A Very Brief Introduction to Hypothesis Testing

Hypothesis testing framework

There are always two hypotheses:

H_A: Research (Alternative) Hypothesis

- What we aim to gather evidence of
- Typically, that there **is** a difference/effect/relationship etc.

H_0: Null Hypothesis

- What we assume is true to begin with
- Typically, that there is **no** difference/effect/relationship etc.

For example, if we wanted to test whether a new drug was effective in increasing weight loss, we might carry out trials on two groups of people, one having the drug and the other taking a placebo. These groups of people are SAMPLES taken from the whole POPULATION of interest. We use information from the samples to say something about the population, so need to take into account that samples may vary, and our particular sample is only one possible sample. The means and variance of the two groups can be calculated and the mean weight loss compared. The hypotheses would be as follows.

H_0: Null Hypothesis *(What we assume is true)*

NO difference in mean weight loss between New Drug and Placebo in the **POPULATION,**
i.e. $\mu_{\text{Placebo}} = \mu_{\text{NewDrug}}$

(Note the use of the Greek letter μ to represent the population mean; the sample mean is $\bar{x}$.)

H_A: Research (Alternative) Hypothesis *(What we look to find evidence of)*

There **IS a** difference in mean weight loss between New Drug and Placebo in **POPULATION,**
i.e. $\mu_{\text{Placebo}} \neq \mu_{\text{NewDrug}}$

We now use a statistical package (such as SPSS or R) to calculate the test statistic, which in this example would be for the comparison of two means; the t statistic if the data is from a normal distribution, or the Mann-Whitney U value if the data is not normal. We also calculate the significance or p value, which is the probability of getting this value for the test statistic **if the null hypothesis were true.** A high p value means that our result is likely, so we cannot reject H_0. A low p value (usually taken to be less than 0.05) means that it is unlikely that we would get this result if the null hypotheses were true, so we reject H_0 and accept the alternative, i.e. we have some evidence for our alternative hypothesis.

- The p value is the probability of getting our data **if** the null hypothesis is true.
- If p is “big”, then it is quite likely, so we can’t reject the null hypothesis.
- If p is “small”, then the situation is unlikely, so we reject H_0 and accept the alternative hypothesis.
- We usually define “small” as 0.05 (a 1 in 20 chance)

An alternative approach is to calculate confidence intervals. In our example we could calculate a confidence interval for the difference in means between the two groups. This interval provides a range which we are 95% confident (i.e. for 95 samples out of 100) will contain the true difference in means for the population. If the confidence interval does not contain zero, then we can conclude that there IS evidence of a difference in the means.

What can go wrong?

Extreme or “bad” samples

We might, by chance, pick a group of people in the “drug” group who had a large weight loss compared to the “placebo” group, even though the drug actually had no effect. We would reject the null hypothesis and conclude the drug was effective. This is a “false positive”, a Type I error, and the probability of it occurring is the cut-off level we chose (0.05 or 5%).

Suppose the drug actually WAS effective. Just by chance we might get a sample where the group of people taking the drug had a similar weight loss to those on the placebo, so we would not reject the null hypothesis, and would conclude that there was no benefit from the drug, even though if we had taken a different sample of people there would have been a difference. This is a “false negative”, a Type II error.

There is a balance to be made between the two types of error: reducing the probability of a Type I error increases the probability of a Type II error (making it less likely to conclude there IS an effect when there is NO effect, increases the probability of NOT finding an effect when there actually is one.) The best ways of reducing both types of error are by increasing sample size and by careful design of the data collection.

There are many different tests for different types of data and hypotheses. The next section gives a summary of some of the tests and when their use is appropriate.

WHAT STATISTICAL TEST DO I NEED? From www.statstutor.ac.uk

Deciding on appropriate statistical methods for your research:

- What is your research question?

- Which variables will help you answer your research question and which is the dependent variable?

- What type of variables are they?

- Which statistical test is most appropriate? Should a parametric or non-parametric test be used?

KEYWORDS:

VARIABLE: Characteristic which varies between independent subjects.

CATEGORICAL VARIABLES: variables such as gender with limited values. They can be further categorised into NOMINAL (naming variables where one category is no better than another, e.g. hair-colour) and ORDINAL (where there is some order to the categories e.g. 1st, 2nd 3rd etc.)

CONTINUOUS (SCALE) VARIABLES: Measurements on a proper scale such as age, height etc.

INDEPENDENT VARIABLE: The variable we think has an effect on the dependent variable.

DEPENDENT VARIABLE: The variable of interest which could be influenced by independent variables.

PARAMETRIC TESTS: there are various assumptions for parametric tests including the assumption that continuous dependent variables are normally distributed. There are specific tests for this within packages such as SPSS but plotting a histogram is also a good guide. As long as the histogram of the dependent variable peaks in the middle and is roughly symmetrical about the mean, we can assume the data is normally distributed (see examples below).

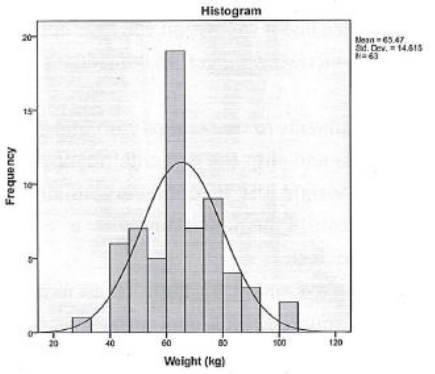
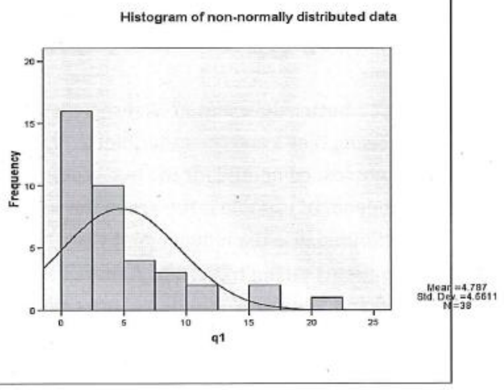
Example of data which is approximately normally distributed	Example of skewed data
	

Table of tests	**Dependent (outcome) variable**	**Independent (explanatory) variable**	**Parametric test**	**Non-parametric alternative**
Comparing means				
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 averages of 2 paired (matched) samples e.g. weight before and after a diet	Scale	Nominal Time/condition variable	Paired t-test	Wilcoxon signed rank test
The 3+ measurements on the same subject	Scale	Nominal	Repeated measures ANOVA	Friedman test
Investigating relationships				
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	Scale	Any number of scale or binary	Simple Linear Regression	Transform the data
	Binary	Any number of scale or binary	Logistic regression	
Assessing the relationship between 2 Nominal variables	Nominal	Nominal		Chi-squared test

Note: The table only shows the most common tests for simple analysis of data.

Examples:

- **Are height and weight related?** Both are continuous variables so Pearson's Correlation Co-efficient would be appropriate if the variables are both normally distributed.

- **Can height predict weight?** You cannot determine height from weight, but you could estimate weight given height, so height is the continuous independent variable. Simple linear regression will help decide if weight is a good predictor of height and produce an equation to predict weight given an individual's height.

- **Is Diet 1 better than Diet 2?** A researcher would randomly allocate subjects to two groups with one group following Diet 1 and the other Diet 2. Weight would be taken before and after the diet and the mean weight loss compared for the two groups. The dependent variable 'weight lost' is continuous and the independent variable is the group the subject is in, which is categorical. If the data is normally distributed, use the independent t-test, if not use the Mann-Whitney test.

- **Are patients taking treatment A more likely to recover than those on treatment B?** Both 'Treatment' (A or B) and 'Recovery' (Yes or No) are categorical variables, so the Chi-squared test is appropriate.

Tests when there are several independent variables, and the dependent variable is a scale variable.

Either multiple regression or Analysis of variance

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

Regression or ANOVA? 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.

Both techniques are based on a linear model, where values of the dependent variable are explained by differences in the values of the independent variables. Regression is often used to *predict* the value of the dependent variable, given values for the independent variables (e.g. predicting weight based on height, gender, age and income). ANOVA is used to *explain the variation* in the dependent variable by identifying the factors (independent variables) which have the most effect.

Resources

- The SPSS Help menu: Topics search, Tutorials and Statistics Coach can be very useful.
- Books: ask colleagues for recommendations or just look in the library for something that suits you.
- Two suggested books are: “SPSS for Psychologists” by Brace, Kemp & Snelgar and “Discovering Statistics using IBM SPSS Statistics” (or “Discovering Statistics using R”) by Andy Field.
- The internet: the web has many examples, tutorials, guides, etc. that various people have put on, with the intention of helping people in a similar position to you. They can be very useful if you’re not sure about something related to either SPSS or statistics in general. Try
    - Statstutor, http://www.statstutor.ac.uk/, which has worksheets and guides on many commonly used statistical techniques
    - Laerd Statistics: e.g. information on data input is found at https://statistics.laerd.com/spss-tutorials/entering-data-in-spss-statistics.php

Mary Lorimer, Jan 2023

How to use the Flow Chart of Statistical tests

1. **First consider the type of the dependent variable,**
2. **then the type and number of the independent variables (predictors),**
3. **then consider what is the aim of the analysis.**

In more detail:

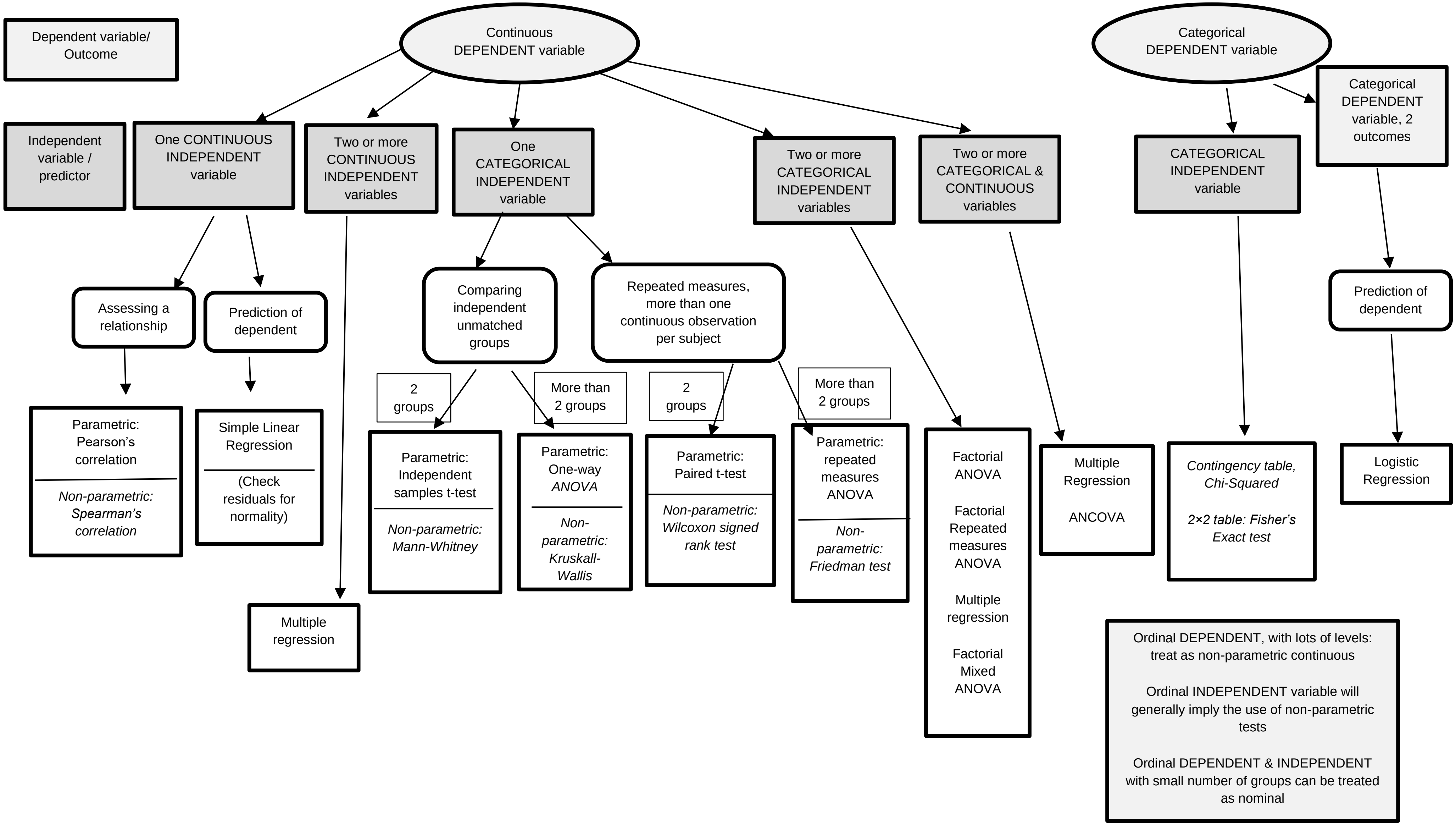
1. Is the dependent variable scale, ordinal or categorical?

2. If it is categorical then there is the choice of
    a. a Chi-squared test of association to check for association between two categorical variables
    b. logistic regression to predict which category based on values for the independent variables
    c. loglinear analysis to determine if there is a statistically significant relationship among three or more discrete variables. It is typically used if none of the variables in the analysis are considered dependent variables, but rather all variables are considered variables of interest

3. Is the dependent variable ordinal?
    a. If it has only a few possible values then treat it as categorical
    b. If it has a large number of possible values then treat it as continuous (scale)

4. Is the dependent variable a continuous (scale) variable? This is where there is most choice of test. To decide what would be appropriate the flow chart leads you through these three questions:
    a. what type of variable (continuous or categorical) are the independent variables?
    b. is there one independent variable or more than one independent variable?
    c. what type of question are we trying to address (relationship, prediction, difference)?
    d. Do we have independent groups or more than one observation per subject (repeated measures)?
    e. Can we do parametric tests or must we use a non-parametric test.

Flow Chart for statistical tests



Dependent variable/ Outcome

Independent variable / predictor

Continuous DEPENDENT variable

One CONTINUOUS INDEPENDENT variable

Assessing a relationship

Parametric: Pearson's correlation

Non-parametric: Spearman's correlation

Prediction of dependent

Simple Linear Regression

(Check residuals for normality)

Two or more CONTINUOUS INDEPENDENT variables

Multiple regression

One CATEGORICAL INDEPENDENT variable

Comparing independent unmatched groups

2 groups

Parametric: Independent samples t-test

Non-parametric: Mann-Whitney

More than 2 groups

Parametric: One-way *ANOVA*

Non-parametric: Kruskall-Wallis

Repeated measures, more than one continuous observation per subject

2 groups

Parametric: Paired t-test

Non-parametric: Wilcoxon signed rank test

More than 2 groups

Parametric: repeated measures ANOVA

Non-parametric: Friedman test

Two or more CATEGORICAL INDEPENDENT variables

Factorial ANOVA

Factorial Repeated measures ANOVA

Multiple regression

Factorial Mixed ANOVA

Two or more CATEGORICAL & CONTINUOUS variables

Multiple Regression

ANCOVA

Categorical DEPENDENT variable

CATEGORICAL INDEPENDENT variable

Contingency table, Chi-Squared

2×2 table: Fisher's Exact test

Categorical DEPENDENT variable, 2 outcomes

Prediction of dependent

Logistic Regression

Ordinal DEPENDENT, with lots of levels: treat as non-parametric continuous

Ordinal INDEPENDENT variable will generally imply the use of non-parametric tests

Ordinal DEPENDENT & INDEPENDENT with small number of groups can be treated as nominal