ARDEN UNIVERSITY

Session Handout

Study Skills Optional Workshop

Advanced Statistics

Guides to selecting the right statistical tests and graphs

Summary of descriptive and graphical statistics

Chart	Variable type	Purpose	Summary Statistics
Pie Chart or bar chart	One Categorical	Shows frequencies/ proportions/percentages	Class percentages
Stacked / multiple bar	Two categorical	Compares proportions within groups	Percentages within groups
Histogram	One scale	Shows distribution of results	Mean and Standard deviation
Scatter graph	Two scale	Shows relationship between two variables and helps detect outliers	Correlation co-efficient
Boxplot	One scale/ one categorical	Compares spread of values	Median and IQR
Line Chart	Scale by time	Displays changes over time Comparison of groups	Means by time point
Means plot	One scale/ 2 categorical	Looks at combined effect of two categorical variables on the meanof one scale variable	Means
Source: [tutorsquickguidetostatisticsprint2uppamphlet.pdf (statstutor.ac.uk)]			

Common Single Comparison Tests

Comparing:	Dependent (outcome) variable	Independent (explanatory) variable	Parametric test (data is normally distributed)	Non-parametric test (ordinal/ skewed data)
The averages of two INDEPENDENT groups	Scale	Nominal (Binary)	Independent t-test	Mann-Whitney test/ Wilcoxon rank sum
The averages of 3+ independent groups	Scale	Nominal	One-way ANOVA	Kruskal-Wallis test
The average difference between paired (matched) samples e.g. weight before and after a diet	Scale	Time/ Condition variable	Paired t-test	Wilcoxon signed rank test
The 3+ measurements on the same subject	Scale	Time/ condition variable	Repeated measures ANOVA	Friedman test
Source: *tutorsquickguidetostatisticsprint2uppamphlet.pdf (statstutor.ac.uk)*				

Tests of Association

Comparing:	Dependent (outcome) variable	Independent (explanatory) variable	Parametric test (data is normally distributed)	Non-parametric test (ordinal/ skewed data)
Relationship between 2 continuous variables	Scale	Scale	Pearson's Correlation Coefficient	Spearman's Correlation Coefficient
Predicting the value of one variable from the value of a predictor variable or looking for significant relationships	Scale	Any	Simple Linear Regression	Transform the data
	Nominal (Binary)	Any	Logistic regression	
Assessing the relationship between two categorical variables	Categorical	Categorical		Chi-squared test
Source: *tutorsquickguidetostatisticsprint2uppamphlet.pdf (statstutor.ac.uk)*				

One scale dependent and several independent variables

1st independent	2nd independent	Test
Scale	Scale/ binary	Multiple regression
Nominal (Independent groups)	Nominal (Independent groups)	2 way ANOVA
Nominal (repeated measures)	Nominal (repeated measures)	2 way repeated measures ANOVA
Nominal (Independent groups)	Nominal (repeated measures)	Mixed ANOVA
Nominal	Scale	ANCOVA
Source: tutorsquickguidetostatisticsprint2uppamphlet.pdf (statstutor.ac.uk)		

Which one of Regression or ANOVA to use? Use regression if you have only scale or binary independent variables. Categorical variables can be recoded to dummy binary variables but if there are a lot of categories, ANOVA is preferable.

Hypothesis testing for Inferential Statistics

Inferential Statistics is used to make conclusions/inferences and take decisions about data. Such conclusions and decisions include comparing data of a dependent variable across various groups to check if they are the same or if differences exist across the groups (i.e., **test of comparison**), checking whether there is association/relationship between two or more variables (i.e., **test of association** or **correlation**), evaluating the nature of the relationship between variables in form of equation/formula (**Regression**).

Processes in Hypothesis Testing

(1) Understand the research or study, independent variable(s), dependent variable(s) and the end goals or purpose of the research properly.
(2) Form the **Research Question** **correctly**: This is usually to explore a comparison, association, or relationship
(3) Determine the **correct types of variables** and **level of measurement** of the variables: Usually Categorical (Ordinal or Nominal) or Scale (Continuous)
(4) Form the **Null hypothesis (H_o)** and **Alternative Hypothesis (H_1)**.

Null Hypothesis (H_o) usually assumes no difference, association or relationship between the variables. It also means the hypothesis test parameter is not significant. The **Alternative Hypothesis** (H_1) assumes difference, association or relationship between the variables. It also means the hypothesis test parameter is significant.

Examples of Null Hypothesis (Ho) are:

- For **comparing the means** of a dependent variable (e.g., response time) across various levels (Teenagers and Adults) of an independent variable, Age Group.

Ho: There is no significant difference in the mean response time of teenagers and adult participants
H_1(two-sided): There is significant difference in the mean response time of teenagers and adult participants
H_1(one-sided): The mean response time of teenager is faster than that of adult OR The mean response time of adult is faster than that of teenager

- For checking **association and relationship** between variables. If there is association between an independent variable (x) and a dependent variable (y), we obtain a formula or equation relating them together in the form of $y = mx + c$ where **m** is the slope and **c** is the intercept. If truly a relationship exists between the variables, **m** and **c** shouldn't both be 0. Sometimes **c** can still be 0 but **m** should never be 0 because it is **m** we use to multiply the value of the independent variable (x). So, we need to put the values of **m** and **c** obtained by regression analysis to test to know **whether m** and **c** are **truly those values obtained** (i.e., **whether they are significant**) or **whether they are just random occurrence and rather zero** (i.e., **whether they are not significant**).

Ho: The correlation and regression between the variables is not significant **OR** The regression values (m and c) are not significant **OR** The regression values (m and c) are zero
H1: The correlation and regression between the variables is significant **OR** The regression values (m and c) are significant **OR** The regression values (m and c) are not zero.

- For checking **normality** of a dependent variable (e.g., response time).

Ho: The data for the dependent variable is normally distributed. Accepting H_o means that the normality test is satisfied, and parametric tests can be used.
H1: The data for the dependent variable is not normally distributed. Accepting H_1 means that the normality test is not satisfied, and parametric tests may not be used.

- For checking **homogeneity (i.e., equality of population variances)** of a dependent variable (e.g., response time) across various levels (Teenagers and Adults) of an independent variable, Age Group.

Ho: There is no significant difference in the population variance of response time of teenagers and adult participants. Accepting H_o means that the homogeneity test is satisfied, and parametric tests can be used.
H1(two-sided): There is significant difference in the population variance of response time of teenagers and adult participants. Accepting H_1 means that the homogeneity test is not satisfied, and parametric tests may not be used.

(5) **Perform the inferential statistics**. Decide on whether parametric or non-parametric tests should be used. Parametric tests make assumption that the data follows normal distribution, while non-parametric tests don't make assumption about the distribution of the data.
(6) Interpret the results by checking the probability **p-value** which can be found under **Sig** column in the table of output.

A decision between the two hypotheses is made by viewing the 'p-value' or 'Sig-value', which is the probability (or chance) of getting the collected data (or more extreme) under the assumption of the null hypothesis.

If this probability is small, 'H_o is rejected in favour of H_1', termed a 'statistically significant result'; otherwise, we 'don't reject H_o' which is termed a 'non-statistically significant result'.

What is 'small'?

Conventionally, we use p = 0.05 meaning Significance level of 5% or confidence level of 95%; hence, if

- $p \leq 0.05$ ('p is less than or equal to 0.05'), reject H_o in favour of H_1
- $p > 0.05$ ('p is greater than 0.05'), fail to reject H_o

Sometimes, when the p-value is close to 0.05, it might limit the level of trust we have in the output. So, we can extend it sometimes to:

Alternatively, interpret the p-value so that:

- $p > 0.1$ implies **no evidence to reject H_o**
- $0.05 < p < 0.1$, that is when the value of p is in between 0.05 and 0.1, it implies **some weak evidence to reject H_o**
- $0.01 < p < 0.05$, that is when the value of p is in between 0.01 and 0.05, implies **evidence to reject H_o**
- $p < 0.01$ implies **strong evidence to reject H_o**

(7) Always relate outcome of the hypothesis testing and statistics to the particular variables in the study; don't just conclude with 'reject the null hypothesis'. For example, state that, there is strong statistical evidence that there is no significant difference in the mean response time of teenagers and adult participants. Always interpret results within the context of the research or study.