PL2132 Textbook Notes

Tan Wee Han

January 31, 2020

Contents

Chapter 1

Spine of Statistics

1.1 Spine of Statistics acronym

Statistical models
Parameters
Interval Estimates (Confidence Intervals)
Null hypothesis significance testing
Estimation

1.2 Statistical models

All outcome generally boils down to one equation:

$outcome_i = model + error_i$, Where i is the different observations in a variable

1.3 Parameters

Parameter is any numerical quantity that characterizes a given population or some aspect of it. Parameters tells us something about the whole population. There is no inferences made.

Common parameters include, *central tendency (mean, median, mode).*

If we only want to just summarise the outcome of experiment and not use the experiment to predict real world, our model would not need variables in it (e.g. when we are just computing the mean). We can write the equation as:

$$outcome_i = (\hat{b}_0) + error_i,$$

where $\hat{b}_0$ is a constant, in this case it is the mean of the outcome. we use ˆto make explicit that the values underneath them are estimates.

If we want to predict an outcome from a variable, we need to expand our model to include a variable.

$$outcome_i = b_0 + (b_1 x_{1i} + b_2 x_{2i}) + error_i$$

where b_0 is a constant, predictor variable (x_{1i}, x_{2i})

Often, we can only use parameter estimates instead of calculating exact values of things, this is because we can never know the parameter of a population, experiments

only uses samples. so we can only use sample data to *estimate* what the population parameters are.

meanings of error: Residual, Error, Deviance, Deviation

1.4 Assessing fit of model: sum of squares and variance

When we are assessing errors, some errors might be positive and others negative. So we use sum of squared errors in order to prevent errors from canceling out each other.

$$\text{Sum of squared errors (SS)} = \sum_{i=1}^{n}(outcome_i - model_i)^2$$

e.g. If our model was calculating the mean, symbolised by $\bar{x}$, and outcome was replaced by x, then you will get:

$$\sum_{i=1}^{n}(outcome_i - model_i)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

If we are interested in the error in the model in the population and not the sample, we can estimate the mean error of the population by dividing by the degrees of freedom. This equation will be the unbiased estimate of the population variance.

$$\text{Mean squared error} = \frac{SS}{df} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

1.5 Estimating Parameters

Estimation is the "E" in the SPINE of statistics

Whenever we are estimating parameters, we want our model to have the minimised sum of squared errors, i.e. ordinary least squared (OLS). Our model for calculating least squared error for a data is using the mean. But we can also randomly guess the average. However, our random guesses will always have a larger squared error than when using the mean.

1.6 Standard Error

Standard error is the "S" in SPINE of statistics

Sampling distribution of mean tells us about the behaviour of samples from the population, the distribution is centred at the mean of the population. If we take the average value of all sample means, we will get the population mean. Standard deviation is a measure of how representative the mean was of the observed data. Small data points means most data point is close to the mean. Similarly standard deviation of sample means tells us how widely spread the sample means are around their average (which is the population mean). So it tells us whether the sample means are representative of the population mean.

Standard error of the mean is the standard deviation of the sample means.

A large standard error means that there is a lot of variability between the means of different samples, so the sample mean we have may not be representative of the population. Small standard error means most sample means are similar to population mean.

1.7 I is for confidence interval

think of confidence interval this way, if we collected 100 of the mean and confidence intervals. 95 of these samples will contain the true mean of the population while 5 of the samples will not contain the true mean in the confidence interval.

1.8 N is for Null hypothesis significance testing

hypothesis testing arose out of: 1) Ronald Fisher's idea of computing probabilities to evaluate evidence, 2) Jerzy Neyman and Egon Pearson's idea of competing hypothesis.

p value is a long run probability: it is computed by working out how often you get specific values of the test statistics (in this case t) if you repeated your exact sampling process an infinite times. it is is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

important to collect the amount of data you set out to collect otherwise, p value obtained will not be correct. If you cut data collection short by arbitrary reason, the p value that you end up with will not be the value that you want

1.9 test statistics

NHST relies on fitting a model to the data and then evaluating the probability of this model, given the assumption that no effect exists.

In explaining the data using a model there is systematic and unsystematic variation.

systematic variation (signal): variation that can be explained by the model that we fit to the data

unsystematic variation (noise): variation that cannot be explained by the model.

test statistics is usually just a measure of signal to noise ratio:

$$\frac{signal}{noise} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

Most test statistics represent similar thing: signal to noise ratio.
Different types of test statistics: t, χ^2, F

1.10 One tailed vs two tailed tests

important thing to note. If you do a one tailed test and the result turn out to be in the opposite direction to what you predicted, you must ignore, and cannot interpret them, and accept the null hypothesis. If you *don't* do this, then you have done a two-tailed test using a different level of significance from the one you set out to use.

one tailed tests encourage cheating, if you find your p to be .06 when you do a two a two tailed tst you will conclude that the results is not significant (as 0.06 is bigger than the critical value of .05). However, if you had done a one tailed test, p value will be .03 (half of the two tailed test) and this is less than .05. Therefore if we find a two tailed p that is just non significant, we might be tempted to pretend we always intended to do a one tailed test because our "one tailed" p value is significant. (for two tailed test, you can either half the alpha value or twice the p value.)

1.11 type I and type II errors

Type 1 error: Rejecting null hypothesis even when null hypothesis is true Type 2 error: Accepting null hypothesis when null hypothesis is false.

Cohen suggested that the maximum acceptable probability of a Type II error is .2 (20%), this is called the β level.

As probability of Type I error decreases, probability of Type II error increases.

1.12 Inflated error rates

When we do multiple tests, probability of having no Type I errors decreases. if we do three tests, probability of not having Type I error is $.95^3 = .857$. So probability of making at least one Type I error is 1 - .857 = .143, which is more than the initial .05%

Error rate across statistical tests conducted on the same data is called familywise or experimentwise error rate.

Familywise error $= 1 - (1 - \alpha)^n$, where n is the number of test conducted.

In order to combat the build up of errors, we can use Bonferroni correction to adjust the level of significance for individual tests such that the overall Type I error rate (α) remains at .05.

Bonferroni correction: $P_{\text{crit}} = \frac{\alpha}{k}$, where k is the number of comparisons

However, trade off for controlling the familywise error rate is the loss of statistical power.

1.13 Statistical Power

Power: probability that a given test is significant assuming that null hypothesis is false. $(1 - \beta)$
Power of statistical test depends on:

1. Effect size
2. How strict are we in deciding that an effect is significant. if we apply Bonferroni correction, tests will have less power to dectect effects
3. Sample Size

Given that power, alpha, sample size, effect size are all related, if we know three of the elements we can calculate the remaining one.

1.14 Confidence intervals and statistical significance

1. 95% confidence intervals that just touch end to end represents a p value of .01 (top left figure)
2. If there is a gap between the upper end of one 95% confidence interval and the lower end of another, then $p < .01$ (top right)
3. p-value of .05 is represented by a moderate of the confidence interval bars (bottom left)

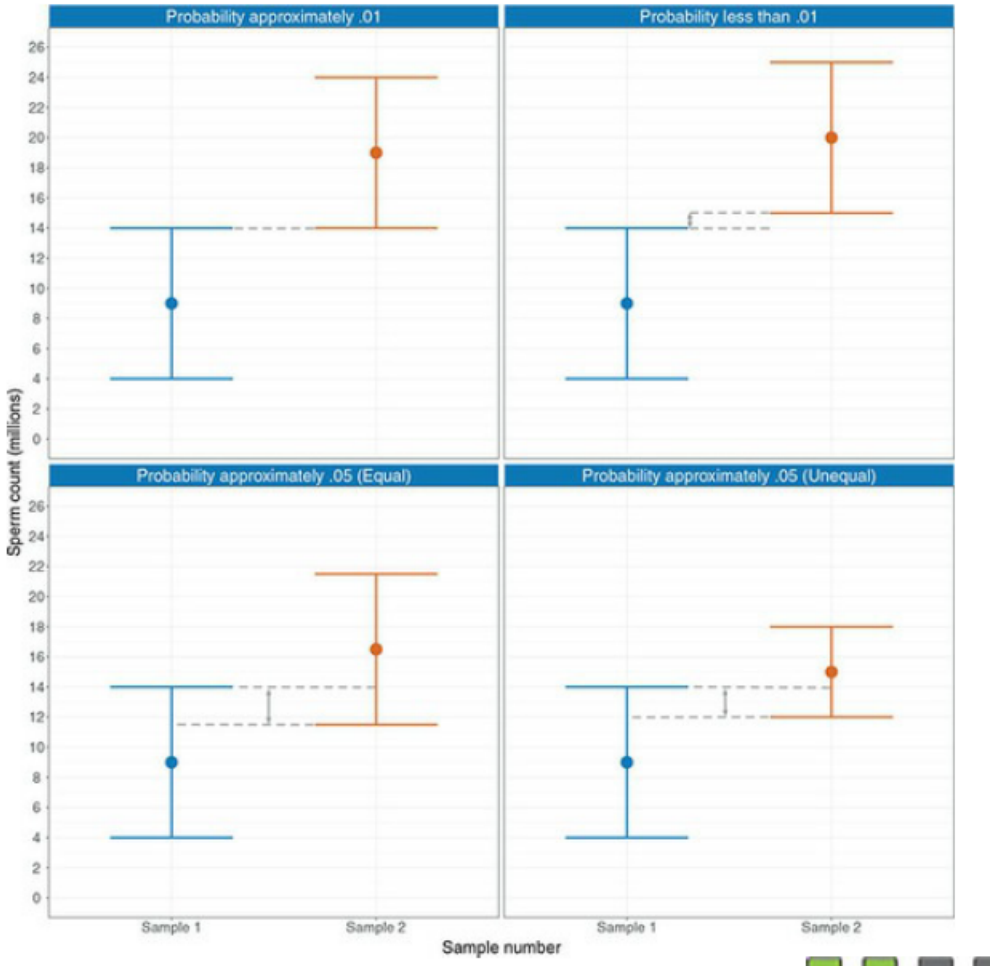

Figure 1.1: Confidence interval and p value

1.15 Sample Size and Statistical Significance

Sample size affects statistical significance via standard error. Even if standard deviation is the same, confidence interval is computed by the mean ± 1.96 * standard error. So as sample size gets larger, the standard error will become smaller. So a tiny difference in mean will result in being statistically significant, if samle size was large enough

Chapter 2

The Pheonix of Statistics: Problem in Statistics

2.1 Problems with NHST

> The almost universal reliance on merely refuting the null hypothesis is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.
>
> Meehl, 1978

Misconceptions about statistical significance

1. A significant result means that the effect is important.

    - statistical significance is not the same thing as importance as p value is affected by sample size. small unimportant effects can also be statistically significant, if sufficiently large data is collected.

2. A non significant result means that the null hypothesis is true

    - Non significant result only tell use that the effect is not big enough to be found (given our sample size). Does not mean effect size is 0. As a sufficiently large sample size can make an infinitesimal effect size significant. so non significant result canot be interpreted as no difference between means or no relationship between variables.

3. A significant result means null hypothesis is false.

    - NHST is the result of trying to find a system that can test which of two competing hypotheses (the null or the alternative) is likely to be correct, it fails because the significance of the test provides no evidence about either hypothesis. It only shows which hypothesis is likely to be correct

2.2 All or Nothing thinking

Using the p value of .05 encourages the **all or nothing** thinking: if p<.05, then it is significant, if p > .05, then not significant.

However, this is ridiculous as sometimes, p value differ only very little yet people treat them as complete opposites (e.g. p = .0499 vs p = .0501, p value differ only by .0002)

e.g. if you do 10 studies, only 4/10 of the studies produce statistically significant result. All or nothing thinking will make you will say that the study is inconclusive. However if you look at the studies confidence intervals, even if some studies did not produce statistically significant result, they all are consistently positive. By looking at confidence intervals, we can have good reason that the effect may be greater than zero.

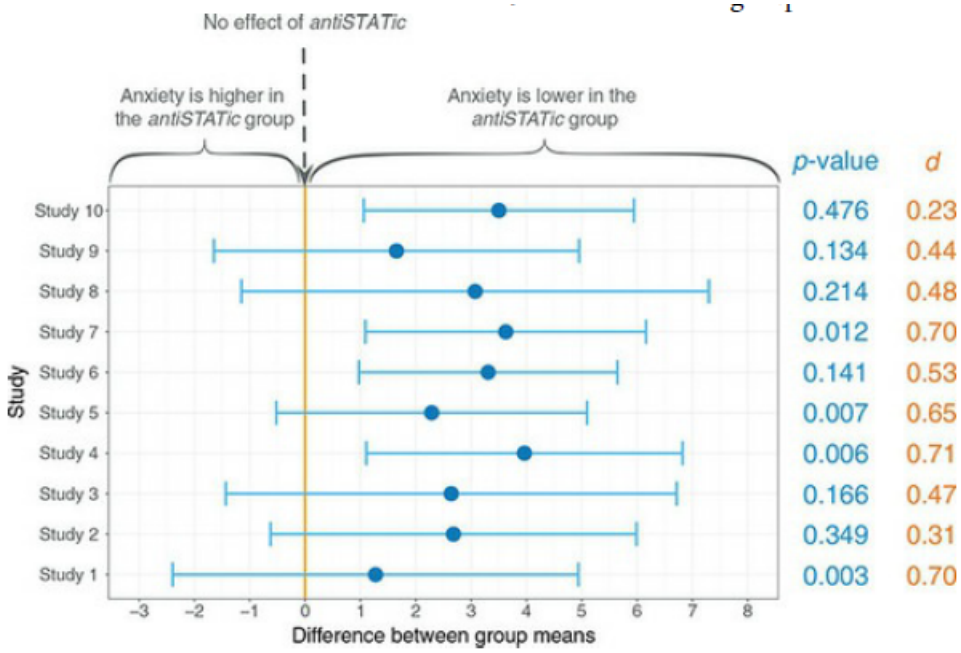

Figure 2.1: 4/10 are not statistically significant, but all the studies have positive difference between group means

2.3 NHST is influenced by the intentions of the scientists

NHST works on the principle that you make a Type I error in 5% in an infinite number of repeated identical experiments (long run probability). Both Type I and Type 2 errors are long run probability. It does not apply to individual studies. in an individual study, the probability is not .05 for alpha, it is either you have a Type I error (p = 1) or you don't (p=0).

p-value is the probability of getting a test statistics at least as large as the one observed relative to all possible values of t_{null} *from an infinite number of identical replications of the experiments*

p-value is the frequency of the observed test statistics relative to all possible values that could be observed in the collective of identical experiments, *with the exact same sampling procedure.*

E.g. You aim to collect 100 participants data, but only can find 93 participants. If you change your decision rule of computing p value based on 100 people to based on 93 people, result will change. i.e. if you compute p value based on df of 99 instead of df of 92, it is wrong. As you end up computing the relative frequency of the test statistics compared to all possible t_{null}s from experiments of size 93, but what you set out to do is to compare test statistic to all possible t_{null}s from experiments of size 100. So the space of possible t_{null}s has been influenced by an arbitrary variable of *availability of participants* rather than sticking to original scheme. The proper p-value you need to compute should be based on relative frequency of the observed test statistic compared to all possible t_{null}s from the collective of experiments where the intention was to collect 100 participants but (for the same reasons as in your experiment) only 93 participants were available. However, this p-value is too idiosyncratic to compute.

2.4 Incentive Structures and Publication Bias

Articles that have a significant finding are 7 times more likely to be published than non-significant ones. This is known as **publicatoin bias**, this bias is driven by reviewers rejecting non-significant result, and scientists not submitting articles with non significant results.

"Publish or Perish" mentality and also incentive structures in Science only award individuals who have successful studies. and "success" is largely defined by a study being signiificant

2.5 Researchers degrees of freedom

Researcher degrees of freedom refers to the fact that a scientist has many decisions to make when designing and analysis a study. e.g. alpha, power, sample size, which statistical model to fit, how to deal with extreme scores, what variables to consider, what measures to use.

Researchers degrees of freedom can be misused to exclude cases to make the result significant.

NHST nurtures these temptations by fostering black and white thinking, in which significant results garner much greater personal rewards than non-significant ones.

2.6 p-hacking and HARKing

p-hacking refers to researcher degrees of freedoms that lead to the selective reporting of significant p-values.

ways to p-hack:

1. deciding to stop collecting data at a point other than when the predetermined (prior to data collection) sample size is reached

2. including (or not) data based on the effect they have on the p-value.

3. including (or excluding) variables in an analysis based on how those variables affect the p-value

4. measuring multiple outcomes or predictor variables but reporting only those for which the effects are significant

5. merging groups of variables or scores to yield significant results, and transforming or otherwise manipulating, scores to yield significant p-values

HARKing refers to the practice in research articles of presenting a hypothesis that was made after data collection as though it were made before data collection.

In both cases of p-hacking and HARKing, you are not controlling the Type I error rate as you are deviating from the process that ensures that it is control, so Type 1 errors will definitely be more than 5%

2.7 Effect Sizes

Significance does not tell us about the importance of an effect. The solution to this criticism is to measure the size of the effect in a standardized way.

Effect size is an objective and standardized measure of the magnitude of observed effect. standardized means that we can compare effect sizes across different studies that have measured different variables, or have used different scales of measurement.

Most common effect sizes are: 1) Cohen's d, 2) Pearson's r, 3) odds ratio

2.7.1 Odds ratio

The odds of an event occuring are defined as the probability of an event occuring divided by the probability of an event not occuring.

odds $= \frac{\text{P(event}}{\text{P(no event)}}$

	Yes	No	Total
Singing	12	88	100
Conversation	26	74	100
Total	**38**	**162**	**200**

Table 2.1: data about effects of singing on saying yes to dates

e.g. We collected a data to undersrtand how much likely a person will say "yes" to a singer than to someone who starts a conversation.

$$\begin{aligned}\text{odds}_{\text{yes to a singer}} &= \frac{\text{Number of yes responses to a singer}}{\text{number of no responses to a singer}} \\ &= \frac{12}{88} = 0.14\end{aligned}$$

$$\begin{aligned}\text{odds}_{\text{yes to a talker}} &= \frac{\text{Number of yes responses to a talker}}{\text{number of no responses to a talker}} \\ &= \frac{26}{74} = 0.35\end{aligned}$$

$$\begin{aligned}\text{odds}_{\text{ratio}} &= \frac{\text{odds}_{\text{yes to a singer}}}{\text{odds}_{\text{yes to a talker}}} \\ &= \frac{.14}{.35} = .4\end{aligned}$$

2.8 Effect sizes compared to NHST

- Encourage interpreting effects on a continuum and not applying a categorical decision rule such as "significant" or "not significant".
- While effect sizes are affected by sample sizes (larger sample yield better estimates of population size effect size), but unlike p-value, there is no decision rule attached to effect sizes, so the interpretation of effect sizes is not confounded by sample size (although it is important in contextualising the degree to which an effect size might affect the populaiton). i.e. effect sizes are less affected than p-values by things like early or late termination of data collection, or sampling over a time period, rather than until a set sample size is reached.
- While researcher degree of freedom still exists in that researchers could maximize effect sizes, but there is less incentives to do so as effect sizes are not tied to a decision in which effects of either side of a certain threshold have qualitatively opposite interpretations.

Chapter 3

The Beast of Bias

3.1 What is bias

We oftern obtain values for parameters ina model using method of least squares. These parameter values in our sample estimate the parameter values in the population because we want to draw conclusions that extend beyond our sample. For each parameter in the model, we also compute an estimate of how well it represents the population such as standard error, or confidence interval. parameters can then be used to test hypothesis by converting them to a test statistic with an associated probability (p-value).

Statistical bias enters the process in 3 ways:

1. Things that bias the parameter estimates (including effect sizes);
2. Things that bias standard error and confidence intervals;
3. things that bias test statistics and p-values.

2 and 3 are linked by standard error, so if standard error is biased, then corresponding confidence interval and p-value will be biased too.

3.2 Outliers

Outliers have a big effect on mean. This in turn have a big effect on sum of squared errors. This is because any bias created by the outlier is magnified by the fact that the deviations are squared.

The effect of outliers on sum of squared errors is important as it is used to compute standard deviation, which is in turn used to estimate the standard error, which is used to calculate confidence intervals and test statistics.

3.3 Overview of Assumptions

An assumption is a condition that ensures that what you are attempting to do works. E.g. when we assess a model using a test statistic, we usually will have made some assumptions and if these assumptions are true then we know that we can take the test statistics and p-value at face value. If not, then they will be inaccurate

Many statistical models have idiosyncratic assumptions, but most are derived from the linear model and so they share a common set of assumptions.

Main assumptions to look at:

- additivity and linearity;
- normality of something or other;
- homoscedasticity/homogeneity of variance;
- independence.

3.3.1 Additivity and Linearity

Assumption of linearity and additivity means that relationship between outcome variable and predictors is accurately described by the equation:

$$\text{outcome}_i = (b_0 + b_1 X_{1i} + b_2 X_{2i}) + error_i \tag{3.1}$$

Assumption of linearity: This means that scores on the outcome variable are in reality linearly related to any predictors and that if you have several predictors then their combined effect is best described by adding their effects together.

So if relationship between variables is curvilinear, then describing it with a linear model is wrong. And there is no point interpreting its parameter estimates or worrying about significance tests.

3.3.2 Normally distributed something or other

Many people wrongly take the "assumption of normality to mean that the data need to be normally distributed. In fact, it relates in different ways to things we want to do when fitting models and assessing them.

1. **Parameter estimates**: The mean is a parameter, extreme scores bias it. This shows that estimates of parameters are affected by non-normal distributions (such as those with outliers). Parameter estimates differ by in how much they are biased in a non normal distribution: e.g. median is affected less by skewed distribution.

   All models include some error, wont predict outcome perfectly for each case. So for each case, there is an error term (deviance or residual). If residuals are normally distributed in the population, then using least squares to estimate parameters (the bs in equation 3.1) will produce better estimates than other methods.

2. **Confidence intervals**: For confidence intervals a parameter estimate (e.g. the mean or b in equation) to be accurate, the estimate must have a normal sampling distribution.

3. **Null Hypothesis Significance Testing (NHST)**: For significance tests of models (and the parameter estimates that define them) to be accurate, the sampling distribution of what's being tested must be normal. e.g. if testing for whether two means are different, the data themselves do not need to be normally distributed. However, the *sampling distribution of means* (or difference between means) must be normal.

   When looking at relationship between variables, significance tests of the parameter estimates that define the relationship (b in the equation) will only be accurate when sampling distribution of the estimate is normal.

3.4 Central Limit Theorem

Central Limit Theorem: regardless of the shape of the population, parameter estimates of that population will have a normal distribution if the samples are large enough.

3.4.1 When does the assumption of normality matter?

Central limit theorem means that there is a variety of situations in which we can assume normality regardless of the shape of our sample data.

1. For **confidence intervals around a parameter estimate** to be accurate (parameter estimate refers to the mean or b in equation), estimate must come from normal sampling distribution. But CLT tells us that estimate will come from a normal distribution regardless of what the sample or population data looks like. No need to worry about assumption of normality if *sample sizes large enough*

2. For **statistic models** to be accurate, sampling distribution must be normal. But CLT tells us no matter shape of data, if sample test large enough sampling distribution will be normal

3. For **estimates of model parameters** (b in equation) to be optimal using least squared method, residuals in the population must be normally distributed.

   method of least squares will always give you an estimate of the model parameters that minimizes error, so you no need to assume normality to fit a linear model and estimate the parameters that define it.

3.5 Homoscedasticity/Homogeneity of Variance

Homogeneity of variance affect two things:

1. **parameters**: using method of least squares to estimate parameters in the model, we get optimal estimates if the variance of the outcome variable is equal across different values of the predictor variable.

2. **NHST**: Test statistics often assume the variance of the outcome variable is equal across different values of the predictor variable. If this is not the case, then test statistics will be inaccurate.

3.5.1 Why does homogeneity of variance matter?

If we assume equality of variance, then the parameter estimates for a linear model are optimal using method of least squares.

Method of least squares will produce "unbiased" estimates of parameters even when homogeneity of variance can't be assumed, but they won't be optimal. If no homogeneity of variance, better estimates can be achieved using other methods, e.g. using weighted least squares in which each case is weighted by a function of its variance.

If all you care about is estimating the parameters of the model in your sample, then you don't need to worry about homogeneity of variance in most cases.

But unequal variance creates inconsistencies in the estimate of the standard error associated with the parameter estimates in the model. So confidence intervals,

significance tests (p-values) for parameter estimates will be biased, because they are computed using standard error. So if you want to test the significance of the model or its parameter estimates then homogeneity of variance matters.

3.6 Independence

Assumption of **Independence** means that errors in your model are not related to each other.

The equation that we use to estimate **standard error** is only valid when observations are independent.

If we use **method of least squares**, model parameter estimates will still be valid but not optimal (can get better estimates using different method, e.g. multilevel models).

Chapter 4

The Linear Model (Regression)

Equation of Linear regression:

$$Y_i = (\beta_0 + \beta_1 X_1 i) + \varepsilon_i \tag{4.1}$$

Linear Models are simply straight lines. All equation are forms of the equation of a straight line. All straight lines can be defined by two things. 1) slope of the line (usually denoted by b_1) and 2) intercept of the line b_0. In regression model, b_1 and b_0 are known as regression coefficients.

4.1 Linear Model with several predictors

For linear models, we can include as many predictor variables as we like:

$$Y_i = (\beta_0 + \beta_1 X_1 i + \beta_2 X_2 i) + \varepsilon_i \tag{4.2}$$

For two variables, our regression graph will not be a line but a plane with 2 dimensions.

4.2 Estimating the model

Model is usually not a perfect fit of the data. *Residuals*: differences between what the model predicts and the observed data.

We calculate the total error in a model by squaring differences between observed values of the outcome and predicted values that come from the model.

$$\text{total error} = \sum_{i=1}^{n} (\text{observed}_i - \text{model}_i)^2 \tag{4.3}$$

Similarly, to assess the error in a linear model, we use sum of squared errors. This error is called *sum of squared residuals* or residual sum of squares (SS_R)

In order to estimate the b-values in a linear model, we will find the line with the smallest SS_R because it would be the best fitting model. We use the method of least squares to estimate the parameter (b) which define the regression model for which the sum of squared errors is the minimum it can be given the data.

This method is known as the **ordinary least squares (OLS)** regression

4.3 Assessing goodness of fit, sums of squares, R and R^2

In order to assess whether our model is better than nothing, we need to compare the model against a baseline to see whether it "improves" how well we can predict the outcome. We can compare our model with the mean of the outcomes. As the mean of the outcomes is a model of "no relationship" between the variables: as one variable changes the prediction for the other remains constant. i.e. a horizontal line in a graph.

Using the mean of outcome as a baseline model, we can calculate the difference between observed values and the values predicted by the mean. We can squared these differences to give us the sum of squared difference also. We call this sum of squared difference as the **total sum of squares** (denoted by SS_T). It represents how good the mean is as a model of the observed outcomes scores.

$$\begin{aligned} SS_T &= (Y - M_Y)^2 \\ SS_R &= (Y - \hat{Y})^2 \end{aligned} \tag{4.4}$$

Where Y is your observed data, M_Y is the Mean of the observed data and $\hat{Y}$ is the predicted value using your model.

We can use SS_T and SS_R to calculate how much better the linear model is than the baseline model of no relationship. The improvement in prediction resulting from using the linear model than the mean is calculated as the difference between SS_T and SS_R. This improvement is known as the **model sum of squares** (SS_M).

$$SS_M = SS_T - SS_R \tag{4.5}$$

If SS_M is large, then the linear model is very different from using the mean to predict the outcome variable. This implies that the linear model has made a huge prediction in improvement to predicting outcome variable. if SS_M is small, then it is only slightly better than using the mean.

A useful measure arising from the sum of squares is the proportion of improvement due to the model. This is calculated by dividing the sum of squares for the model by the total sum of squares to give a quantity: R^2

$$\text{proportionate reduction in error}, R^2 = \frac{SS_M}{SS_T} \tag{4.6}$$

4.4 Using sum of squares in F test

We use F test to measure the amount of systematic variance divided by the amount of unsystematic variance. F here is based upon the ratio of improvement due to the model (SS_M) and the error in the model (SS_R), however we do not use SS_M or SS_R as their values depend on the number of observations were added up. So instead we use the average sum of squares known as **mean squares** or MS to compute F.

The mean sum of squares is the sum of squares divided by the associated degrees of freedom (similar to calculating variance from sum of squares).

$$\begin{aligned} MS_M &= \frac{SS_M}{k} \\ MS_R &= \frac{SS_R}{N - k - 1} \end{aligned} \tag{4.7}$$

Degrees of freedom for SS_M is number of predictors in the model (k), and for SS_R degree of freedom are the number of observation (N) minus number of parameter estimated (i.e. number of b coefficient including the constant). total number of b will $k+1$ if we include the intercept b_0.

F statistic is computed from these mean squares:

$$F = \frac{MS_M}{MS_R} \tag{4.8}$$

If model is a good model, F test will be larger, due to larger MS_M

F-statistics is also used to calculate the significance of R^2 using the following equation:

$$F = \frac{(N-k-1)R^2}{k(1-R^2} \tag{4.9}$$

where N is the number of participants, k is the number of predictors in the model. This F tests the null hypothesis that R^2 is zero (i.e. there is no improvement in the sum of squared error due to fitting the model).

4.5 Assessing individual predictors

A regression coefficient of 0 means (1) A unit change in the predictor variable results in no change in the predicted value of the outcome (predicted value of the outcome is constant). (2) The linear model is "flat" and horizontal.

If a variable significantly predicts an outcome, it should have a b value that is different from zero. This hypothesis is tested using a t statistics that tests the null hypothesis that the value of b is 0. What we are interested in the t test is whether the b we have is big compared to the amount of error in that estimate. Remember that the standard error for b tells us how different b values will be across different samples. If standard error is very small, then most samples are likely to have a b value that is similar to the one in oru sample (as there is little variation across samples).

$$t = \frac{b_{\text{observed}} - b_{\text{expected}}}{SE_h} = \frac{b_{\text{observed}}}{SE_h} \tag{4.10}$$

Null hypothesis b_{expected} is 0. You check your t statistics using the degree of freedom of $N-k-1$

Chapter 5

Comparing the two means: t-test

5.1 Looking at differences

We look at different means when there is a systematic manipulation of independent variables. (This is often done in a between-groups design).

Researchers get tempted to compare artificially created groups by dividing people into groups based on a median score; avoid doing that.

5.2 invisibility and mischief example

Testing the effect of invisibility has on the tendency for mischief. DV is mischief, IV is invisibility. Invisibility is a categorical variable, you either can be invisible or not.

5.3 Categorical predictors in a linear model

If we want to compare differences between the means of two groups, all we are doing is predicting an outcome based on memberships of two groups. This is a **linear model with one dichotomous predictor**. the b for the model reflects the differences between mean levels in the two different groups, and the resulting t-test, will tell us whether the difference between means is different from zero (t-test tests whether $b = 0$). When we are comparing means, we are using a special case of linear model.

Recall all statistical models are version of this idea:

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

When using a linear model the equation above becomes equation in which the model is defined by parameters: b_0 tells you the value of the outcome when the predictor is zero, b_1 quantifies the relationship between the predictor (X_i) and outcome (Y_i):

$$Y_i = (b_0 + b_1 X_{1i}) + \varepsilon_i$$

So in the invisibility and mischief example, the equation will look like this:

$$\text{Mischief}_i = (b_0 + b_1 \text{Cloak}_i) + \varepsilon_i$$

No cloak condition, knowing that they are in that group, the best prediction we could make of the number of mischievous acts would be the group mean because this value is the summary statistics with the least squared error. So the value of Y in the

equation will be the group mean of mischief under no cloak and the value of the cloak variable will be 0 (which is what we code for nominal variable). If we ignore the error term, the equation is:

$$\bar{X}_{\text{noCloak}} = b_0 + (b_1 * 0)$$
$$b_0 = \bar{X}_{\text{noCloak}}$$

Now to predict mischief in people who is invisible. Predicted value of the outcome will be the mean of the group to which the person belonged (cloak). Value of Cloak variable is 1. In the above we see that b_0 is equal to $\bar{X}_{\text{noCloak}}$ placing all the values in the equation, we get:

$$\bar{X}_{\text{Cloak}} = b_0 + (b_1 * 1)$$
$$\bar{X}_{\text{Cloak}} = b_0 + b_1$$
$$\bar{X}_{\text{Cloak}} = \bar{X}_{\text{noCloak}} + b_1$$
$$b_1 = \bar{X}_{\text{Cloak}} - \bar{X}_{\text{noCloak}}$$

So b_1 represents the difference between group means. In a model with a categorical predictor with two predictors, b_1 represents the difference between group means, b_0 represent the group mean of the group coded as 0.

Remember that t-statistic is used to ascertain whether a model parameter (b_1) is equal to 0; in this context, it would test whether the difference between group means is 0.

5.4 t-test

Two variants of t test:

- **Independent t-test**: used when comparing two means that come from conditions consisting of different entities.
- **Paired samples t-test**: or dependent t test, used when you want to compare two means that come from conditions consisting of the same or related entities

5.4.1 Rationale for the t-test

- Both t-tests, two samples of data are collected
- If samples come from the same population, we expect means to be roughly equal. Under null hypothesis, we assume that the experimental manipulation has no effect on the participants' behaviour: therefore, expect means of two samples to be similar.
- Compared difference between sample means we collected to the difference between sample means if null hypothesis was true. We use the standard error to gauge the variability between sample means. If standard error is small, then we expect most samples to have similar means.

  If standard error is large, large differences in sample means are more likely. If the difference between the samples we have collected is larger than we would expect based on the standard error, then one of two things has happened.

1. There is no effect but the sample means from our population fluctuate a lot and we happen to have collected two samples that produce very different means.
2. Two sample means come from the same population, so the difference is indicative of a genuine difference between samples. In other words, null hypothesis is unlikely.

- The larger the observed difference between sample means (relative to the standard error), the more likely it is that the second explanation is correct: that is, that the two sample means differ because of the different testing conditions imposed on each sample.

Remember that most test-statistics are a **signal-to-noise** ratio, "variance explained by the model"/"variance that model can't explain". The signal (the effect) here is the difference between the two group means. And the noise (the error) is the standard error, i.e. error in the estimate of the mean. So we can use the standard error between the two means as an estimate of the error in our model.

$$t = \frac{\text{observed difference between sample means - expected difference between population means if null is true}}{\text{standard error of the difference between two sample means}}$$

Top half of the equation is the model, which is that the difference between means is bigger than the expected difference under null hypothesis, which in most case is 0. Bottom half is error.

5.5 Assumptions of t-test

Both t-test and paired sample t-test are parametric tests and as such are prone to the sources of bias. For paired samples t-test the assumption of normality relates to the sampling distribution of difference scores, not the scores themselves.

Chapter 6

GLM 1: Comparing several independent means, ANOVA

Chapter 10: if we include predictor variable containing two categories into the linear model, then the resulting b for that predictor compares the difference between the mean score for the two categories.

Chapter 11: says that if we want to include a categorical predictor that contains more than two categories, this can be achieved by recoding that variable into several categorical predictors, each of which has only two categories (dummy coding). When we do this, bs for predictors represent differences between means.

Chapter 9: use F-statistics to test the overall fit of linear model, for ANOVA, we can do the same: we can use F to test whether we significantly predict the outcome variable by using group means.

For ANOVA, we can use an F to test whether we significantly predict the outcome variable by using group means (which tells us whether overall, the group means are significantly different) and then use the specific model parameters (bs) to tell us which means differ from which.

ANOVA is simply F-statistics which is use as a test of the fit of a linear model, its just that the linear model consists of group means.Learning ANOVA as a linear model framework allows ANOVA to be extended to more complex situations (e.g. multiple predictors, unequal group sizes) without the need to get bogged down in maths.

Another way to learn ANOVA is through use of F-statistics to compare means known as the variance-ratio method. This approach is fine for simple design analysis but is difficult to use in complex situations such as the analysis of covariance.

6.0.1 Example

	Control	15 minutes	30 minutes
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
$\bar{X}$	**2.20**	**3.20**	**5.00**
s	**1.30**	**1.30**	**1.58**
s^2	**1.70**	**1.70**	**2.50**
Grand mean = **3.467** Grand *SD* = **1.767**			
Grand variance = **3.124**			

Figure 6.1: Predicting happiness from exposure to dog therapy of different durations. 3 groups: Control, 15 mins, 30 mins

This can be incorporated into a linear model by including 2 dummy variable (each assigned a b-value).

When we assign dummy variable, one group must be the base like condition, b_0. In dog therapy example, we use the control group as the baseline category as we are interested in comparing both the 15 and 30 minutes groups to the control group. Thus the equation of the model is:

$$\text{Happiness}_i = b_0 + b_1\text{Long}_i + b_2\text{Short}_i + \varepsilon_i$$

By rearranging the variables, we can find out get the equation for $\bar{X}_{\text{control}}, \bar{X}_{30\text{ mins}}, \bar{X}_{15\text{ mins}}$, which are predicted value of happiness for each group.

$$\begin{aligned} \text{Happiness}_i &= b_0 + (b_1 * 0) + (b_2 * 0) \\ \bar{X}_{\text{control}} &= b_0 \end{aligned} \tag{6.1}$$

substitute b_0, into equation for 30mins:

$$\begin{aligned} \text{Happiness}_i &= b_0 + (b_1 * 1) + (b_2 * 0) \\ &= b_0 + b_1 \\ \bar{X}_{30\text{ mins}} &= \bar{X}_{\text{control}} + b_1 \\ b_1 &= \bar{X}_{30\text{ mins}} - \bar{X}_{\text{control}} \end{aligned} \tag{6.2}$$

substitute b_0, into equation for 15 mins:

$$\begin{aligned} \text{Happiness}_i &= b_0 + (b_1 * 0) + (b_2 * 1) \\ &= b_0 + b_1 \\ \bar{X}_{15\text{ mins}} &= \bar{X}_{\text{control}} + b_2 \\ b_2 &= \bar{X}_{15\text{ mins}} - \bar{X}_{\text{control}} \end{aligned} \tag{6.3}$$

Equation 5.3 shows that the b value for dummy variable representing the 15 minute group is the difference between the means for the 15 minute group and control group.

By coding a new dummy variable for each of the variable, we can use linear regression and F test to test the overall fit of the model.

6.1 Logic of F statistics

F test is an overall test that does not specify differences between specific means. However, the model parameters (b values) do.

F-statistic tests the overall fit of a linear model to a set of observed data. F is the ratio of how good the model is compared to how bad it is (its error). When model is based on group means, our predictions from the model are the means. If group means are the same, then our ability to predict observed data will be poor (F will be small), bt if the means differ, we will be able to better discriminate between cases from different groups (F will be large). In the dog therapy context, F basically tells us whether the group means are significantly different.

From above we know that $b_1 = \bar{X}_{30 \text{ mins}} - \bar{X}_{\text{control}}$ and $b_2 = \bar{X}_{15 \text{ mins}} - \bar{X}_{\text{control}}$. However, if null hypothesis is true, and all the groups have the same group mean, then these b coefficients will be 0 (if group means are equal then differences between them will be 0).

We can apply same logic as for any linear model:

- The model that represents 'no effect' or 'no relationship between the predictor variable and the outcome' is one where the predicted value of the outcome is always the grand mean (the mean of the outcome variable).
- We can fit a different model to the data that represents our alternative hypotheses. We compare the fit of this model to the fit of the null model (i.e., using the grand mean).
- The intercept and one or more parameters (b) describe the model.
- The parameters determine the shape of the model that we have fitted; therefore, the bigger the coefficients, the greater the deviation between the model and the null model (grand mean).
- In experimental research the parameters (b) represent the differences between group means. The bigger the differences between group means, the greater the difference between the model and the null model (grand mean).
- If the differences between group means are large enough, then the resulting model will be a better fit to the data than the null model (grand mean).
- If this is the case we can infer that our model (i.e., predicting scores from the group means) is better than not using a model (i.e., predicting scores from the grand mean). Put another way, our group means are significantly different from the null (that all means are the same).

We use F statistics to compare the improvement in fit due to using model (rather than the null, grand mean model). F statistics is the ratio of the explained to unexplained variation. We calculate this variation using sum of squares: $R^2 = \frac{SS_{total} - SSerror}{SS_total}$

Recall that we can examine the extent to which a model deviates from the observed data using the general form of:

$$\text{Total error} = \sum_{i-1}^{n} (\text{observed}_i - \text{model}_i)^2$$

6.1.1 Total sum of squares SS_T

To find the total variation within our data, we calculate the difference between each observed data point (regardless of which group the data point is in) and the grand mean:

$$\mathrm{SS}_T = \sum_{i-1}^{N}(x_i - \bar{x}_{\mathrm{grand}})^2$$

SS_T is derived from the variance formula of $s^2 = \frac{SS}{N-1}$. Therefore, we can calculate the SS_T from the variance of all observations (grand variance). Grand variance is the variation between all scores, regardless of which group the scores comes from. i.e. take the mean of the total observation and do $\frac{(X-M)^2}{N-1}$

$$\mathrm{SS}_T = s^2_{grand}(N-1) \tag{6.4}$$

6.2 Model sum of squares (SS_M)

Model sum of squares tells us how much of the total variation in the outcome can be explained by the fact that different scores come from entities in different treatment conditions. Model sum of squares is calculated by taking the difference between values predicted by the model and the grand mean.

When making predictions from group membership, the values predicted by the model are the group means. From figure, it shows that model sum of squared error is the sum of squared distances between what the model predicts for each data point. It is the difference between the mean of the group to which the scores belongs, represented by the different horizontal lines and the grand mean (represented by the red line)

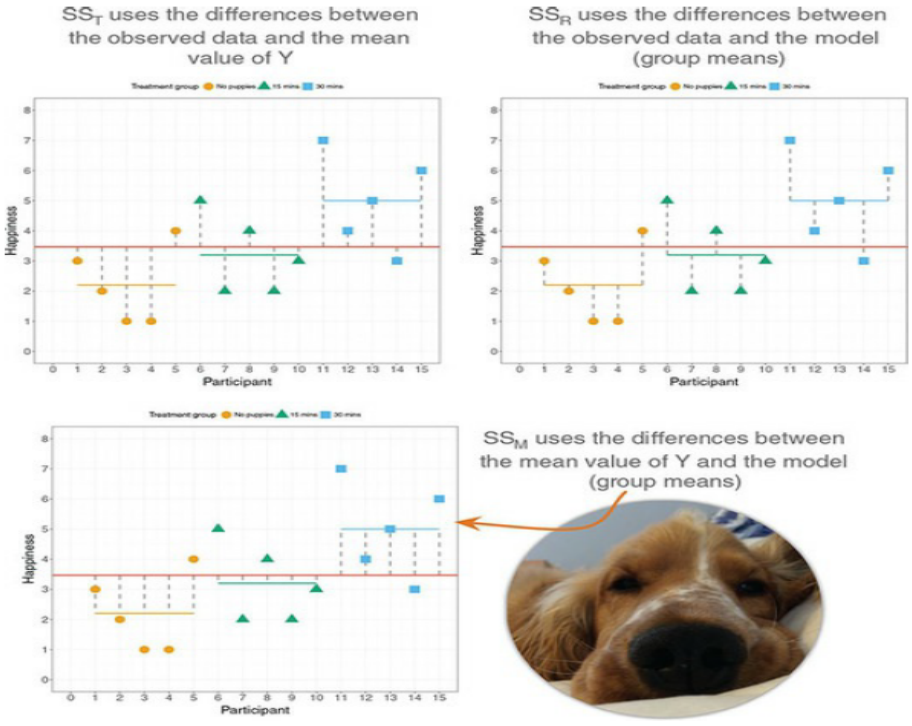

Figure 6.2: graphical representation of SS_T, SS_R, SS_M. Experiment has 3 different groups, each with sample size of 5. Red line represent grand mean, shorter coloured lines represent group mean in the samples.

Equation of SS_M:

$$\mathrm{SS}_M = \sum_{g=1}^{k} n_g(\bar{x}_g - \bar{x}_{grand})^2 \tag{6.5}$$

Equation is saying:

1. calculate the difference between the mean of each group ($\bar{x}_g$) and the grand mean ($\bar{x}_{grand}$).
2. Square each of the differences
3. Multiply each result by number of participants within the group (n_g)
4. Add the values for each group together

For SS_M the **degrees of freedom** (df_M) is the number of groups - 1, denoted by $(k-1)$.

6.3 Residual sum of squares (SS_R)

Residual sum of squares (SS_R) tells you how much variation cannot be explained by the model. This variation is caused by things we haven't measured such as measurement error, random noise and individual differences that may affect the DV other than our IV. Simplest way to calculate SS_R is:

$$SS_R = SS_T - SS_M$$

However this provides little insight to what SS_R represents.

We know that our model predicts the **mean of the group to which the person belongs**. Therefore, SS_R is calculated by looking at the difference between the score obtained by the a person and the mean of the group to which the person belongs. In graphical terms, the dotted vertical lines in the figure represent difference. The distances between each data point and the group mean are squared and added together to give the residual sum of squares, SS_R:

$$SS_R = \sum_{g=1}^{k}\sum_{i=1}^{n}(x_{ig} - \bar{x}_g)^2 \tag{6.6}$$

We can also express SS_R as:

$$SS_R = SS_{group1} + SS_{group2} + ... + SS_{groupk} \tag{6.7}$$

where k is the number of groups.

Given that variance is the sum of squares divided by n -1, we can express SS_R like this:

$$SS_R = \sum_{g=1}^{k} s_g^2(n_g - 1) \tag{6.8}$$

This means that you multiply the variance for each group (s_g^2) by number of people in the group minus 1 ($n_g - 1$, i.e. the df_{within}).

The degrees of freedom for SS_R (df_R) is the total degrees of freedom minus the degrees of freedom for the model.

$$df_R = df_T - df_M \tag{6.9}$$

This is equivalent to N - k, where N is total participants, and k is the number of groups.

6.4 Mean squares

SS_M tells you the total variation that the model explains (i.e. variation due to experimental manipulation), SS_R tells you the total variation that is due to unmeasured factors. However, both SS are sum values,their size then depends on the number of scores. Furthermore SS_M uses the sum of 3 values (group means) while SS_R and SS_R uses the sum of 15 values. We need to use MS instead in order to eliminate the discrepancies in size.

However rather than dividing by the number of scores for each SS, we divide by the df as we are trying to extrapolate to a population and so some parameters within that population will be held constant

$$\begin{aligned} MS_M &= \frac{SS_M}{df_M} \\ MS_R &= \frac{SS_R}{df_R} \end{aligned} \tag{6.10}$$

6.5 F statistic

F statistic is the measure of the ratio of variation explained by the model and the variation attributable to unsystematic factors. It is calculated by:

$$F = \frac{MS_M}{MS_R} \tag{6.11}$$

If F is less than 1 it means that MS_R is greater than MS_M and that there is more unsystematic variance than systematic variance. This means that our experimental manipulation is unsuccessful and F will be non-significant.

6.6 Interpreting F

F is an **omnibus** test. A significant F tells us that the different groups are not equal but then, it does not tell you whether all of them are different or only some of them are different.

However if we want to see whether specific group means differ, we cannot just fit lots of model each of which compares only two means at a time; This is inflate the type 1 error rate.

The reason why F-test is useful is that as a single test, it controls Type I error rate. Having established that overall group means differ (i.e. the outcome can be significantly predicted using the group means), we can use the parameters of the model (b-values) to tell us where the differences lie.

6.7 Assumptions when comparing means

Given that we are still using a linear model, all assumptions that we have when doing normal hypothesis testing also applies. Normality is tested within groups, not across entire samples.

6.7.1 Homogeneity of variance

We assume that variance of the outcome is steady as the predictor changes, (i.e. variances in the groups are equal). If group sizes are unequal, violations of the assumption of homogeneity of variance can have serious consequence. F test can be corrected to account for deviations from homogeneity.

1. Brown-Forsythe F

2. Welsch's F

6.7.2 Brown-Forsythe F

F is conservative when group sizes are unequal and the large groups have the biggest variance. If sample size is too large, it will inflate value of SS_R as it is calculated by variance multiplied by sample size -1. If SS_R is big, F will be smaller.

Brown Forsythe got around this problem by weighting the group variances not by their sample size, but by the inverse of their sample size (they use $\frac{n}{N}$ which is the sample as the sample size as a proportion of the total sample size. This adjustment reduces the impact of large sample sizes with large variances.

$$\begin{aligned} F_{BF} &= \frac{SS_M}{SS_{R_{BF}}} = \frac{SS_M}{\sum s_k^2(1-\frac{n_k}{N})} \\ &= \frac{SS_M}{s^2_{group1}(1-\frac{n_{group1}}{N}) + s^2_{group2}(1-\frac{n_{group2}}{N}) + s^2_{groupn}(1-\frac{n_{groupn}}{N})} \end{aligned} \tag{6.12}$$

6.8 Is Anova Robust?

Robust means that even if we break assumptions, the test will still be accurate. Questions to ask:

1. Does F control the Type I error rate or is it significant even when there are no differences between means?

2. Does F have enough power to detect differences when they are there?

Recent simulations shows that differences in skewness, non-normality and heteroscedasticity of variance interact in complicated ways that impact power. In the absence of normality, violations of homoscedasticity will affect F even when group sizes are equal.

F can only be considered robust if **group distributions are identical**; e.g. all the groups if they are not normal, must be skewed in an identical way.

Violations of normality: Heavy tailed distributions are problematic: if you set up a situation with power of 0.9 to detect an effect from a normal distribution and contaminate the distribution with 10% scores from a normal distribution with a heavier tail, power drops to 0.28. Heavy tailed distribution would need a greater sample size in order for central limit theorem to work (160 instead of usual 30).

Violations of Assumptions of independence is also very serious: if three groups of 10 observation per group are correlated at r = .5 , Type I error rate is 0.74 (when we thought it to be 0.05).

6.9 What to do when assumptions are violated?

If you routinely use Welch's F, then you no need to think about homogenity of variance, you can bootstrap parameter estimates which wont affect F itself and you know at least model parameters are robust.

If you can also use a robust test in all situation and not worry about the assumptions.

If you use the usual F statistic, at least conduct a sensitivity analysis, apply a robust test to check that your conclusion doesn't change.

6.10 Planned contrast (contrast coding)

When we are using t test to check whether b is significantly from 0 (b is the differences between means), the trouble comes when we have 2 dummy variables and we need to do 2 t-tests, this will inflate the familywise error rate.

Another problem is that dummy variables might not make all the comparisons we want to make. (in our example it doesnt compare 15 and 30 mins).

Solution is to use contrast coding rather than dummy coding. Contrast coding is a way of assigning weights to groups in dummy variables to carry out planned contrasts (known as planned comparisons). Weights are assigned in such a way that the contrasts are independent, which means Type I error rate is controlled.

Another solution is to do a post hoc test: compare every mean group to all others (to conduct several overlapping tests using a t-statistic each time) but using a stricter acceptance criterion that keeps family wise error rate at .05.

6.11 Choosing which contrast to use

Chapter 7

GLM 4: Repeated Measures Designs

7.1 Example

Celebrity	Stick Insect	Kangaroo Testicle	Fish Eye	Witchetty Grub	Mean	s^2
1	8	7	1	6	5.50	9.67
2	9	5	2	5	5.25	8.25
3	6	2	3	8	4.75	7.58
4	5	3	1	9	4.50	11.67
5	8	4	5	8	6.25	4.25
6	7	5	6	7	6.25	0.92
7	10	2	7	2	5.25	15.58
8	12	6	8	1	6.75	20.92
Mean	**8.13**	**4.25**	**4.13**	**5.75**		

Figure 7.1: Sample size of 8, each participant goes through 4 different conditions.

For this the equation is:

$$\begin{aligned} \text{Retch}_{gi} &= b_{0i} + b_{1i}X_{gi} + \varepsilon_{gi} \\ b_{0i} &= b_0 + u_{0i} \\ b_{1i} &= b_0 + u_{1i} \end{aligned} \tag{7.1}$$

Subscript i is used to denote each individual. And the different predictors is denoted by g. So we can predict retch time (Y_{gi}) for food g within person i from the specific food eaten (X_{gi}).

However, we need to factor in that individuals will vary in their constitution. We can do this by adding a variance term to the intercept. Intercept represents time to wretch when predictor is 0; so if we allow this parameter b_{0i} to vary across individuals, we are effectively modelling the possibility that different people will have different retching latency. We define b_{0i} as made up of the group level intercept (b_0) plus the deviation of the individual's intercept from the group-level intercept (u_{0i}). So u_{0i} is the individual differences in retching.

Similarly, for the slope, we also have to account for individual's deviation of the group slope. So for b_1, we also split the variables into b_1 and u_{1i}. u_i reflects individual's differences in the effect of food on retching.

First line of equation (7.1) models the individual. Second and third line models the group level effects.

So we assume that the variation within conditions is similar and that no two conditions are any more dependent than any other two.

7.2 Assumption of sphericity

Sphericity is about assuming that the relationship between scores in pairs of treatment conditions is similar (i.e. level of dependence between means is roughly equal).

The assumption of sphericity (denoted by ε and sometimes referred to as circularity) can be likened to the assumption of homogeneity of variance in between-group designs. It is a form of compound symmetry, which holds true when both the variances across conditions are equal. and the covariances between pairs of conditions are equal.

Sphericity is a more general, less restrictive form of compound symmetry and refers to the equality of variances of the differences between treatment levels

Condition A	Condition B	Condition C	A - B	A - C	B - C
10	12	8	-2	2	4
15	15	12	0	3	3
25	30	20	-5	5	10
35	30	28	5	7	2
30	27	20	3	10	7
		Variance:	**15.7**	**10.3**	**10.7**

Figure 7.2: 3 different conditions and differences of each conditions

Sphericity will hold when:

$$\text{variance}_{A-B} \approx \text{variance}_{A-C} \approx \text{variance}_{B-C}$$

In this example, there is some deviation from sphericity in the data as variance of A-B is greater than the rest. However, the data has local circularity (local sphericity) as two of the variances of differences are very similar, which means that sphericity can be assumed for any multiple comparisons involving these conditions.

But how do we know whether the difference in variance of differences is large enough to be a problem?

7.3 Assessing the severity of departures from sphericity

We can use Mauchly's test to assess the hypothesis that the variances of the differences between conditions are equal. If Mauchly's test statistic is significant, it implies that there are significant differences between the variances of differences and,therefore, sphericity is not met. If Mauchly's test is non significant ($p > .05$), then it means variances of differences are roughly equal and sphericity is met.

Downside of Mauchly test is that it is a significance test so it depends on sample size. So in large sample size, Mauchly test could mean just a small departure from

sphericity that we have a lot of power to detect, and in small sample size, Mauchly test can mean a large departure from sphericity but we do not have power to detect. So either ways we always just use the Greenhouse-Geisser correction

Instead, we can estimate the degree of sphericity using Greenhouse-Geisser estimate $\hat{\varepsilon}$.

The degree to which sphericity is present, or not, is represented by a statistic called epsilon ($\hat{\varepsilon}$). An epsilon of 1 (i.e., $\hat{\varepsilon} = 1$) indicates that the condition of sphericity is exactly met. The further epsilon decreases below 1 (i.e., $\hat{\varepsilon} < 1$), the greater the violation of sphericity. Therefore, you can think of epsilon as a statistic that describes the degree to which sphericity has been violated. The lowest value that $\hat{\varepsilon}$ can take is called the lower-bound estimate, upper limit is 1.

Greenhouse-Geisser estimate varies between 1/(k-1) and 1, where k is the number of repeated conditions. e.g. if there are 5 conditions, the lower limit of $\hat{\varepsilon}$ is 1/(5-1) = 0.25, and upper limit is 1.

7.4 Effect of violating the assumption of sphericity

Sphericity creates a loss of power and an F-statistic that does not have the distribution that it's supposed to have (i.e. an F-distribution).

Lack of sphericity causes some complications for post hoc tests. If you dont want to worry about what these complications are when sphericity is violated, then the Bonferroni method is the most robust in terms of power and control of Type I error rate.

When sphericity is not violated, Tukey's test can be used.

Note: sphericity is not relevant if you are comparing only two means. The assumption is that the variances of difference scores between pairs of treatment levels are equal, and with only two conditions, you have only one set of difference scores and so only 1 variance. You need at least 3 conditions for sphericity to be an issue.

7.5 What to do if you violate sphericity?

If you violate sphericity, just adjust the degrees of freedom of any F-statistics affected. You multiply the degrees of freedom for an affected F by this estimate of sphericity. The result is that when you have sphericity, the degrees of freedom don't change as you are multiplying by 1. But if you dont have sphericity, your degrees of freedom will be smaller (as you are multiplying by less than 1).

Greater violation of sphericity, smaller the estimate, smaller the degrees of freedom. Smaller df make the p value associated with F-statistic less significant. By adjusting degrees of freedom, we make F-statistics more conservative, and so Type I error is controlled.

7.6 F statistics for repeated measures designs

In repeated measures design, the effect of the experiment (Independent variable) is shown up in within participant variance (rather than in the between-group variance).

In between subject design, the within participant variance is the residual sum of squares (SS_R); variance is created by individual differences in performance.

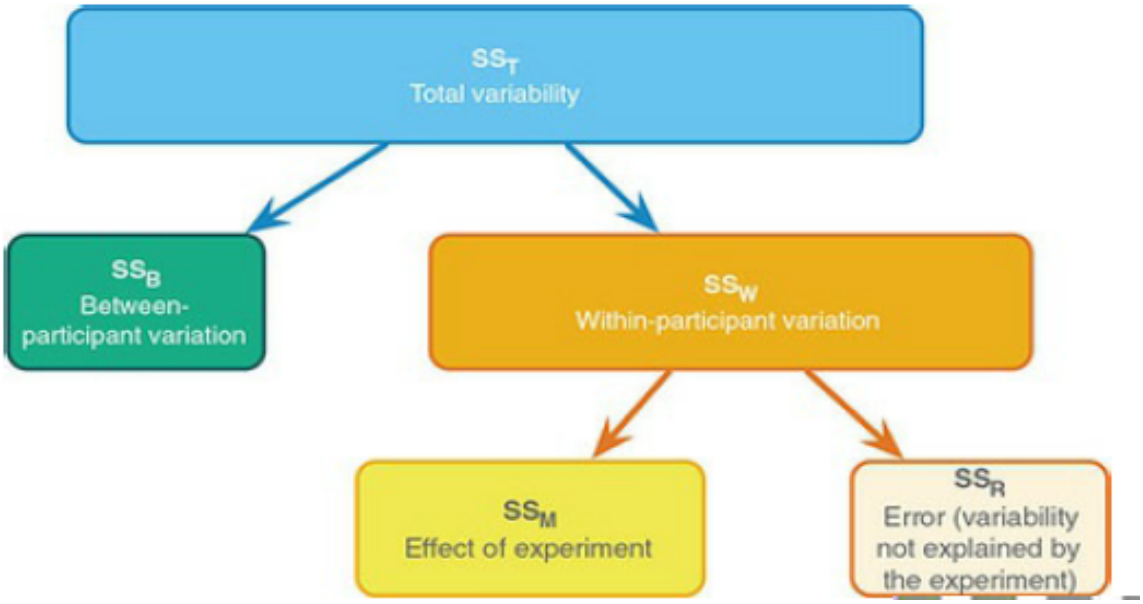

Figure 7.3: SS_M and SS_R both comes from within participant variation in repeated measures design

However for repeated measures design, experimental manipulation is carried on the same entities, within participant variance will be made up of not just individual differences in the performance but also effect of the manipulation.

7.7 Total sum of squares, SS_T

Similar to one way independent design, SS_T is calculated as:

$$\mathrm{SS}_T = s^2_{grand}(N-1) \tag{7.2}$$

df for SS_T is also N - 1

7.8 Within participant sum of squares,SS_W

For SS_R in a repeated measures design, the most crucial difference is that there is a within participant variance component, which represents individual differences within participants.

In independent design, there are different participants within each condition, we calculated SS_R within each condition and added these to get a total value.These individual differences were quantified with residual sum of squares (SS_R) using the equation:

For independent design :

$$\begin{aligned} SS_R &= \sum_{g=1}^{k}\sum_{i=1}^{n}(x_{ig}-\bar{x}_g)^2 \\ &= \sum_{g=1}^{k} s^2_g(n_g-1) \\ &= s^2_{group1}(n_1-1) + s^2_{group2}(n_2-1) + ... + s^2_{groupn}(n_n-1) \end{aligned} \tag{7.3}$$

However, in repeated measures design, we subjected entities to more than one experimental condition, we are interested in the variation not within a condition but *within an entity*. We can adapt similar equation to look within participants rather than groups. We will call this equation SS_W (for within participant SS).

$$\textbf{For repeated measures design}: \\ SS_W = s^2_{entity1}(n_1 - 1) + s^2_{entity2}(n_2 - 1) + ... + s^2_{entityn}(n_n - 1) \tag{7.4}$$

In our celebrity example it will be:

$$\begin{aligned} SS_W &= 9.67(4-1) + 8.25(4-1) + 7.58(4-1) + 11.67(4-1) + 4.25(4-1) + 0.92(4-1) + \\ &\quad 15.58(4-1) + 20.92(4-1) \\ &= 236.50 \end{aligned}$$

df for each entity is n-1 (i.e. number of conditions minus 1). Total *df* is just the sum of df of each participants. e.g. 8 participants with 3 df each, there will be 24 df in total.

7.9 Model sum of squares, SS_M

Note that SS_M for within participant design is part of SS_W. For independent design, we worked out how much variation could be explained by our experiment (model SS) by looking at the means for each group and comparing these to the overall mean. We do the same in repeated measures design:

$$\text{SS}_M = \sum_{g=1}^{k} n_g(\bar{x}_g - \bar{x}_{grand})^2 \tag{7.5}$$

$\bar{x}_g$ is the group mean of each condition, $\bar{x}_{grand}$ is the grand mean of all condition. n_g is the number of participants in each condition.

In our example, SS_M is:

$$\begin{aligned} SS_M &= 8(8.13 - 5.56)^2 + 8(4.25 - 5.56)^2 + 8(4.13 - 5.56)^2 + 8(5.75 - 5.56)^2 \\ &= 83.13 \end{aligned}$$

df for SS_M is the number of conditions (k) - 1 (similar to independent design).

7.10 Residual sum of squares, SS_R

simplest way is to subtract SS_M from SS_W.

$$SS_R = SS_W - SS_M \tag{7.6}$$

In our example, SS_R = 236.50 - 83.13 = 153.37 similarly df for SS_R is calculated in similar way:

$$df_R = df_W - df_M \tag{7.7}$$

In our example, df is 24-3= 21.

7.11 Mean Squares and F-statistics

$$MS_M = \frac{SS_M}{df_M}$$
$$MS_R = \frac{SS_R}{df_R} \tag{7.8}$$

$$F = \frac{MS_M}{MS_R} \tag{7.9}$$

In our example: $MS_M = \frac{83.13}{3} = 27.71, MS_R = \frac{153.37}{21} = 7.30$
$F = \frac{27.71}{7.30} = 3.79$

7.12 Between Participant sum of squares, SS_B

Just need to subtract within participant sum of squares with total sum of squares

$$SS_B = SS_T - SS_W \tag{7.10}$$

Chapter 8

GLM 3: Factorial Designs

Factorial designs have two or more predictors (independent variables).

Types of Factorial Design:

1. Independent Factorial Design: Several Independent Variables or predictors each has been measured using different entities (between groups) (discussed in this chapter)

2. Repeated-Measures factorial design: Several independent variables or predictors have been measured, but the same entities have been used in all conditions.

3. Mixed design: Several independent variables or predictors have been measured; some have been measured with different entities, others used the same entities.

8.1 Example

Alcohol	Placebo		Low dose		High dose	
FaceType	Attractive	Unattractive	Attractive	Unattractive	Attractive	Unattractive
	6	2	7	3	5	5
	7	4	6	5	6	6
	6	3	8	7	7	8
	7	3	7	5	5	6
	6	4	6	4	7	7
	5	6	7	4	6	8
	8	5	6	5	5	7
	6	1	5	6	8	6
Total	51	28	52	39	49	53
Mean	6.375	3.500	6.500	4.875	6.125	6.625
Variance	0.839	2.571	0.857	1.554	1.268	1.125
	Grand mean = 5.667					
	Grand variance = 2.525					

Figure 8.1: Factorial design: 2 independent variable (Face type and dosage)

General Linear Model takes the form:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + ... + b_n X_{ni} + \varepsilon_i \tag{8.1}$$

Just like the puppy therapy example, we can code participant's category membership on these variables with 0 and 1. However, in the alcohol and face attractiveness

example, our linear equation we also need to consider about interaction effects. so the model will become:

$$\begin{aligned} \text{Attractiveness}_i &= b_0 + b_1 A_i + b_2 B_i + b_3 AB_i + \varepsilon_i \\ &= b_0 + b_1 \text{FaceType}_i + b_2 \text{Alcohol}_i + b_n \text{Interaction}_i + \varepsilon_i \end{aligned} \tag{8.2}$$

Type of faces rated	Alcohol	Dummy (FaceType)	Dummy (Alcohol)	Interaction	Mean
Unattractive	Placebo	0	0	0	3.500
Unattractive	High dose	0	1	0	6.625
Attractive	Placebo	1	0	0	6.375
Attractive	High dose	1	1	1	6.125

Figure 8.2: Coding scheme for the Linear Model

You code the interaction variable by multiplying the face dummy variable with the alcohol dummy variable. e.g. for someone receiving high dose of alcohol and rating unattractive faces, their dummy variables would be Facetype = 0, Alcohol = 0, Interaction = 0 x 0 = 0.

Similar to independent design, b_0 represents **mean of the group for which all variables are coded as 0**, i.e. it is the mean value of the baseline category (in our example, it is the placebo group rating unattractive faces). Similar to independent design, the predicted value of the outcome is our group mean.

$$\begin{aligned} \text{Attractiveness}_i &= b_0 + b_1 A_i + b_2 B_i + b_3 AB_i + \varepsilon_i \\ \bar{X}_{\text{unattractive, placebo}} &= b_0 + (b_1 \times 0) + (b_2 \times 0) + (b_3 \times 0) \\ b_0 &= \bar{X}_{\text{unattractive, placebo}} \\ &= 3.5 \end{aligned} \tag{8.3}$$

For participants in placebo group rating attractive faces, b_1 represent difference between ratings of attractive and unattractive face in placebo condition. Or we can say it is the effect of type of face for baseline category of alcohol. It is model in the equation like this:

$$\begin{aligned} \bar{X}_{\text{attractive, placebo}} &= b_0 + (b_1 \times 1) + (b_2 \times 0) + (b_3 \times 0) \\ &= b_0 + b_1 \\ &= \bar{X}_{\text{unattractive, placebo}} + b_1 \\ b_1 &= \bar{X}_{\text{attractive, placebo}} - \bar{X}_{\text{unattractive, placebo}} \\ &= 6.375 - 3.5 \\ &= 2.875 \end{aligned} \tag{8.4}$$

For participants in high dose of alcohol and rating unattractive faces. b_2 shows difference between ratings of unattractive faces after high alcohol dose. It is the effect of alcohol on baseline category type of face (i.e. Face type coded with 0). The model becomes:

$$\begin{aligned}
\bar{X}_{\text{unattractive, high dose}} &= b_0 + (b_1 \times 0) + (b_2 \times 1) + (b_3 \times 0) \\
&= b_0 + b_2 \\
&= \bar{X}_{\text{unattractive, placebo}} + b_2 \\
b_2 &= \bar{X}_{\text{unattractive, high dose}} - \bar{X}_{\text{unattractive, placebo}} \\
&= 6.625 - 3.5 \\
&= 3.125
\end{aligned} \tag{8.5}$$

For participants rating attractive faces after high dose of alcohol. We replace b_0, b_1, b_2 to get the equation:

$$\begin{aligned}
\bar{X}_{\text{attractive, high dose}} &= b_0 + (b_1 \times 1) + (b_2 \times 1) + (b_3 \times 1) \\
&= b_0 + b_1 + b_2 + b_3 \\
&= \bar{X}_{\text{unattractive, placebo}} + (\bar{X}_{\text{attractive, placebo}} - \bar{X}_{\text{unattractive, placebo}}) + \\
&(\bar{X}_{\text{unattractive, high dose}} - \bar{X}_{\text{unattractive, placebo}}) + b_3 \\
b_3 &= \bar{X}_{\text{unattractive, placebo}} - \bar{X}_{\text{attractive, placebo}} + \bar{X}_{\text{attractive, high dose}} - \bar{X}_{\text{unattractive, high dose}} \\
&= 3.5 - 6.375 + 6.125 - 6.625 \\
&= -3.375
\end{aligned} \tag{8.6}$$

b_3 compares the difference between ratings of unattractive and attractive faces in the placebo group to the same difference in the high dose group. More generally, it compares the effect of type of face after a placebo drink to the effect of type of face after a high dose of alcohol.

If you rearrange the terms in the equation you can phrase the interaction in the opposite way: it represents effect of alcohol on ratings of attractiveness for attractive faces compared to unattractive ones.

Below is the plot of the interaction graph: Difference between ratings of unattractive and attractive faces in the placebo group is 2.875. Difference between unattractive and attractive faces in high dose group is -0.500. If we plotted the difference values in the new graph, we will get the bottom left graph. The slope of the bottom left graph will be the value of b_3 (-.5 - -2.875 = -3.375)

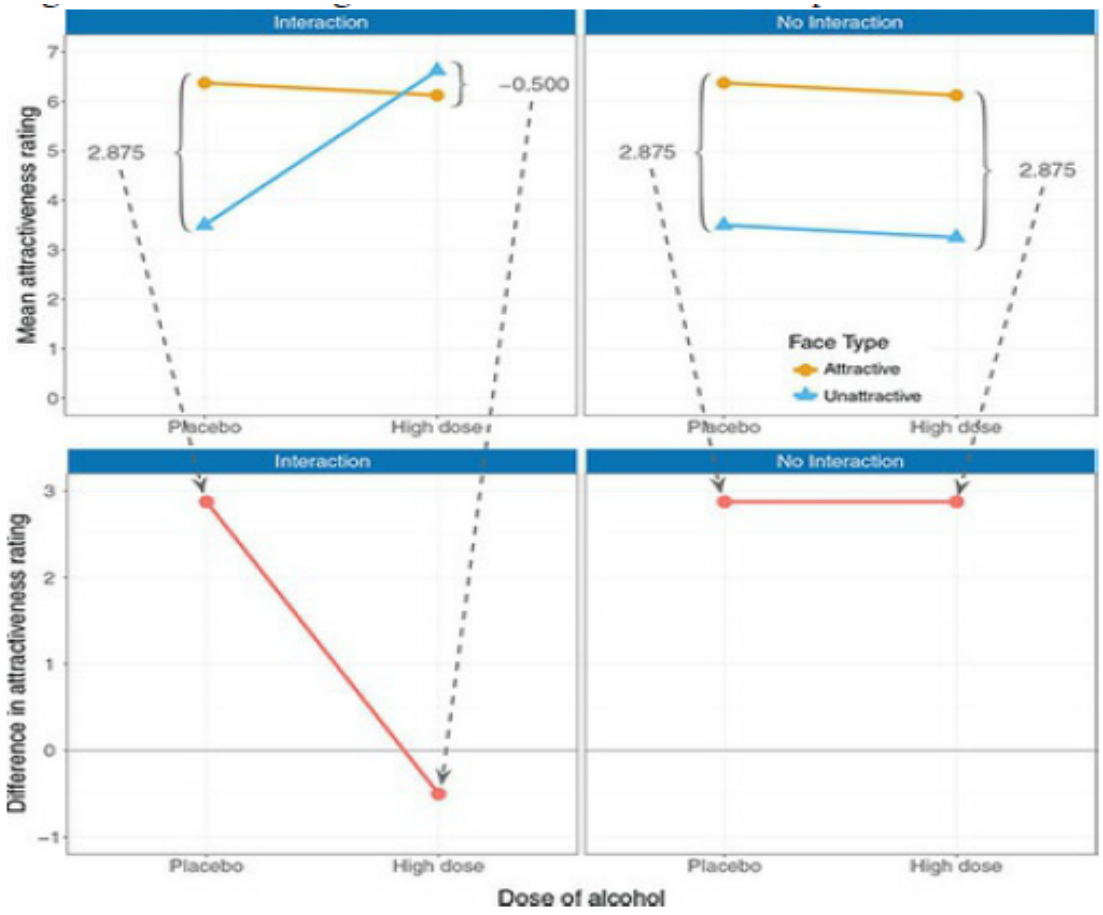

Figure 8.3: Interaction graph that plot the 4 different means. Top left is the actual graph, top right is an example showing if there is no interaction effect (differences between attractiveness treatment group is the same)

8.2 Behind the scenes of factorial designs

Calculating F statistics for factorial design is similar in calculation as for independent design. Just that for SS_M it is further subdivided to SS_A, SS_B and $SS_{A\times B}$

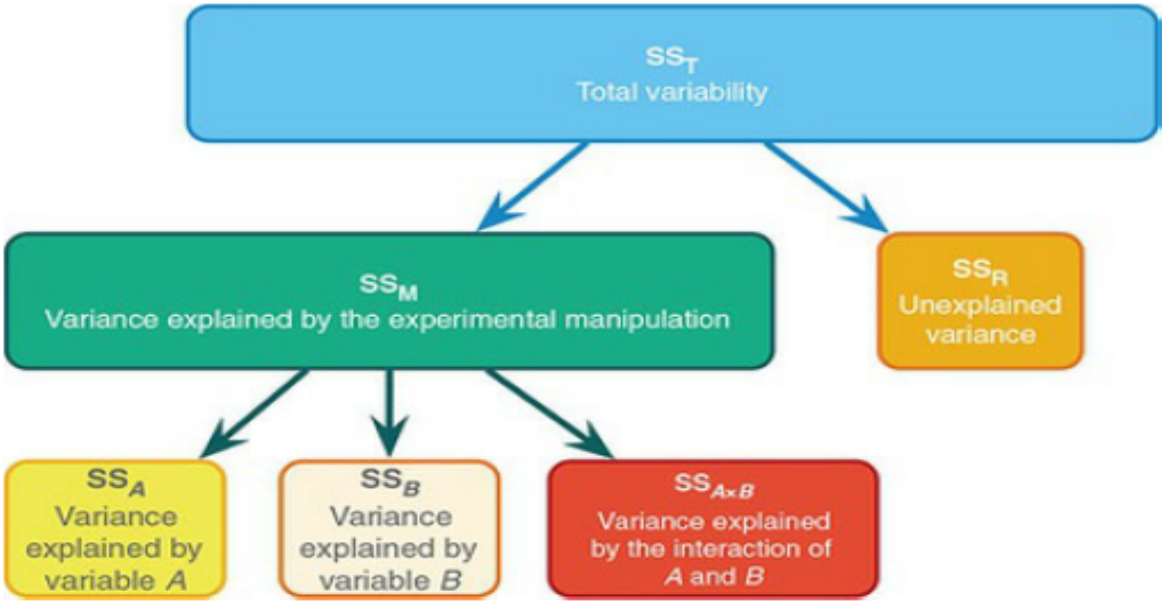

Figure 8.4: Breaking down the variance in a two-way factorial design

8.3 Total of Squares, SS_T

$$\begin{aligned} \text{SS}_T &\sum_{i=1}^{N}(x_i - \bar{x}_{\text{grand}})^2 \\ &= s^2_{grand}(N-1) \end{aligned} \tag{8.7}$$

In our example then, $SS_T = 2.525(48-1) = 118.675$

8.4 Model sum of squares, SS_M

SS_M in factorial design can be broken down to SS_A and SS_B and $SS_{A\times B}$. Recall SS_M is the difference between what the model predicts and the overall mean of the variable outcome. We also seen that with predictors that represent group membership, what the model predicts is the group means. So We work out the model sum of squares by looking at the difference between each group mean and the overall mean.

$$\text{SS}_M = \sum_{g=1}^{k} n_g(\bar{x}_g - \bar{x}_{grand})^2 \tag{8.8}$$

In our alcohol example, we have 6 separate groups for 2 IV. So,

$SS_M = 8(6.375-5.667)^2 + 8(3.5-5.667)^2 + 8(6.5-5.667)^2 + 8(4.875-5.667)^2 + 8(6.125-5.667)^2 + 8(6.625-5.667)^2 = 61.17$

df_M is $k-1$. So in this case $df = 5$.

8.5 Main effect for face type, SS_A

To calculate SS_A, you group the scores into attractive and unattractive faces, regardless of alcohol dose. (i.e. group all the scores into 2 different groups). Then we apply the same equation for SS_M to our data of 2 groups.

$$\mathrm{SS}_A = \sum_{g=1}^{k} n_g(\bar{x}_g - \bar{x}_{grand})^2 \tag{8.9}$$

$\bar{x}_g$ is the group mean of either attractive and unattractive faces. n_g is the number of scores in each group. i.e. number of people who rate unattractive and attractive. In our example, both groups have 24 participants.

$SS_{\text{facetype}} = 24(6.33 - 5.667)^2 + 24(5 - 5.667)^2$

df for SS_A is k -1. So df is 1 in our example.

8.6 Main effect of alcohol, SS_B

Similar to SS_A.

$$\mathrm{SS}_A = \sum_{g=1}^{k} n_g(\bar{x}_g - \bar{x}_{grand})^2 \tag{8.10}$$

Alcohol split into 3 groups, no dose, low dose and high dose.

$SS_{textalcohol} = 16(4.938 - 5.667)^2 + 16(5.668 - 5.667)^2 + 16(6.375 - 5.667)^2 = 16.53$

Similarly, *df* is k -1 . So in this example, $df = 2$

8.7 Interaction effect, $SS_{A\times B}$

We can get the interaction effect by subtracting.

$$SS_{A\times B} = SS_M - SS_A - SS_B \tag{8.11}$$

In our example, it will be 61.17-21.32-16.53 = 23.32.

df can be calculated in 2 ways. Both will yield the same results:

1. $df_{A\times B} = df_M - df_A - df_B$
2. $df_{A\times B} = df_A \times df_B$

8.8 Residual sum of squares, SS_R

SS_R is calculated the same way. It represents error in the prediction from the model, but in experimental designs this also reflects individual differences in performance or variance that cannot be explained by factors that were systematically manipulated.

$$\begin{aligned} SS_R &= \sum_{g=1}^{k} s_g^2(n_g - 1) \\ &= s^2_{group1}(n_1 - 1) + s^2_{group2}(n_2 - 1) + ... + s^2_{groupn}(n_n - 1) \end{aligned} \tag{8.12}$$

$df_R = df_1 + df_2 + ... + df_n$

8.9 F statistics

$$\begin{aligned} MS_A &= \frac{SS_A}{df_A} \\ MS_B &= \frac{SS_B}{df_B} \\ MS_{A\times B} &= \frac{SS_{A\times B}}{df_{A\times B}} \\ MS_R &= \frac{SS_R}{df_R} \end{aligned} \tag{8.13}$$

$$\begin{aligned} F_A &= \frac{MS_A}{MS_R} \\ F_B &= \frac{MS_B}{MS_R} \\ F_{A\times B} &= \frac{MS_{A\times B}}{MS_R} \end{aligned} \tag{8.14}$$