****

statstutor community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-marshallowen-6b

The following resources are associated:

Data\_for\_tutor\_training\_SPSS\_workbook.xls

Tutor\_training\_SPSS\_workbook.pdf

**SOLUTIONS**

**to**

**SPSS workbook**

**for New Statistics Tutors**

**(Based on SPSS Versions 21 and 22)**

**This workbook is aimed as a learning aid for new statistics tutors**

**in mathematics support centres**

# Contents

[Exercise 1: Data types 3](#_Toc406346117)

[Exercise 2: Chi-squared 4](#_Toc406346118)

[Exercise 3: Nationality and survival 5](#_Toc406346120)

[Exercise 4: Comparison of continuous data by group 6](#_Toc406346122)

[Exercise 5: Weight before the diet by gender 7](#_Toc406346123)

[Exercise 6: Confidence intervals 8](#_Toc406346124)

[Exercise 7:ANOVA 8](#_Toc406346125)

[Exercise 8: ANOVA output 9](#_Toc406346126)

[Exercise 9: Testing the assumptions of homogeneity and normality: 11](#_Toc406346127)

[Exercise 10: Two way ANOVA 12](#_Toc406346128)

[Exercise 11: Non-parametric tests 15](#_Toc406346129)

[Exercise 12: Kruskal-Wallis 15](#_Toc406346130)

[Exercise 13: Repeated measures example 17](#_Toc406346131)

[Exercise 14: Friedman example 18](#_Toc406346132)

[Exercise 15: Regression assumptions 19](#_Toc406346133)

[Exercise 16: Scatterplots 19](#_Toc406346134)

[Exercise 17: Correlations 19](#_Toc406346135)

[Exercise 18:Regression 20](#_Toc406346136)

[Exercise 19: Logistic regression 21](#_Toc406346137)

#### Exercise 1: Data types

Identify the type of variables and key questions of interest for the **Titanic** dataset

**Titanic** - Details for passengers travelling on the Titanic when it sank:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **class** | **Survived 0 = died** | **Country of residence** | **Gender** | **Age** | **No. of siblings/ spouses on board** | **No. of parents/ children on board** | **price of ticket** |
| Abbing, Anthony | 3 | 0 | USA | male | 42 | 0 | 0 | 7.55 |
| Abbott, Rosa | 3 | 1 | USA | female | 35 | 1 | 1 | 20.25 |
| Abelseth, Karen | 3 | 1 | UK | female | 16 | 0 | 0 | 7.65 |
| 1. **Type of variable** | Ordinal | Binary (nominal) | Nominal | Binary (nominal) | Continuous (Scale) | Discrete | Discrete | Cont/ Scale |

b) Key question(s)

Were wealthy people more likely to survive? Which variables would you use to investigate this question?

Survival is the outcome. Wealthy could be measured using either class or price of ticket.

**Quick question: What percentage of people survived the sinking of the Titanic? 38.2%**

**Quick question: What percentage of people survived the sinking of the Titanic in each class?**

**1st = 62%, 2nd = 43%, 3rd = 26%**

#### Exercise 2: Chi-squared

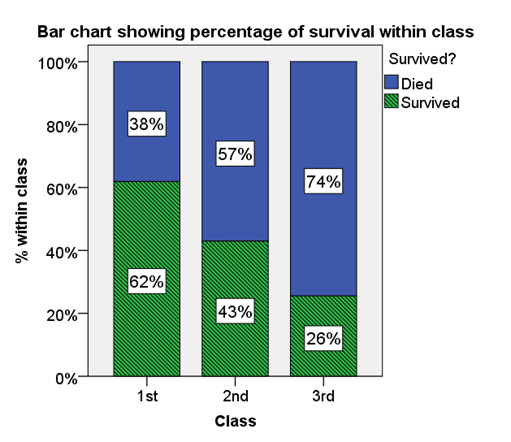
Investigate whether class and survival were related

1. Interpret the results of the Chi-squared test. If there is evidence of a relationship, what is the relationship?



As p < 0.001 for the Pearson’s Chi-squared, there is evidence to suggest that there is a relationship between class and survival.

If you find significant evidence of a relationship, produce some %’s and/ or a chart to show the relationship.



The percentage of those surviving decreases with 2nd and 3rd class. 62% of those in 1st class survived compared to only 38% of those in 3rd class.

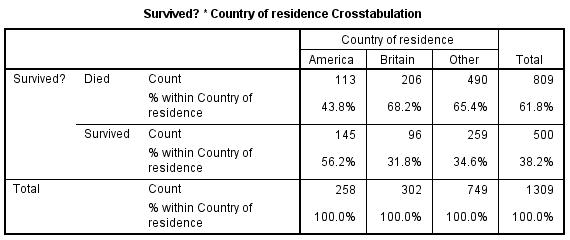
Data collected on 1309 passengers aboard the Titanic was used to investigate whether class had an effect on chances of survival. A chi-squared test gave a p-value of p < 0.001 so there is strong evidence to suggest a relationship between class and survival. *Figure 1* shows that as class decreases, the percentage of those surviving also decreases from 62% in 1st Class to 26% in 3rd Class. *Figure 1: Bar chart showing % survival within classes*

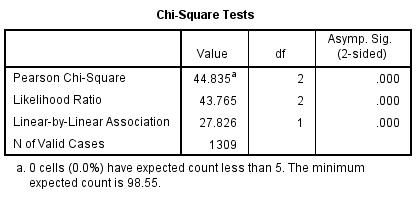
1. What would you do if you had a 2x2 contingency table?

Note: If each variable only has two categories, the contingency table is a 2 x 2 table. In this situation, use the line with ‘Continuity Correction’ in the main SPSS output or the Fishers Exact test to get the p-value for the test. They both automatically appear in the output for a 2x2 table.

Exercise 3: Nationality and survival

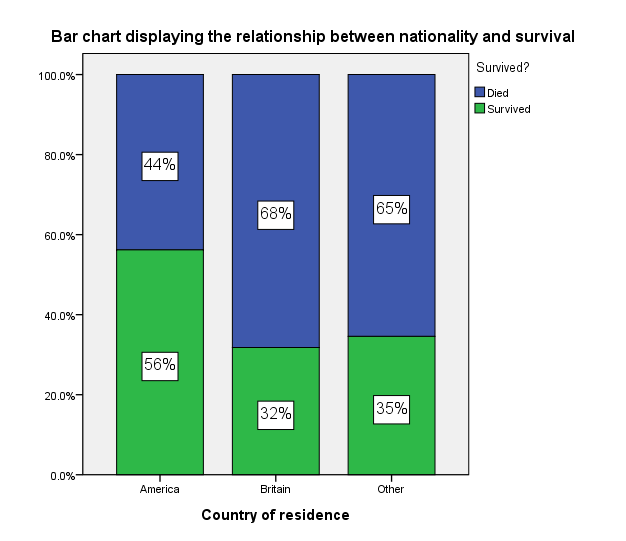
Investigate whether nationality and survival were related

******

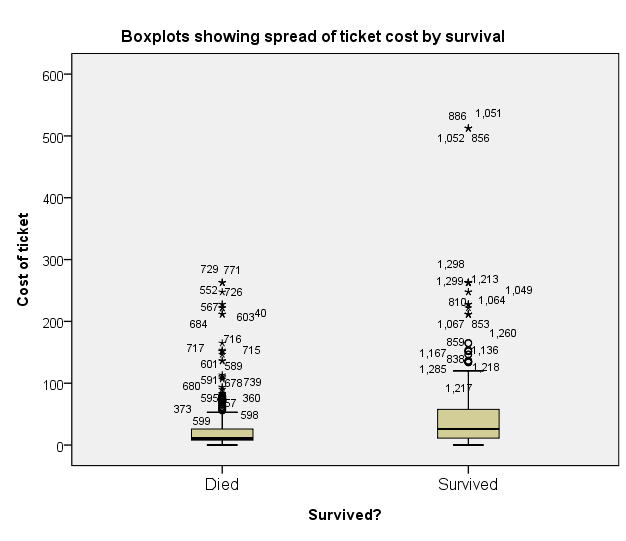


Test Statistic = 44.835, p-value < 0.001, reject null. i.e. there is significant evidence to suggest that there was a relationship between nationality and survival.

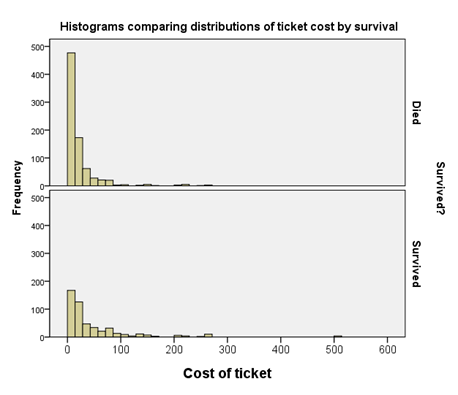
56% of Americans survived compared to 32% of British passengers and 32% of other nationalities.



#### Exercise 4: Comparison of continuous data by group

Did the cost of a ticket affect chances of survival?





1. Is there a big difference in average ticket price by group?

Yes. The mean and median ticket prices are much higher in the group who survived

1. Which group has data which is more spread out?

The standard deviation is double in the group who survived so there is much more variation in that group

1. Is the data skewed? Yes – it’s very positively skewed. There a lot of people with cheap tickets and not so many with expensive tickets. This can be seen clearly from the graphs but also by comparing the mean and the median. If there is a big difference, the data is skewed
2. Is the mean or median a better summary measure? The median as the data is very skewed.

If you are going to transform data, change the y axis in the graph to a log scale via the graph properties to see if taking the log of the variable will help.

#### Exercise 5: Weight before the diet by gender

1. **Fill in the following table using the summary statistics table in the output.**

|  |  |  |
| --- | --- | --- |
|  | Female = 0 | Male = 1 |
| Minimum | -70 | 71 |
| Maximum | 82 | 88 |
| Mean | 64 | 79 |
| Median | 66 | 79 |
| Standard Deviation | 21.6 | 5 |

1. **Interpret the summary statistics by gender. Which group has the higher mean and which group is more spread out?**

**Standard deviation:** The standard deviation for men of 5, is much smaller than the standard deviation for women of 21.6 so the weights for women are more spread out.

**Averages:** Females had a mean weight of 64kg and median of 66kg before the diet. There’s quite a difference between the two measures suggesting that the data may be skewed. Males had a mean and median pre-weight of 79kg suggesting that the data is normally distributed.

**Minimum/ maximum:** Are there any extreme outliers? Someone weighed -70kg before the diet which is clearly an error. Outliers cannot always be removed/ changed but here the real weight might be 70kg so make that adjustment and re-run the analysis. What effect has this had on the summary statistics?

1. **How could the chart be improved and is there anything odd?**

Better labelling of variables. Someone weighed -70kg which is clearly wrong

Before the next section, change the error of -70 to 70. Outliers should not normally be changed unless they are clearly data entry errors as in this case. Give the variables sensible labels and label gender with 0 = Female and 1 = Male. **Re-run explore to see how the change has affected the summary statistics. Which summary statistics have changed the most?**

|  |  |  |
| --- | --- | --- |
|  | Female with outlier | Female after changing outlier |
| Minimum | -70 | 58 |
| Maximum | 82 | 82 |
| Mean | 64 | 67 |
| Median | 66 | 67 |
| Standard Deviation | 21.6 | 5.6 |

The mean, standard deviation, minimum and maximum are more influenced by outliers than the median and interquartile range.

**Which diet seems the best and which diet has the most variation in weight loss? Diet 3 is the best for losing weight. Diet 2 has the most variation but the standard deviations are all similar.**

#### Exercise 6: Confidence intervals

Use Explore to get the confidence intervals of weight lost by diet.

|  |  |  |
| --- | --- | --- |
| Diet | Mean | 95% Confidence interval |
| 1 | 3.3 | (2.35, 4.25) |
| 2 | 3.03 | (2.03, 4.02) |
| 3 | 5.15 | (4.2, 6.1) |

What is the correct definition of a confidence interval?

* A range of values (a, b) which will include the true population mean 95% of the time.
* A 95% CI means that if you could sample an infinite number of times:
  + 95% of the time the CI would contain the true population parameter.
  + 5% of the time the CI would fail to contain the true population parameter.

How would you explain a confidence interval to a student?

We use the sample mean to represent the population mean but a group of different people would have a different mean. To allow for this variation in means, a confidence interval is used when estimating a population mean from a sample. It gives a range of possible values (a, b) within which the true population mean is likely to lie. A 95% confidence interval means that we would expect 95% of confidence intervals to contain the real population mean weight lost on that diet.

The population mean weight lost on diet 1 is likely to be between 2.35 and 4.25 kg.

The confidence Interval plot shows more clearly that on average, those on diet 3 lost more weight than those on diets 1 and 2. ANOVA is the appropriate test to look for significant differences in the mean weight lost by diet but confidence intervals and hypothesis testing are strongly related. If confidence intervals don’t overlap, it’s likely that there will be a significance difference between groups. This chart suggests that there will be no difference between diets 1 and 2 as the confidence intervals overlap a lot but the upper limit of diet 2 is close to the lower limit of diet 3 suggesting that there may be evidence of a difference.

#### Exercise 7:ANOVA

1. Explain briefly why ANOVA is called Analysis of variance instead of Analysis of means.

It is called an ‘Analysis of variance’ test as it uses the ratio of between group variation to within group variation when deciding if there is a difference between the groups. The reason we do this is this compares two different estimates of the population variance, which if the null hypothesis is true should be the same (or similar!)

1. What are the assumptions for ANOVA and how can they be tested?

|  |  |
| --- | --- |
| **Assumption** | **How to check** |
| **Normality:** The residuals (difference between observed and expected values) should be normally distributed | Histograms/ QQ plots/ normality tests of residuals |
| **Homogeneity of variance** (each group should have a similar standard deviation) | Levene’s test |

#### Exercise 8: ANOVA output

The ANOVA table:

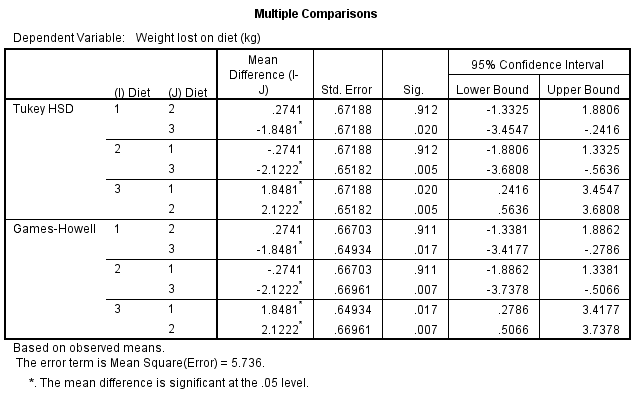
F = Test statistic

MSDiet = 35.547 =6.197

MSerror 5.736

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tests of Between-Subjects Effects** | | | | | |
| Dependent Variable: Weight lost on diet (kg) | | | | | |
| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
| Corrected Model | 71.094a | 2 | 35.547 | 6.197 | .003  P = p-value = sig = P(TS > 6.197)  p = 0.003 |
| Intercept | 1137.494 | 1 | 1137.494 | 198.317 | .000 |
| Diet | 71.094 | 2 | 35.547 | 6.197 | .003 |
| Error | 430.179 | 75 | 5.736 |  |  |
| Total | 1654.350 | 78 |  |  |  |
| Corrected Total | 501.273 | 77 |  |  |  |
| a. R Squared = .142 (Adjusted R Squared = .119) | | | | | |

ANOVA uses the F-test to test the null hypothesis that all the group means are the same. If the p-value is less than 0.05, reject the null hypothesis and conclude that there is a significant difference between at least one pair of means. When reporting the p-value, never report p = 0. Here, the p-value would be reported as p = 0.003, so there is highly significant evidence to suggest a difference between at least one pair of means. To find out where the differences lie, post hoc tests are needed. Here, just the Tukey tests are reported.



**Summary of pairwise comparisons:**

|  |  |
| --- | --- |
| Test | p-value |
| Diet 1 vs Diet 2 | P = 0.912 |
| Diet 1 vs Diet 3 | P = 0.02 |
| Diet 2 vs Diet 3 | P = 0.005 |

There is no significant difference between Diets 1 and 2 but there is between diet 3 and both the others. Looking back at the descriptive statistics, the mean weight lost on Diets 1 (3.3kg) and 2 (3kg) is less than the mean weight lost on diet 3 (5.15kg). Therefore, for those looking to lose weight, Diet 3 would be recommended.

**Explaining a p-value**

If you repeated a study numerous times you would get a variety of test statistics which form a distribution.

p-value = probability of getting a test statistic as extreme as the one calculated, **if the null is true.** The smaller the p-value, the smaller the chance of rejecting the null hypothesis when it is actually true.

p-value example

Null hypothesis: The mean weight lost on each diet is the same

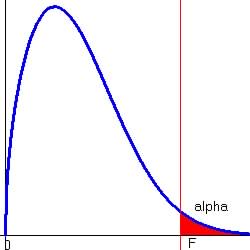
Alternative: The mean weight lost on each diet is NOT the same

Test Statistic = F = Mean between group sum of squared differences

Mean within group sum of squared differences

P – value = probability of getting a test statistic as large as ours (or larger) IF the null is true (i.e. no difference between means). The p-value is calculated using the F-distribution for ANOVA which is a skewed distribution.

Distribution of test statistics if the null is true

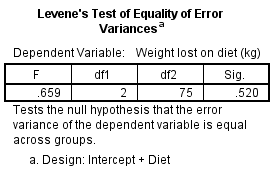


Test Statistic = 6.197

p-value = P(F > Test Statistic)= 0.003

The p-value is the probability of getting a test statistic of at least 6.197 if there really was no difference between the groups. There is a 0.3% chance of rejecting the null when it is actually true so we can be very confident with our decision to reject the null and conclude that there is a difference between the means.

#### Exercise 9: Testing the assumptions of homogeneity and normality:



**Can equal variances be assumed?**

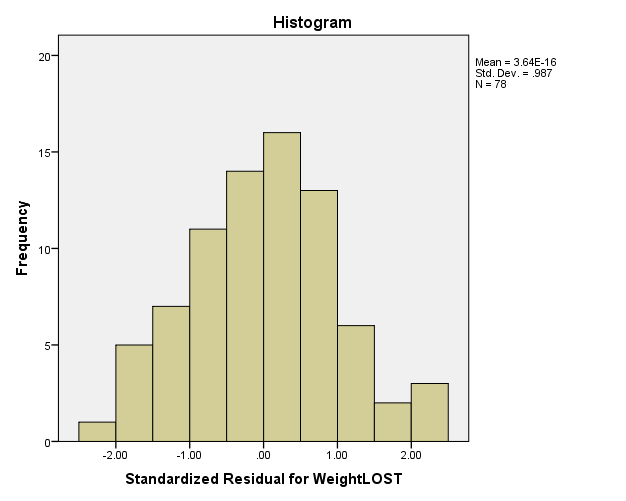
Null: The variances are the same.

Test Statistic: 0.659

p-value = 0.52

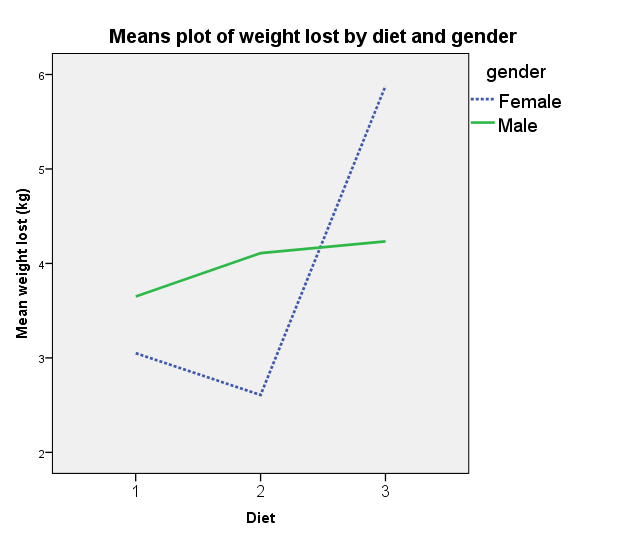
Here p = 0.52, which is above 0.05, so the assumption is met and the normal output is used.

**Are the residuals normally distributed?**



SPSS produces two normality tests but it’s clear from the histogram that the residuals are normally distributed.

**Quick question: Is there an interaction between gender and diet when it comes to weight lost?**



The chart shows how the diets affected weight lost by gender. For males there is not much difference between the diets but for women diet 3 led to a higher weight loss.

An interaction effect is when the effect of one variable on the dependent variable is altered when another variable is taken into consideration. This is important to know when doing statistical analysis. Here there is an interaction between gender and diet.

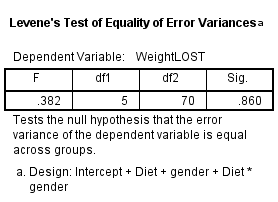
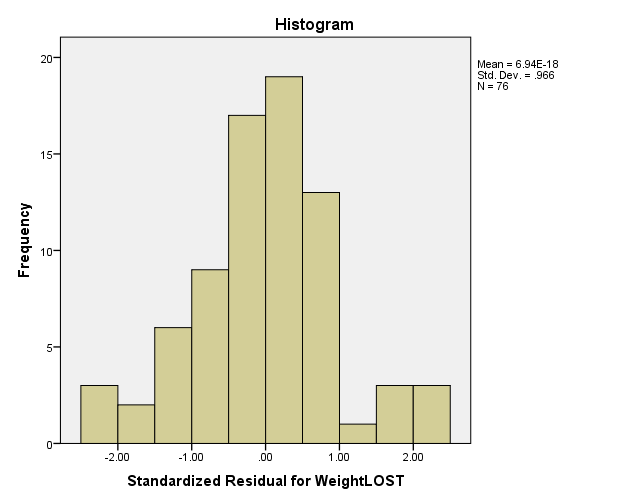
Mean weight lost for females on diet 2

#### Exercise 10: Two way ANOVA

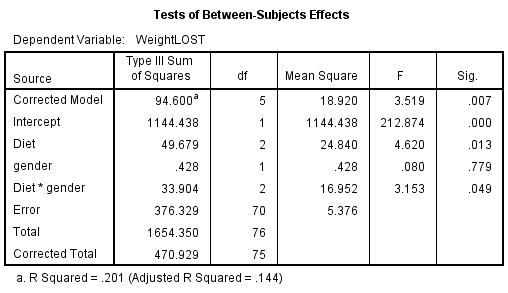
Tests 3 hypotheses:

1. Mean weight loss does not differ by diet
2. Mean weight loss does not differ by gender
3. There is no interaction between diet and gender

Results of two-way ANOVA:

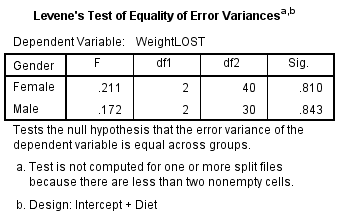
As p > 0.05, equal variances can be assumed so the first assumption has been met

The residuals look a little skewed but you can easily get histograms like this when sampling which is normally distributed. Plus the data peak approximately in the middle so normality can be assumed.



The interaction between diet and gender is significant (p = 0.049) so it is hard to interpret the main effects of diet and gender. Run separate ANOVA’s by gender.

**Results by gender:**

 Equal variances can be assumed for males and females.

|  |  |
| --- | --- |
|  |  |

Both histograms look approximately normally distributed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tests of Between-Subjects Effects** | | | | | | |
| Dependent Variable: WeightLOST | | | | | | |
| Gender | Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
| Female | Corrected Model | 92.320b | 2 | 46.160 | 10.640 | .000 |
| Intercept | 635.277 | 1 | 635.277 | 146.438 | .000 |
| **Diet** | **92.320** | **2** | **46.160** | **10.640** | **.000** |
| Error | 173.528 | 40 | 4.338 |  |  |
| Total | 917.540 | 43 |  |  |  |
| Corrected Total | 265.848 | 42 |  |  |  |
| Male | Corrected Model | 2.002c | 2 | 1.001 | .148 | .863 |
| Intercept | 524.420 | 1 | 524.420 | 77.577 | .000 |
| **Diet** | **2.002** | **2** | **1.001** | **.148** | **.863** |
| Error | 202.801 | 30 | 6.760 |  |  |
| Total | 736.810 | 33 |  |  |  |
| Corrected Total | 204.802 | 32 |  |  |  |

There is a difference between the mean weight lost on the 3 diets for females (p < 0.001) but not for males (p = 0.863). Only the post hoc tests for females should be interpreted.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Multiple Comparisons** | | | | | | | |
| Dependent Variable: WeightLOST | | | | | | | |
| Tukey HSD | | | | | | | |
| Gender | (I) Diet | (J) Diet | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| Female | 1 | 2 | .4429 | .78724 | .841 | -1.4732 | 2.3589 |
| 3 | -2.8300\* | .77401 | .002 | -4.7139 | -.9461 |
| 2 | 1 | -.4429 | .78724 | .841 | -2.3589 | 1.4732 |
| 3 | -3.2729\* | .77401 | .000 | -5.1567 | -1.3890 |
| 3 | 1 | 2.8300\* | .77401 | .002 | .9461 | 4.7139 |
| 2 | 3.2729\* | .77401 | .000 | 1.3890 | 5.1567 |
| Male | 1 | 2 | -.4591 | 1.13602 | .914 | -3.2597 | 2.3415 |
| 3 | -.5833 | 1.11326 | .860 | -3.3278 | 2.1611 |
| 2 | 1 | .4591 | 1.13602 | .914 | -2.3415 | 3.2597 |
| 3 | -.1242 | 1.08530 | .993 | -2.7998 | 2.5513 |
| 3 | 1 | .5833 | 1.11326 | .860 | -2.1611 | 3.3278 |
| 2 | .1242 | 1.08530 | .993 | -2.5513 | 2.7998 |
| Based on observed means.  The error term is Mean Square(Error) = 6.760. | | | | | | | |
| \*. The mean difference is significant at the .05 level. | | | | | | | |

For females, diet 3 is significantly different to diet 1 (p = 0.002) and diet 2 (p < 0.001) but there is no evidence to suggest that diets 1 and 2 differ (p = 0.841).

Use summary statistics to report the differences:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | Diet | | |
| 1 | 2 | 3 |
|  |  | Female | Mean | 3.05 | 2.61 | 5.88 |
| Standard Deviation | 2.07 | 2.29 | 1.89 |
| Count | 14 | 14 | 15 |
| Male | Mean | 3.65 | 4.11 | 4.23 |
| Standard Deviation | 2.54 | 2.53 | 2.72 |
| Count | 10 | 11 | 12 |

For females, the mean diet lost on diet 3 was 5.88kg compared to only 3.05kg and 2.61kg on diets 1 and 2 respectively.

#### Exercise 11: Non-parametric tests

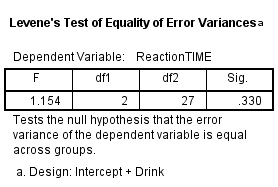
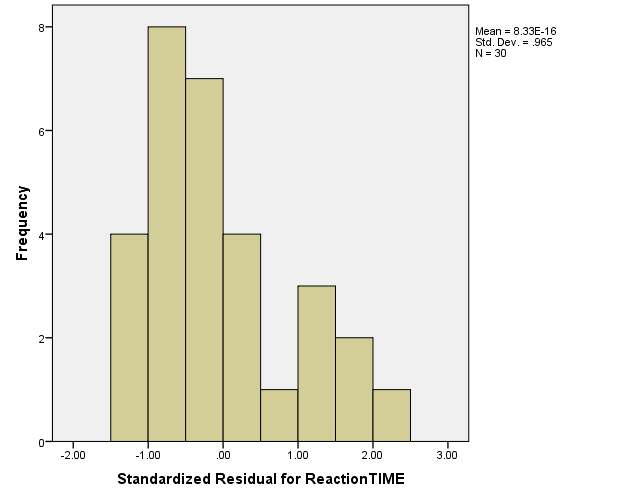
Parametric tests assume you can describe the distribution of the data using a particular distribution e.g normal. They are more likely to lead to a significant result than non-parametric tests, when the assumptions about the distribution of the data are true. Examples of non-parametric tests include t-tests, analysis of variance and linear regression.

Non-parametric tests are alternative data analysis techniques not assuming anything about the shape of the data. They are usually based on ranks or signs and can be used for ordinal, ranked or skewed scale data. Data is ordered and ranked and analysis is carried out on the ranks rather than the actual data.

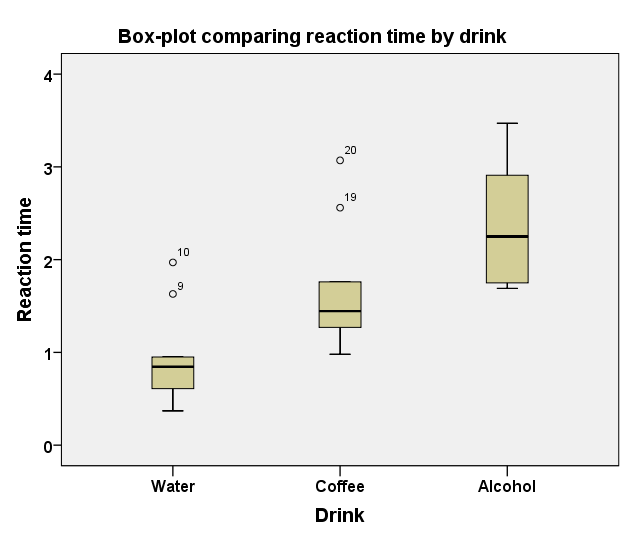
|  |  |  |
| --- | --- | --- |
| **Parametric test** | **What to check for normality** | **Non-parametric test** |
| **Independent t-test** | Dependent variable by group | Mann-Whitney test |
| **Paired t-test** | Paired differences | Wilcoxon signed rank test |
| **One-way ANOVA** | Residuals/Dependent | Kruskal-Wallis test |
| **Repeated measures ANOVA** | Residuals | Friedman test |
| **Pearson’s Correlation**  **Co-efficient** | At least one of the variables should be normal | Spearman’s Correlation Co-efficient |
| **Linear Regression** | Residuals | None – transform the data |

#### Exercise 12: Kruskal-Wallis

Enter the data into a new SPSS sheet in a suitable way to be analysed using ANOVA/ Kruskall-Wallis. Carry out a one-way ANOVA and check the assumptions. Have the assumptions been met?



Assumption of equal variances has been met but the residuals are not normally distributed so a Kruskal-Wallis test should be carried out.

Produce suitable summary statistics and follow the instructions below to perform the Kruskall-Wallis test.

A box-plot and the median with interquartile range are used when carrying out a non-parametric test.

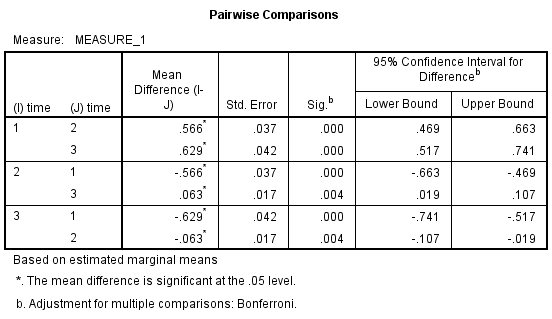
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Drink | | |
| Water | Coffee | Alcohol |
| Reaction time | Median | .85 | 1.44 | 2.25 |
| Percentile 25 | .61 | 1.27 | 1.75 |
| Percentile 75 | .95 | 1.76 | 2.91 |

It’s clear from the plot that those consuming alcohol have the slowest reaction times and the most spread out scores.

#### Exercise 13: Repeated measures example

Interpret the post hoc tests and check the assumption of normality.

There is a significant difference between each combination of time point. Cholesterol reduces by 0.566 after 4 weeks (p < 0.001) and then decreases by an additional 0.063 between 4 and 8 weeks (p = 0.004).



Does the change in mean cholesterol look meaningful? To assess this, look at the starting mean. Cholesterol drops by approximately 9% after 4 weeks which is meaningful but only drops by approximately 1% between 4 and 8 weeks. This seems less meaningful. The reason such a small change is significant is the small standard error for the differences.

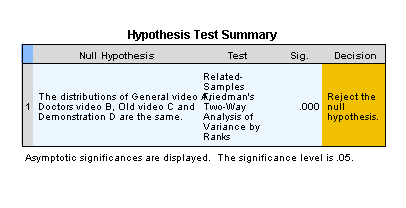
|  |  |  |
| --- | --- | --- |
|  | Mean | Standard Deviation |
| Before | 6.41 | 1.19 |
| After 4 weeks | 5.84 | 1.12 |
| After 8 weeks | 5.78 | 1.10 |

What test would you use instead if the assumption of normality has not been met?

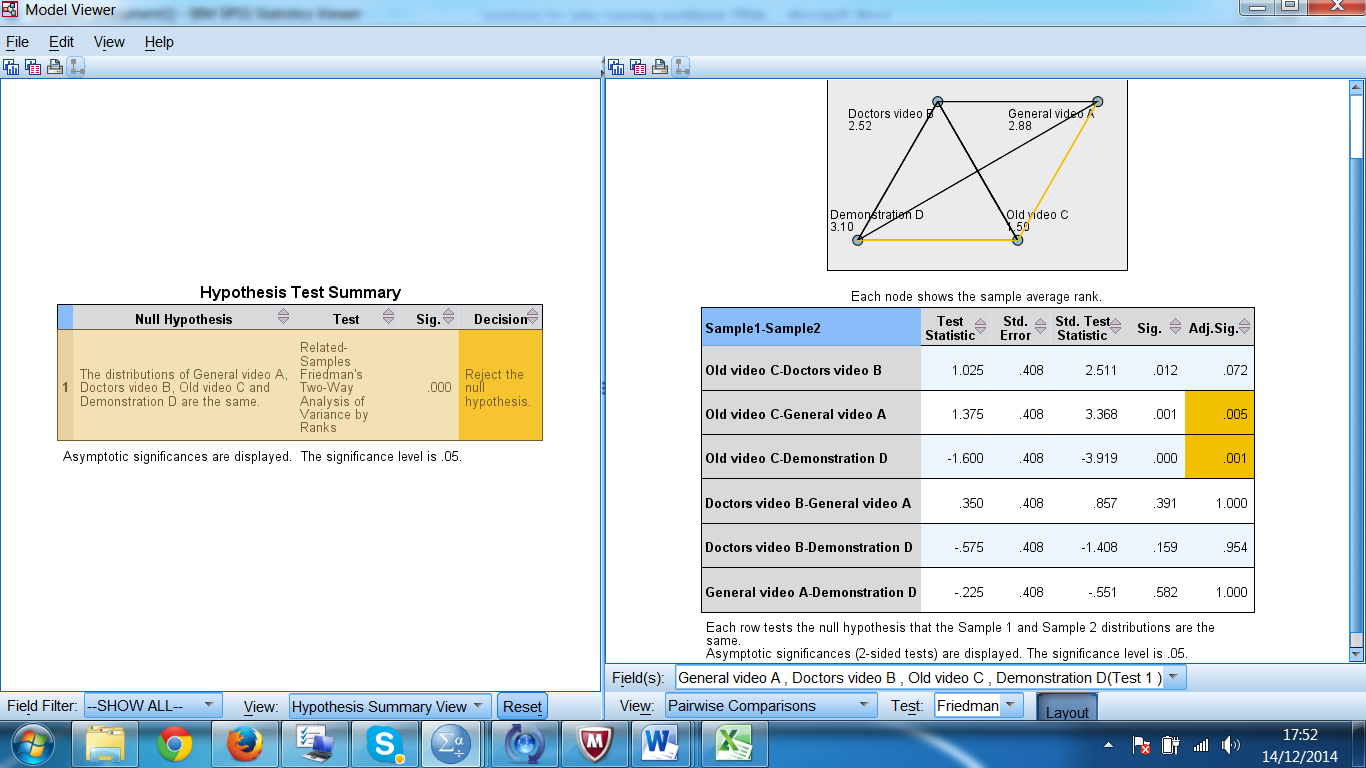
Friedman

#### Exercise 14: Friedman example

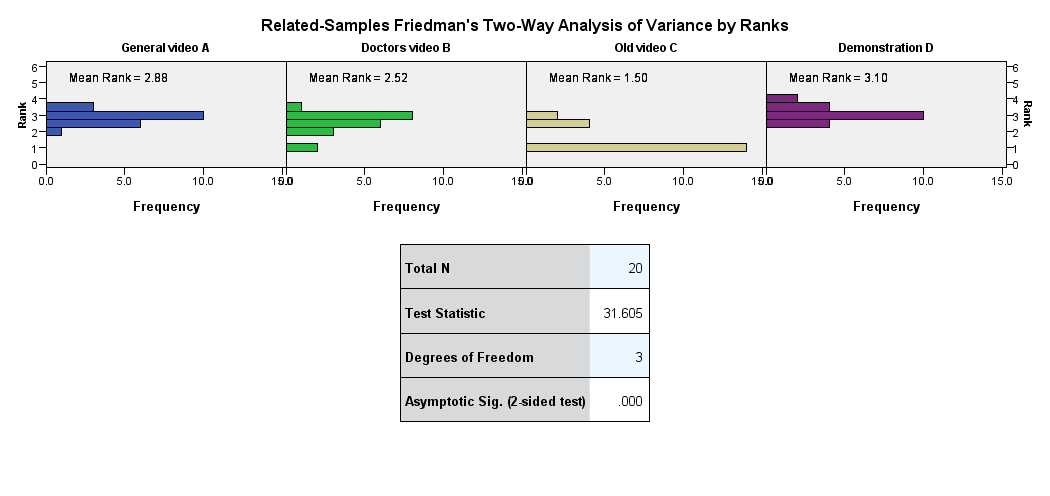
***Carry out the Friedman test and interpret the output including the post hoc tests***



The Friedman test was significant (p < 0.001) so the distributions of the scores for the videos being compared are different. Bonferroni post hoc tests were carried out and there were significant differences between the Old video C and the General video A (p = 0.005), and also between the old video C and the demonstration D (p = 0.001). The difference between the Old video C and Doctors video B (which was a more technical video aimed at doctors) was borderline not significant (p=0.072).



The bar charts below show the distribution of scores for each product. The mean rank for the Demonstration video D was highest for understanding of the condition, followed by the General video A and then the Doctors video B.



#### Exercise 15: Regression assumptions

What are the assumptions for multiple regression?

|  |  |
| --- | --- |
| **Assumption** | **Plot to check** |
| The relationship between the independent and dependent variables is linear. | Original scatter plot of the independent and dependent variables |
| Homoscedasticity: The variance of the residuals about predicted responses should be the same for all predicted responses. | Scatterplot of standardised predicted values and residuals |
| The residuals are normally distributed | Plot the residuals in a histogram |
| The residuals are independent. Are adjacent observations related? Example: Weather by day | If you suspect that the data may be autocorrelated you can use the Durbin Watson statistic. Note: Time series is beyond the scope of most students |

#### Exercise 16: Scatterplots

Describe the relationship between gestational age, smoking and birth weight. Does it look like there is an interaction between smoking and gestational age?

There is a strong positive relationship between gestational age and birth weight. It looks like smokers may have lighter babies which is consistent as gestational age increases. There doesn’t appear to be an interaction between smoking and gestational age

#### Exercise 17: Correlations

Interpret the correlations between birth weight, gestational age, height and weight of mother.

Strong relationship: Birthweight and gestational age, Height and weight of mother

Moderate relationship: Birth weight with height and weight of mother

Weak relationship: Gestational age with height and weight of mother

***Quick question: What is being tested in regression?***

***The slope. ***

#### Exercise 18:Regression

***Interpret the output from the regression including answering the following questions:***

1. Which independent variables are significant and what is their relationship with the dependent variable?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | -7.165 | 2.107 |  | -3.400 | .002 |
| Gestational age at birth (weeks) | .313 | .053 | .623 | 5.926 | .000 |
| Smoker | -.665 | .268 | -.253 | -2.485 | .017 |
| Mothers pre-pregnancy weight (lbs) | .020 | .009 | .237 | 2.261 | .030 |
| a. Dependent Variable: Weight of baby at birth (lbs) | | | | | | |

Gestational age (p < 0.001), being a smoker (p = 0.017) and weight of the mother (p = 0.03) are all significant predictors of birth weight. Birthweight increases with gestational age and weight of the mother but decreases for smokers whose babies are 0.67 lbs lighter on average.

1. What is the equation of the model?



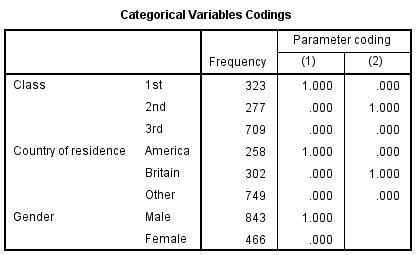
1. How reliable is the model?

R2=0.61 so the model explains 61% of the variation in birthweight which is fairly reliable but there is still 39% of the variation unaccounted for.

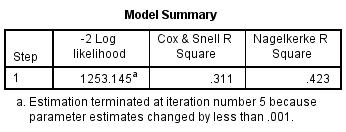
#### Exercise 19: Logistic regression

Look at the relationship between nationality and survival but control for gender and class.

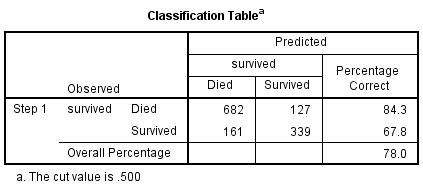
Which variables are significant? Interpret the odds ratio for those variables.



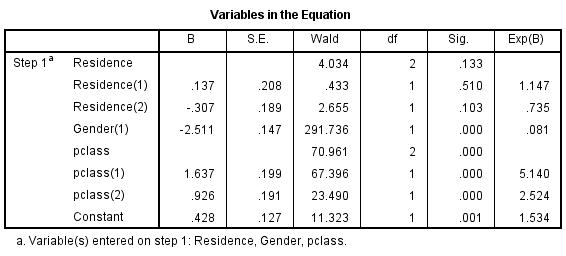
Reference categories are 3rd class, ‘Other’ nationality and female.



From the table above, we can conclude that based in Nagelkerke’s R2, 42.3% of the variation in survival can be explained by the model including nationality, gender and class.



The model correctly classifies 78% of those in the dataset.



Whilst gender and class are significant, nationality becomes insignificant once these two factors are controlled for.

Since 3rd class is the reference category for class, the ratio of 5.14 for pclass(1) means that the odds of survival for those in 1st class were 5.14 times that for those in 3rd class. Similarly, the odds of survival for those in 2nd class were 2.5 times that for those in 3rd class

Since females are the reference category for gender, the odds ratio of 0.081 for Gender(1) means that the odds of survival for men was 0.081 times that for females (i.e. they were less likely to survive!).