Chapter 1

1.1

From now on we will compare two or more independent samples, with two or more means. To do this, we will use the ANOVA model, (analysis of variance model). We use the variance to compare the two means, because when the sample means are all the same, the variance will be 0. The simplest form of the ANOVA model is the one-factor ANOVA model. In the ANOVA model we look at the X (independent variable) and the Y (dependent variable). The researcher is interested in the influence of the independent variable on the dependent variable.

When you are interested in comparing the means of two independent samples, you could simply use the independent t test. However, when you are interested in comparing more than two independent samples, you need to use the multiple independent t tests. This means you will get these null hypotheses:

$$\mu 1 = \mu 2, \mu 1 = \mu 3, \mu 1 = \mu 4, \mu 2 = \mu 3, \mu 2 = \mu 4, \mu 3 = \mu 4.$$

With these hypotheses you have to conduct different independent t tests. However, so many independent tests will give problems with the probability of making a *Type I error* (researcher incorrectly rejects a true null hypothesis). You can set an level for every individual test. However, the overall for the entire set of tests (experiment wise Type I error rate) is larger than the level for each individual test. The level will be higher, because the type I error accumulates across the tests. For each test you take a risk, the more tests you conduct, the more risk you take. The formula for the experimentwise error is:

$$\alpha\,(total) = 1 - (1 - \alpha)\hat{}C \qquad (1)$$

In this formula,

$\alpha(total)$ = experimentwise error. Alpha = Probability of type I error and C= the number of independent tests.

Because of the increasing risk of a Type I error, when doing more and more independent t tests, we do not want to use this.

The experimentwise error rate for C dependent tests is more difficult to determine, so we use:

$$\alpha \leq \alpha(total) \leq C\alpha \qquad (2)$$

To maintain as less risk as possible, we need to control the overall level, but also we need to maximize the power (probability of correctly rejecting a false null hypothesis). Conducting an overall test, the omnibus test, can do this. This test is used in ANOVA.

The one-factor ANOVA has only one independent variable or factor with two or more levels.

In the random-effects model, the researcher samples randomly some levels of the independent variable from the population of levels. From this generalizations can be made about all the levels of the population. The random selection of levels is almost the same as random selection of individuals or objects.

From now on we will use the fixed-effects model. This means that when the levels of the independent variable are selected, subjects are randomly assigned to the levels of the independent variable. In some situations researcher can control the assignment of subjects to groups, in other situations they cannot. So a distinction between these two needs to be made, because the analysis will not change, however the interpretation of the results will be. When a researcher has control over group assignments, the extent to which they can generalize their findings is greater than for those who do not have control. This is the difference between true experimental designs and quasi-experimental designs.

A repeated-measures model deals with subjects that are exposed to multiple levels of an independent variable.

A last characteristic of the ANOVA model is the measurement scale of independent and dependent variables. Because ANOVA is a test of means, the scale of measurement on the dependent variable is at the interval or ratio level. When the dependent variable is at a ordinal level, the Kruskal-Wallis test should be uses (discussed later). If the independent variable has characteristics of both ordinal and interval levels, both the ANOVA and Kruskal-Wallis procedures could be considered.

ANOVA is most often used with independent variables that are categorical – nominal or ordinal in scale. However it can also be used with interval of ratio values that are discrete (discrete variables are the ones that can only take on a certain value and that arise from the counting process).

The one-factor ANOVA is often called the completely randomized design.

- In summary, the characteristics of the ANOVA model are:
- The omnibus test controls the experimentwise error
- There is one independent variable with two ore more levels
- The researcher fixes the levels of the independent variable
- Random assignment of subject to the different levels
- Exposure of subjects to only one level of independent variable

Dependent variable is measured at least at the interval level, however the Kurskal-Wallis one-factor ANOVA can be used for dependent variables at ordinal level.

1.2

See layout of the data (see table 1.1 (page 6)). Each observation is called Yij, where the j shows us what group or level the observation belongs to and the i tells us the observation or identification number within that group. (Y34 means: the third observation in the fourth group, or level of the independent variable). The first subscript i ranges from i=1,…,,n, and second subscript j= 1,…,J. Thus there are J levels of the independent variable and n subjects in each group, for a total of Jn = N total observations.

$\bar{Y}_{.}$

The ,j is the sample group mean, and the overall sample mean is .

1.3

In ANOVA, mean differences are tested by looking at the variability of the means. We show how it is done. We begin with the hypotheses we will test with the ANOVA.

Hypotheses for a two-group situation of the independent t test, the null and alternative hypotheses for a two-tailed (nondirectional) tests are as follows:

$H_0: \mu 1 = \mu 2$
$H_1: \mu 1 \neq \mu 2$

With more than two groups, we use the omnibus test, as concluded before. The hypotheses for the omnibus ANOVA tests are as follows:

$H_0: \mu 1 = \mu 2 = \mu 3 = \cdots = \mu J$
H_1: not all the μj are equal.

The alternative hypothesis for the omnibus is written in a general form, this is to cover the multitude of the possible mean differences that could arise. These differences range from only two different means, to all of the means being different. If H0 would be rejected, a multiple comparison procedure (MCP) could be considered. This is used to determine which means or combination of means is significantly different (discussed in chapter 2).

We test mean differences by looking at the variability of the means. This is because with more groups is becomes complicated to look only at the mean. When doing a test of three groups, the null hypothesis is actually true, than the means are equal so no variability among the three groups. But the null hypothesis can also be false, this means the three samples are not from the same population, but from three different populations with a different mean. This means there is variability among the three group means. In short this means, that the statistical question becomes whether the difference between the sample means is due to the usual sampling variability expected from a single population, or the result of a true difference between the sample means from different populations.

Within-group variability is the variability of the observations within a group combined across groups. Between-group variability is the variability between the groups.

A low variability both within and between groups shows that performance is very consistent. When both variability are low, is it unlikely that the null hypothesis would be rejected. Also when within-group variability exceeds between-group variability is it unlikely one would reject the null hypothesis. When between-group variability exceeds within-group variability it is likely one would reject the null hypothesis. When within and between-group variability are both high, one may or may not reject the null hypothesis.

A low within-group variability is seen in the diagram because the distribution is very compact with little spread. So high within-group variability is seen by more spread. A High variability between groups is seen when each distribution is nearly isolated from one another with very little overlap. A low variability between groups is seen when there is a lot of overlap. (see figure 1.1 (page 8)).

The partitioning of the sums of squares in ANOVA is a new concept. The total sum of squares in Y is denoted as SStotal.This represents the amount of total variation in Y. Next step is to make from total variation the variation between the groups, denoted as SSbetw, and variation within the groups, denoted as SSwith. In one-factor ANOVA partition SStotal is as follows:

SStotal = SSbetw + SSwith
(3)

or

$$\sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}..)^2 = \sum_{i=1}^{n}\sum_{j=1}^{J}(\bar{Y}_{.j} - \bar{Y}..)^2 + \sum_{i=1}^{n}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}.j)^2 \qquad (4)$$

Where

- SStotal is the total sum of squares due to variation among all of the observations without regard to group membership
- SSbetw is the between-groups sum of squares due to the variation between the groups
- SSwith is the within-group sum of squares due to the variation within the groups combined across groups

This formulation of the portioned sums of squares is referred to as the definitional (or conceptual) formula because each term defined a form of variation. The definitional formula is rarely used with real data, because of the computational error. Instead a computational formula is used for hand computations.

Anova summary table (see table 1) is the result of the analysis. The first column lists the sources of variation. The second column notes the sums of square terms computed for each source. The third column shows the degrees of freedom for each source. The degrees of freedom are in general the number of observations that are free to vary. For the between-group source the degrees of freedom is J-1. Where J is the number of groups or categories or levels of the independent variable. For the within-group source the degrees of freedom are N-J. Because in each group there are n observations, so there are in each group n-1 degrees of freedom. There are J groups, so the dfwith (the degrees of freedom denominator) is J(n-1), or simply N-J.

To compare studies, you cannot simply use the SS. This would be unfair because the number of observations influences the SS. A fair comparison would be to compare the SS terms by their respective number of degrees of freedom. This is called a mean square term (MS). These numbers are in the fourth column.

$\text{MS}_{\text{betw}} = \text{SS}_{\text{betw}}/df_{\text{betw}}$

(5)

$\text{MS}_{\text{with}} = \text{SS}_{\text{with}}/df_{\text{with}}.$

(6)

These mean square terms are also variance estimates because they represent the sum of the squared deviations from a mean divided by their degree of freedom.

The last column is the F value, the summary test statistic of the summary table. The F value is computed by dividing the two mean squares or variance terms:

F = MSbetw/MSwith

(7)

Source	SS	Df	MS	F
Between groups	SSbetw	J-1	MSbetw	MSbetw/MSwith
Within groups	SSwith	N-J	MSwith	
Total	SStotal	N-1		

The F ratio tells whether there is more variation between groups than there is withing groups, which is required if we are to reject H0. F will be larger than 1 when there is more variation between groups than there is within groups. If the variation between and within groups is about the same, the F ratio will be approximately 1. In order to be able to reject the null hypothesis, we need to find large F values. The F value from the table is compared with the critical F value from the table (F-table, or Table A.4) to make a decision about the null hypothesis. To find the critical F value, you need to look at F (J-1, N-J). Thus the level of you choose (there are different tables for different ones). The dfbetw= J-1 is the numerator of the F ratio and the dfwith = N-J is the denominator of the F ratio. The test is a one-tailed test in order to be consistent with the alternative hypothesis. The null hypothesis is rejected when F test statistic exceeds the F critical value. This omnibus F test provides evidence of the extent to which there is at least one statistically significant mean difference between groups.

Again when the null hypothesis is rejected, it is not clear where the difference among the mean lies, because there are more then two groups. Than MCP can be used (chapter 2).

For a two-group test it is clear the difference lies between the two groups. Also for a two-group situation the F and t statistic follow the rule F=t2, for a nondirectional alternative hypothesis in the independent t test.

1.4

We will look at the ANOVA linear model, the estimation of parameter of the model, effect size measures, confidence intervals (CIs), power, and the expected mean squares.

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad (8)$$

The one-factor ANOVA fixed-effects model can be written in terms of population parameters:

Where:

- Y is observed score on the dependent (or criterion) variable for individual i in group j.
- μ is overall or grand population mean
- α is group effect for group j
- ε is random residual error for individual I in group j

Residual error can be due to individual differences, measurement error or other factors. The population group effect and residual error are computed as

$$\alpha_j = \mu_{.j} - \mu \qquad (9)$$

$$\varepsilon_{ij} = Y_{ij} - \mu_{.j} \qquad (10)$$

μ is the population mean for group j. The group effect (αj) can also be seen as the average effect of being a member of a particular group. A positive group effect shows a group mean greater than the overall mean. A negative group effect shows a group mean smaller than the overall mean.

In the one-factor fixed-effects model, the population group effects sum to 0. The residual in ANOVA represents that portion of Y not accounted for by X.

The parameter in the model,

$$, \mu, \alpha_j \text{ and } \varepsilon_{ij},$$

need to be estimated. The sample estimates are represented by , aj and eij. The last two are computed as:

Y represents the overall sample mean. The double dot subscript indicates we have averaged across both the i and j subscripts. Y.j represents the sample mean for group j, where the initial dot subscripts indicates we have averaged across all i individuals in group j.

There are various effect size measures, these indicate the strength of association between X and Y, that is the relative strength of the group effect. There are three measures:

1. the correlation ratio (generalization of R^2) and represents the portion of variation in Y explained by the group mean differences in X. It ranges from 0 to 1. When it is 0 it means that none of the total variance in the dependent variable is due to differences between the groups. When it is 1, all the variance of the dependent variable is due to group mean differences. It is a positively biased statistic, this bias is most evident for n < 30.

$$\eta^2 = \frac{SS_{betw}}{SS_{total}} \qquad (13)$$

2. ω2 is interpreted the same way as the first one. However it is less biased than the correlation ratio .

$$\omega^2 = \frac{SS_{betw} - (J-1)MS_{with}}{SS_{total} + MS_{with}} \qquad (14)$$

3. f is the final size measure. It can take values from 0 (when means are equal) to an infinitely large positive value. This effect is interpreted as an approximate correlation index, but can also be interpreted as the standard deviation of the standardized means.

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \qquad (15)$$

We can use f to compute the effect size, d. The formulas from computing d out of f depend on whether there is minimum, moderate or maximum variability between the means of the groups.

The effect sizes are interpret as:

- small effect: f= .1, η^2 or ω^2 = .01
- medium effect: f=.25, η^2 or ω^2 = .06
- large effect: f= .40, η^2 or ω^2 = .14

Confidence interval procedures are mostly useful in showing an interval estimate of a population parameter. These allow us to determine the accuracy of the sample estimate.

The probability of correctly rejecting a false null hypothesis, or power, can be considered as planned power (a priori) or observed power (post hoc). In the context of ANOVA we know that power is a function of , the sample size and the effect size. Planned power is mostly used to determine the adequate sample sizes in the model. So this method shows the right sample size for the desired level of power.

(Example of all the theory is at page 15 and 16)

The expected mean square for a source of variation represents the average mean square value for that source obtained if the same study were to be repeated an infinite number of times. The expected value of MSbetw is denoted as E(MSbetw). Now there are two situations. The first H0 is actually true, the second that H0 is actually false. In the first situation there is really no difference between the population group means, the expected mean squares are:

$$E(MS_{betw}) = \sigma_\varepsilon^2$$
$$E(MS_{with}) = \sigma_\varepsilon^2 \qquad (16)$$

The signma is the population variance of the residual errors.

So the ratio of expected mean squares is:

E(MSbetw) / E(MSwith) = 1
(17)

From this we can create the expected value of F:

E(F) = dfwith/(dfwith-2)
(18)

However in the second situation where H0 is actually false, so there are differences between the population group means, then the expected mean squares are:

$$E(MS_{betw}) = \sigma_\varepsilon^2 + (n \sum_{j=1}^{J} \alpha_j^2)/ (J-1)$$
$$E(MS_{with}) = \sigma_\varepsilon^2 \qquad (19)$$

So the ratio of expected mean squares is:

E(MSbetw)/E(MSwith) > 1
(20)

From this we have the expected value of F:

E(F) > dfwith/(dfwith-2)
(21)

So there is a difference in expected mean square between (E(MSbetw), when H0 is true and when H0 is false. The summation term that is added is the sum of the squared group effects.

From all this we can conclude that the F ratio represents the following:

F = (systematic variability + error variability)/ (error variability) (22)

The systematic variability is the variability between the groups and the error variability is the variability within groups.

1.5

There are three assumptions that are made in ANOVA models:

Independence

Observations need to be independent from one another, withing samples and across samples. To meet this assumption you need to (a) keeping assignment of individuals to groups separate through design of experiment and (b) keeping individuals separate from one another can meet this assumption.

Random sampling is needed in ANOVA, because the F ratio is very sensitive to violation of this assumption, because it will increase the likelihood of Type I and Type II error. Moreover a violation of this assumption can affect the standard error of the sample means.

The easiest way to check the independence assumption is to examine residual plots by groups. If the assumption is met, the residuals should fall into a random display of points for each group. The Durbin-Watson statistic can be uses to test for autocorrelation.

Violation of this assumption generally occurs in three situations:

1. Observations are collected over time
2. Observations are made within blocks
3. When observation involves replication.

Homogeneity of variance

The variance of each population are equal is known as homogeneity of variance or homoscedasticity. Violation of this assumption leads to bias in the SSwith term and increases the likelihood of type I or type II error. It is shown that not meeting this assumption or heterogeneity has severe effects. Dealing with such violations include:

- Using alternative procedures (Welch, Brown-Forsythe, and James procedures)
- Reduce
- Transform Y

Normality

Each of the populations should follow the normal distribution. The F test is relatively robust to moderate violations of this assumption. Effects of the violation will be minimal expect for small n's, unequal n's and/or for extreme nonnormality. The violation of this assumption can be caused by outliers. There are different graphical techniques that can be used to detect violations of this assumption:

- Frequency distribution sof the scores or the residuals for each group (boxplot, histogram etc.)
- Normal probability or quantile-quantile (Q-Q) plot
- A plot of group means versus group variances.
- Skewness and kurtosis of residuals
- Also a random display of points in a residual plot shows normality.

1.6

Until now we have assumed that there are equal n's. However this is not always the case. Unequal n's or unbalanced case also can be used. The interpretation of this analysis, assumptions, and so forth are the same as equal n's case.

1.7

There are different alternative for the one-factor fixed-effects ANOVA:

Kruskal-Wallis test

This test makes no normality assumption about the population distributions, but does still assume equal variances across the groups. When normality assumption is met or nearly met, the ANOVA is more powerful (less likelihood of type II error). Otherwise the Kruskal-Wallis test is more powerful.

The test works as follows. The observations on the dependent measure are ranked from the highest to lowest, disregarding group membership. This tests if the mean ranks are different across the groups such that they are unlikely to represent random samples from the same population. So H0 is that mean rank is the same for each group and H1 is that the mean rank is not the same across groups. The test statistic is denoted as H, and is compared to the critical value (table A.3). Then H > critical value the null hypothesis is rejected.

Two situations need to be considered. First, the chi-squared critical values is only useful when there are at least three groups and at least five observations per group. Second is that when there are tied ranks, the sampling distribution H can be affected.

The Welch test, Brown and Forsythe procedure and the James first and second-order procedures are for the heteroscedasticity condition. These do not require homogeneity. Research suggests that: (a) under homogeneity the F test is a little bit more powerful than any of these procedures and (b) under heterogeneity, each of these alternatives is more powerful than the F-test.

1.8

(Explanation SPSS, see textbook for the screenshots)

To conduct a one-way ANOVA test in SPSS, you need to have a dataset that consists of at least two variables or columns. One will indicate the levels or categories of the independent variable; the other is the dependent variable. To conduct the ANOVA test in SPSS this are the steps:

1. Go to “Analyze” and select “General Linear Model” and select “Univariate”.
2. Move the dependent variable into the box “Dependent variable”, and the same for the independent variable by moving It into the “Fixed Factors”.
3. Click on “Options”, this allows you to select options. Such as “Descriptive statistics”, “Homogeneity tests” etc.
4. Go back to the “Univariate” dialoge box, and click “Plots”. The independent variable needs to be placed on the “Horizontal axis”. Then click “add” to move the variable into the “Plots” box at the bottom of the dialog box.
5. From the “Univariate” box click on “save” to select the elements you want to save. From the “Univariate” box click on “ok” to return to generate the output.

From this we will generate the output, and these results need to be interpreted. (see page 28 and 29).

The table labelled “Between Subjects Factors” shows the sample sizes for each of the categories of the independent variable.

The table labelled “descriptive Statistics” provides the basic statistics, such as means, standard deviations and sample sizes.

The F test (and the p-value) for the Levene’s Test for Equality of Error Variances is reviewed to determine whether equal variances can be assumed. We meet the assumption when p>.

The df1 are the degrees of freedom from the numerator, so J-1. The df2 are the degrees of freedom from the denominator, so N-J.

For the one-way ANOVA test the shape of the residuals should be a normal distribution. In SPSS this can be examined by clicking on “Explore” and making a histogram of the residuals, explained in Introduction to statistical concepts. By looking at the skewness and kurtosis we can see whether it is a normal distribution.

A formal test for normality is the S-W test. It shows the evidence of the extend to which our sample distribution is statistically different from a normal distribution.

Moreover Q-Q plots are used to determine normality. These are graphs that plot the quantiles of the theoretical normal distribution against quantiles of the sample distribution. Points that fall on or close to the diagonal line suggest evidence of normality.

The boxplot is also a way to examine normality.

We also need to test for the assumption of independence. In some cases there is random assignment, however that is not always possible. By plotting the residuals against levels of our independent variable using a scatterplot gives an idea whether or not there is a pattern in the data. This can indicate if we have met the assumption.

The “Simple Scatterplot” screen can do this. The residual variable needs to be in the box “Y Axis”. The independent variable needs to be moved in the box “X axis”.

This scatterplot needs to be interpreted whether there is or there is not independence. When the points fall relatively randomly above and below the reference line there is evidence for independence.

Also a Kruskal-Wallis test can be done in SPSS, these are the steps:

1. Go to “Analyze” and select “Nonparametric Tests”, select “Legacy Dialogs” and finally select “K Independent Samples”.
2. From the main “Tests for Several Independent Samples” dialog box, click the dependent variable and move it to the “Test Variable List” box. Next click the grouping variable in the “Grouping Variable” box. In the grouping variable you need to show which categories of the grouping variable you want to include. You do this by clicking on “Define Range”. Then check the “Kurskall-Wallis H” in the “Test Type” labelled screen.

The Kurskal-Wallis test has a null hypothesis that the mean ranks of the groups of the independent variable will not be significantly different. By looking at the p value you can reject or not reject the null hypothesis.

Another test is the Welch and Brown-Forsythe test. These are the steps:

1. Go to “Analyze” and select “Compare Means”. Select “One-way ANOVA”.
2. Click the dependent variable in the “Dependent list”, and move the independent variable in the “Factor” box.
3. Click on “Options”, this will show you different statistics. Make sure you check Brown-Forsythe and if you want you can check more.
4. Generate the output.

There is a p-value that can or reject or not reject the null hypothesis. When the null hypothesis is rejected there is a statistically significant difference in the mean number of the different groups.

To compute the post hoc power for the One-way ANOVA, we will use the G*Power. To find the one-way ANOVA, we select “Tests”, than select “Means” and then “Many groups: ANOVA: One-way (one independent variable)”. The “Type of Power Analysis” needs to be selected. We select “Post hoc: Compute achieved power-given , sample size and effect size”.

(The default setting for the “Test family” is the “t-test”. The default selection for “Statistical Test” is “Correlation: Point biserial model”. Following the steps this will change automatically).

The “Input Parameters” must be specified including:

- Effect size f
- Alpha level
- Total sample size
- Number of groups in the independent variable.

After specifying the input parameters click on “Calculate” to find the power statistics. The “Output parameters” provide the relevant statistics given the input statistics. Eventually you get the post hoc power of the test – the probability of rejecting the null hypothesis when it is really false. A sufficient power is >0.80 (80%).

Conducting a power analysis is recommended so that you avoid a situation where the sample size was not sufficient to reach the desired level of power.

For a priori power, we can determine the total sample size needed given an estimated effect size, alpha level, desired power, and number of groups of our independent variable.