+(27-1)} = 49.52$$

- As the null hypothesis states that $\mu_1 = \mu_2$ can assume that $\mu_1 - \mu_2 = 0$

$$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(25.85-23.15)-(0)}{\sqrt{49.52}\sqrt{\frac{1}{24}+\frac{1}{27}}} = 1.37$$

Group	n	X bar (mpg)	Variance
4 cylinder	24	25.85	50.43
6 cylinder	27	23.15	48.71

Hypothesis Test: Two Groups, Continuous Data

- Degrees of freedom:

$$df = (n_1 + n_2 - 2) = 24 + 27 - 2 = 49$$

- The critical z score is approx. 1.68 using a t-distribution table and the p value just less than 0.1. We therefore cannot reject the null hypothesis
- If the examples for each group are greater than 30 we can use:

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

Hypothesis Test: Two Groups, Categorical Data

- Claim: A new drug reduces the number of strokes
  - H_0: $\pi_1 = \pi_2$
  - H_a: $\pi_1 < \pi_2$
  - Where : π_1 is the proportion of patients who take the new medicine and : π_2 is the proportion of patients who take a placebo
- Contingency table outlining the results:

	Takes Medicine	Takes Placebo	Total
Has Strokes	213	342	555
No Strokes	9,791	9,671	19.462
Totals	10,004	10,013	20,017

Hypothesis Test: Two Groups, Categorical Data

- We first calculate the total proportions of patients who had a stroke:

$$\rho = \frac{X_1 + X_2}{n_1 + n_2} = \frac{213 + 342}{10004 + 10013} = 0.0277$$

- Then we calculate the proportions for each group:

$$\rho_1 = \frac{X_1}{n_1} = \frac{213}{10004} = 0.0213 \qquad \rho_2 = \frac{X_2}{n_2} = \frac{342}{10013} = 0.0342$$

Hypothesis Test: Two Groups, Categorical Data

- The null hypothesis states that there is no difference in proportions of strokes. This now needs to be checked statistically
- We can use the following equation:

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$z = \frac{(0.0213 - 0.0342) - 0}{\sqrt{0.0277(1-0.0277)(\frac{1}{10004} + \frac{1}{10013})}} = -5.54$$

Hypothesis Test: Two Groups, Categorical Data

- Checking the z-score for 5.54 in the normal distribution table we extract a value of virtually zero and can therefore reject the hypothesis with confidence.
- We can therefore conclude (with statistical significance) that the medicine works

Paired Test

- The paired test can be used to investigate whether the sample mean comes a different population than the one specified?
- Two possible reasons for the difference:
  - The sample comes from a different population (reject Null Hypothesis)
  - The mean varies by chance (fail to reject the Null Hypothesis)
- The statistic you need to decide whether to reject or fail to reject the Null Hypothesis is the one-sample t test.
- t-score is like a z-score and tells you if the sample mean is far away from the population mean.

Paired Test

- Let's use a widely quoted example:
  - There is no difference in the wear of shoes made from material X compared to shoes made out of material Y. We can therefore define the Hypotheses to be:
    - H_0: $\mu_D = 0$
    - H_a: $\mu_D \neq 0$
  - Where μ_D is the difference between the shoes made out of material X and Y
  - 10 boys wore a shoe made out of material X and one shoe made out of material Y
  - The amount of wear of each shoe was then recorded
  - This study requires a 90% confidence level

Paired Test

- The average difference is 0.41 (D bar) and the standard deviation was calculated to be 0.386 (S_D).
- We will use a t-distribution as the number of examples is small
- Therefore:

$$t = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} = \frac{0.41 - 0}{\frac{0.386}{\sqrt{10}}} = 3.36$$

- The degrees of freedom are calculated is (n-1)=(10-1)=9
- p value (=0.0082) is less than 0.1 and we therefore reject the null hypothesis and conclude that there is a difference.

Errors

- Since hypothesis tests are based on samples and samples vary there exists the possibility of errors. There are two potential errors:
- Type I error:
  - The null hypothesis is rejected when it really should not be. Setting the value of alpha low minimises these errors
- Type II error:
  - The null hypothesis is not rejected when it should have been. We can minimise this type of error by increasing the number of examples in a sample
- The real life example will identify which type of error should be avoided more and subsequent decisions will need to focus on the avoidance of this type of error.
- This is similar to the confusion matrix where we also have false positive and false negatives.
- **Power (1-beta)** is the probability of correctly rejecting a false Null Hypothesis

Exercise

- See exercises 1 to 7

Chi Square

- We can use a Chi Square test to set a hypothesis test to use with variables measured on a nominal or ordinal scale. It facilitates an investigation as to whether there is a relationship between two categorical variables.
- It is all about whether proportions differ or not.
- Generally the null hypothesis states 'there is no relationship' and the alternative hypothesis states 'there is a relationship'
- Example:

		Washing Powder Brand			
		Brand X	Brand Y	Brand Z	
Zipcode	43221	5521	4597	4642	**14760**
	43029	4522	4716	5047	**14285**
	43212	4424	5124	4784	**14332**
		14467	**14437**	**14473**	**43377**

Chi Square

- Chi Square test compares the observed frequencies with the expected frequencies.
- The expected frequencies are calculated using $E_{r,c} = \frac{r \times c}{n}$
- E.g. $E_{43221,Brand\_x} = \frac{r \times c}{n} = \frac{14760 \times 14467}{43377} = 4{,}923$
- Expected purchases:

		Washing Powder Brand			
		Brand X	BrandY	Brand Z	
Zipcode	43221	4923	4913	4925	**14760**
	43029	4764	4754	4766	**14285**
	43212	4780	4770	4782	**14332**
		14467	**14437**	**14473**	**43377**

Chi Square

- Having calculated all expected values we can calculate the chi square test using the following equation:

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

- Where k is the number of all categories O is the observed cell frequency and E is the estimated cell frequency.

Chi Square

- Calculation Table of chi square

k	Category		Observed (O)	Estimated (E)	((O-E)^2)/E
1	Brand X	43221	5521	4923	72.63945
2	Brand X	43029	4522	4764	12.29303
3	Brand X	43212	4424	4780	26.51381
4	Brand Y	43221	4597	4913	20.32485
5	Brand Y	43029	4716	4754	0.303744
6	Brand Y	43212	5124	4770	26.2717
7	Brand Z	43221	4642	4925	16.26173
8	Brand Z	43029	5047	4766	16.56756
9	Brand Z	43212	4784	4782	0.000836
				Sum=	**191.1767**

Chi Square

- Degrees of freedom

$$df = (r-1) \times (c-1) = (3-1)(3-1) = 4$$

- We can now look up the critical value in the chi square distribution table for df=4 and alpha = 0.05. The critical value is 9.488.
- Since the critical value of 9.488 is less than 191.2 we reject the null hypothesis and conclude that there is a relationship between zip codes and brands.
- Note: Chi square test only tells you whether a relationship exists or not but it does not tell you what type of relationship exists.

One-way Analysis of Variance

- Compares means from three or more different groups
- The test determines whether there is a difference between the groups or not
- Can be applied to cases where the groups are independent and random, the distributions are normal and the populations have similar variances.
- For example an online retailer has four call centres of similar size which process a certain amount of calls each day
- An analysis of the different call centres based on average number of calls each day is required.
- The null hypothesis states that the means are equal and the alternative hypothesis states that they are not equal

One-way Analysis of Variance

- The test checks whether there is a difference or not between the means or whether the difference is due to random variation.
- The following steps need to be applied:
  - Calculate group means and standard deviations
  - Determine the within group variation
  - Determine the between group variation
  - Determine the F-statistic
  - Test the significance of the F-statistic

One-way Analysis of Variance

- Calculate Group Means and Variances
  - Form the average of all group means (139.1+129.9+142.4+153.7)/4

	Call Center A	Call Center B	Call Center C	Call Center D	
	136	124	142	149	
	145	131	145	157	
	139	128	139	154	
	132	130	145	155	
	141	129	143	151	
	143	135	141	156	
	138	132	138		
	139		146		**TOTAL**
Count	8	7	8	6	**29**
Mean	**139.1**	**129.9**	**142.4**	**153.7**	141.3
Variance	16.4	11.8	8.6	9.5	

One-way Analysis of Variance

- Determine the Within Group Variation
  - This is defined by statistic within group variance or mean square within (MSW) and can be calculated using

$$MSW = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N - k}$$

  - Where N is the total number of observations, k is the number of groups and si^2 is the variance of group i

$$MSW = \frac{(8-1)\times 16.4 + (7-1)\times 11.8 + (8-1)\times 8.6 + (6-1)\times 9.5}{(29-4)} = 11.73$$

	A	B	C	D	Total
Count	8	7	8	6	**29**
Mean	139.1	129.9	142.4	153.7	141.3
Variance	16.4	11.8	8.6	9.5	

One-way Analysis of Variance

- Determine the Between Group Variation or also known as Mean Square Between (MSB)
- The MSB is the variance between the group means
- We can calculate it by using a weighted sum of the squared differences between the group mean (xi bar) and the average of the means (x bar bar). We then divide by the degrees of freedom (k-1) where k is the number of groups.
- Therefore

$$MSB = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

One-way Analysis of Variance

$$MSB = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

$$MSB = \frac{(8 \times (139.1 - 141.3)^2) + (7 \times (129.9 - 141.3)^2) + (8 \times (142.4 - 141.3)^2) + (6 \times (153.7 - 141.3)^2)}{4-1}$$

$$= 626.89$$

	A	B	C	D	Total
Count	8	7	8	6	**29**
Mean	139.1	129.9	142.4	153.7	141.3
Variance	16.4	11.8	8.6	9.5	

One-way Analysis of Variance

- Now we can calculate the F statistic

$$F = \frac{MSB}{MSW} \qquad F = \frac{626.89}{11.73} = 53.44$$

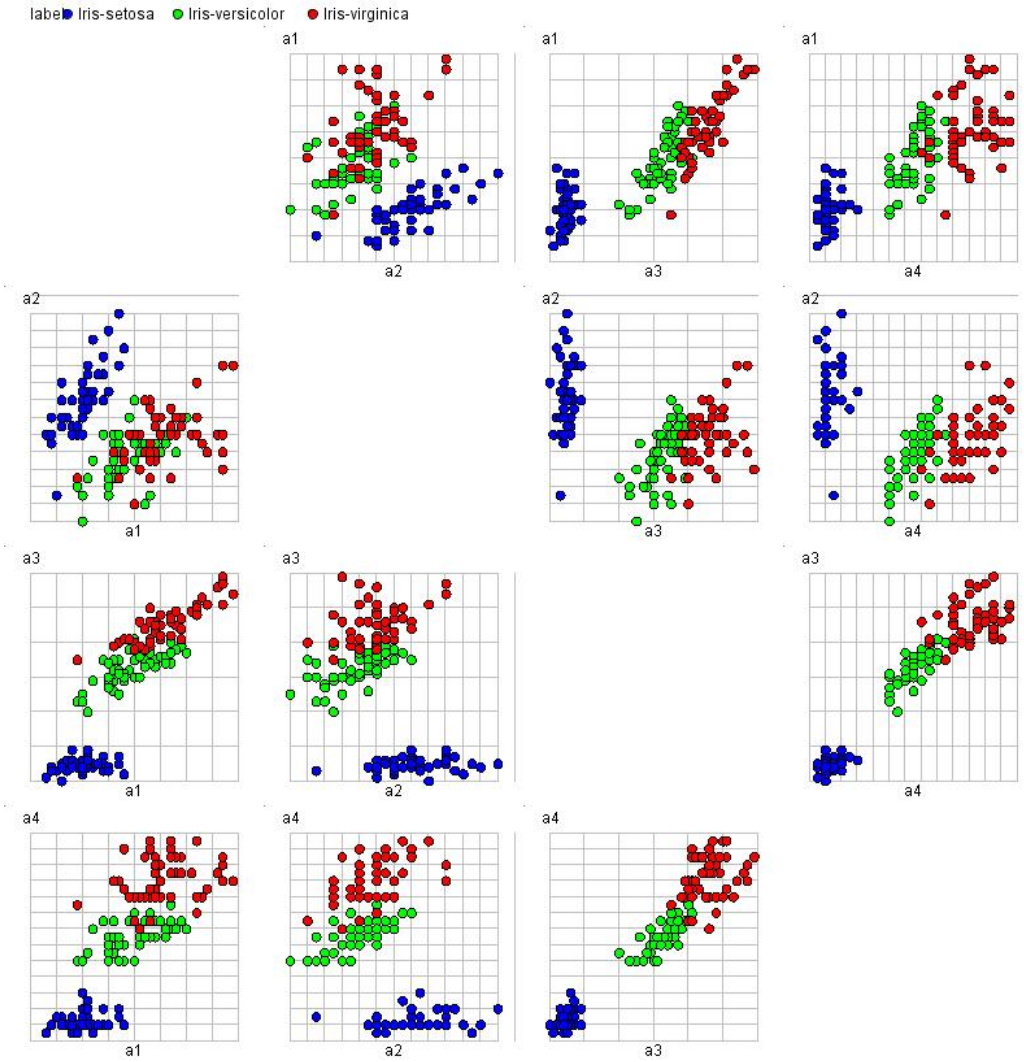
- Test the Significance of the F-Statistic
  - First we need to determine the degrees of freedom
    - DF within: $df_{within} = N - k = 29 - 4 = 25$
    - DF between: $df_{between} = k - 1 = 4 - 1 = 3$

Using the F-distribution table we can look up the critical z value using alpha of 0.05 and the degrees of freedom. The z-value we extract is 2.99 which is smaller than the calculated statistic and we therefore reject the null hypothesis. The means are not equal.

Comparative Statistics

- Relationships between variables can be complex but some characteristics can be measured:
  - Shape: a relationship is linear when it is drawn in a straight line. As values of one variable change the values of the second variable change proportionally.
  - Direction: Positive relationships exist if higher values in the first variable coincide with higher values in the second variable. In addition, lower values of the first variable coincide with lower values in the second variable. Negative relationships occur if there is a shift so that lower values of the first variable coincide with higher values of the second variable and vice versa. More complex scenarios exist also.

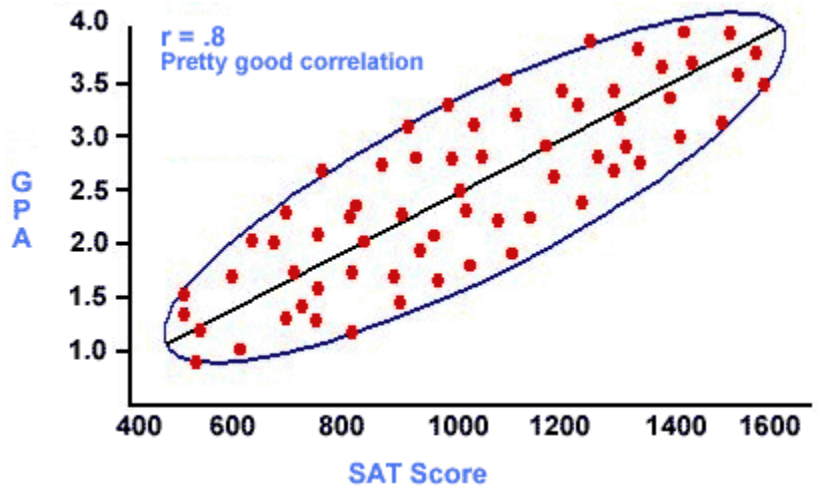
Visualising Relationships

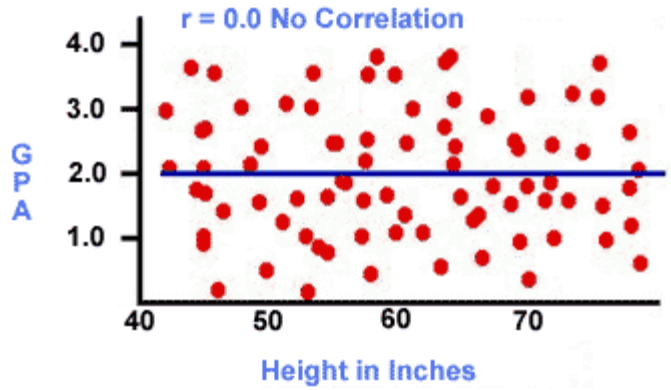


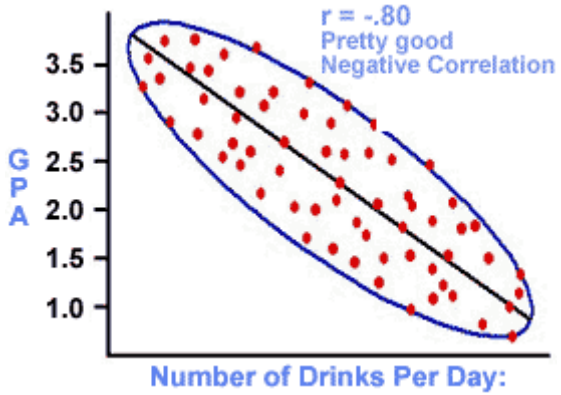
Correlation Coefficient (r)

- The correlation coefficient can be calculated for pairs of variables that are measured on the ration/interval scale.
- The value of the coefficient ranges from -1 to 1 and determines the level of correlation.
- If an optimal straight line is drawn through the points of a scatterplot then the value r reflects on how close to this line the points lie.
- The closer the value r is to zero the less likely it is that there is a linear relationship
- It has the ability to identify linear relationships (can one score predict another)

Correlation Coefficient (r)







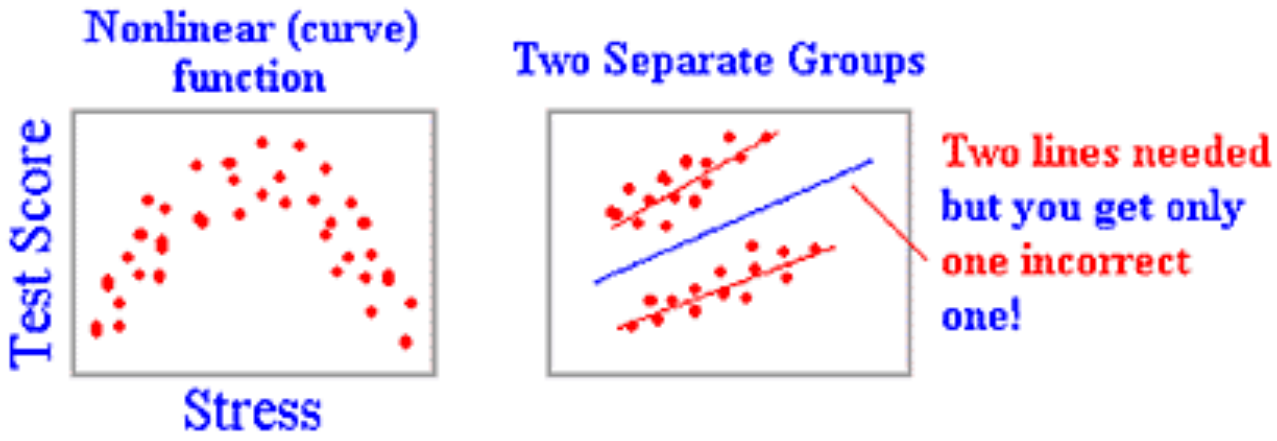
Source:
http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/correlation/correlation_06.html

Correlation Coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

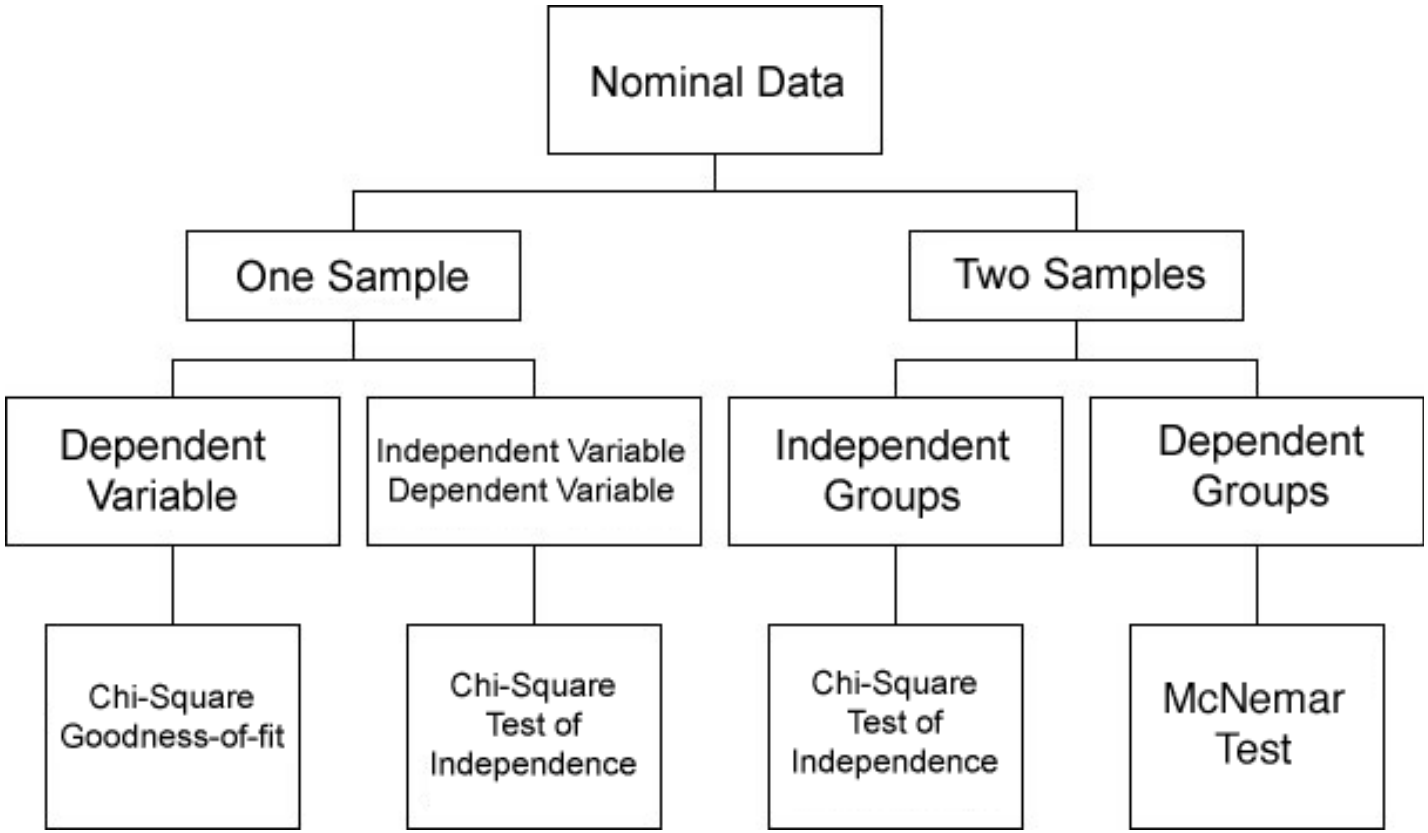
- where s_x is the standard deviation of variable x and s_y is the standard deviation of variable y

When does the correlation coefficient not work



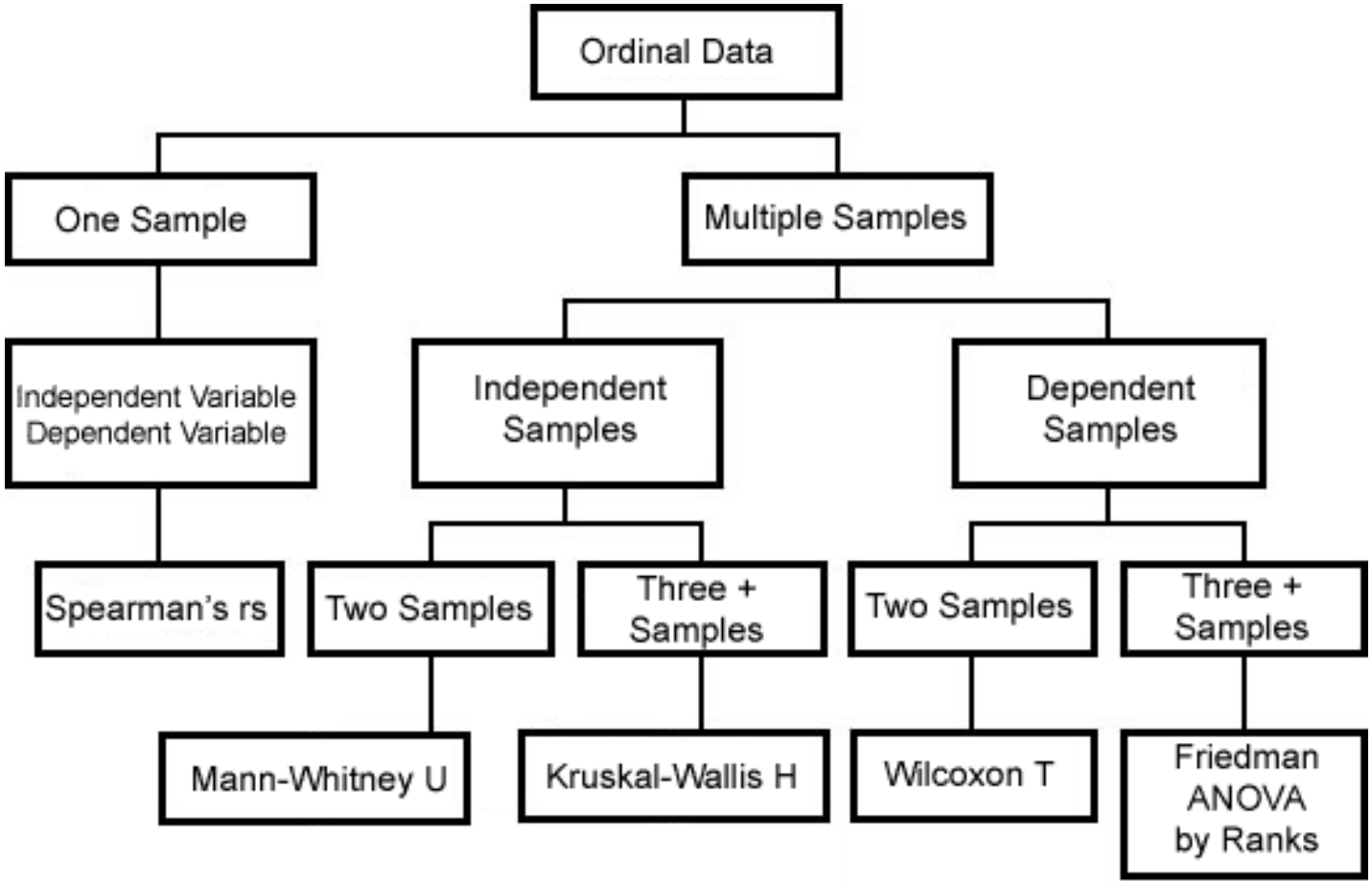
Source:
http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/correlation/correlation_18.html

Which Test for Nominal Data?



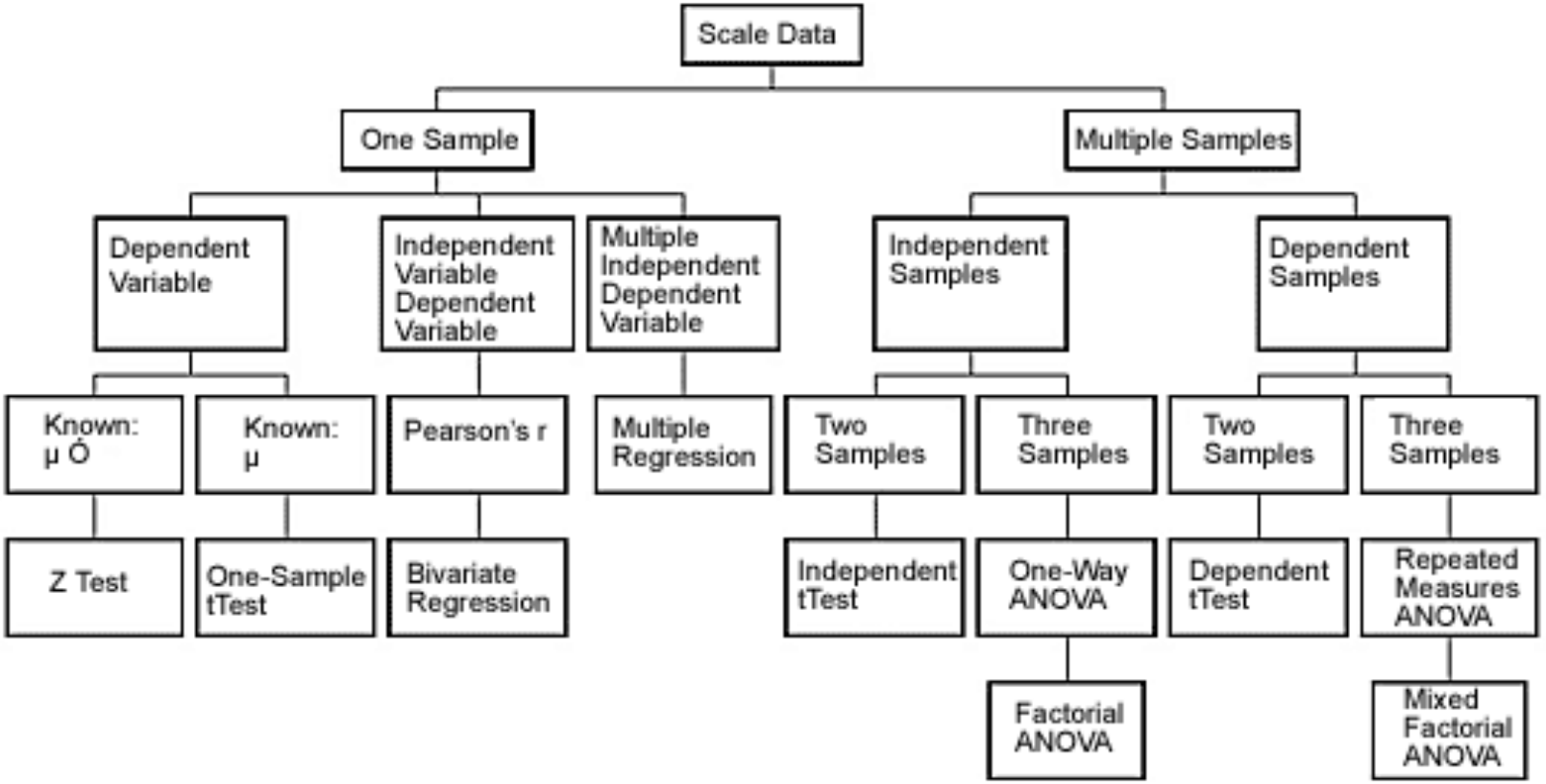
http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/chose_stat/chose_stat_23.html

Which Test for Ordinal Data?



http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/chose_stat/chose_stat_24.html

Which Test for Scale Data?



http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/chose_stat/chose_stat_25.html

Summary

	Continuous	Categorical	Number of groups	Number of Variables
Confidence Intervals	Yes	Yes	1	1
Hypothesis test	Yes	Yes	1 or 2	1
Chi Square	No	Yes	2+	2
One way analysis of variance	Yes	No	3+	1

Exercise

- See exercises 8 and 9