Mini-Dissertation Write-up Guide

Part 05 - Data Pre-Processing

Dr. Gordon Wright

February 5, 2024

Lab 15 Missing Data and Assumption Testing in ANOVA

Remember what Jan told you about assumptions...

By the end of the session, you will have:

• Considered your missing data

• Familiarised yourself with the assumptions you need to meet to perform an ANOVA (depending on your design)

Assumptions when using ANOVA

ANOVA is a 'parametric' test, meaning that it assumes the data you are analysing conforms to a series of underlying parameters, or features. ANOVA doesn't usually collapse when these assumptions are not met, it's what we call 'robust', but you'd normally consider possible corrections, or alternative approaches to inferential test, such as non-parametric tests. But you know all this. Missing data Firstly, and this isn't really an assumption, but you need to make sure you don't have any missing data. If you have cells in your SPSS data set that are empty, SPSS will exclude that participant from any analyses that rely on the data. It's just a fact that people drop out of studies, or miss questions. So a certain number of participants, who haven't completed large parts of your study will have to be removed.

By running Descriptive Statistics you will be shown the N for each variable included. Obviously, you may have one dependent variable, or you may have two or four. They should all be equal (in an ideal world), but the Valid N (listwise) refers to the number of participants for whom you have complete data, and so would be the number used in an ANOVA.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Individual_Trial_1_RTms	14	111	964	510.93	261.928
Individual_Trial_2_RTms	17	111	777	458.00	200.882
Valid N (listwise)	14				

If you have a few missing values in your data, you can do what's called a mean imputation, which is just a fancy way of replacing the missing values with the mean of all the other values of the same variable from the rest of your data set. You can `Transform` the variable and `Replace Missing Values using Series Mean`. Or calculate the mean yourself and copy it into each missing cell of a particular value.

If you choose to do this, please make sure to document what you did for inclusion in your Mini-Dissertation submission. If you think that you have more than just a few cases, please talk to your Lab Tutor and we can help resolve the issue and offer specific guidance.

Assumptions (based on design and measurement)

When you choose to analyse your data using a two-way ANOVA, a critical part of the procedure is checking that the data you want to analyse can actually be analysed using this test. In fact, the two-way ANOVA has five (or six) assumptions (depending on the flavour of ANOVA) that you have to consider, three of which you can test for using SPSS.

I shall show you how to assess these three assumptions in SPSS and briefly explain how to interpret the results.

But before that, let's look at the first three checks you should perform.

1) You need to have a continuous dependent variable. This should be something you have been aiming for all along, so we can move ahead.

2) You have two categorical independent variables with two groups or levels in each. ANOVA works for more complex designs, with more than two levels of an IV, but again, you should have a 2x2 ANOVA design.

3) Your observations are independent, if you are running an independent or between-groups ANOVA. Independent means exactly what it sounds like, and should originate from separate trials of an individual participant, and not in any way related to another participant's data. This is obviously not the case if you are running a Mixed, or Repeated Measures ANOVA, as the within-participant measures are not independent, they come from the same individual.

Parametric Assumptions (tested via SPSS)

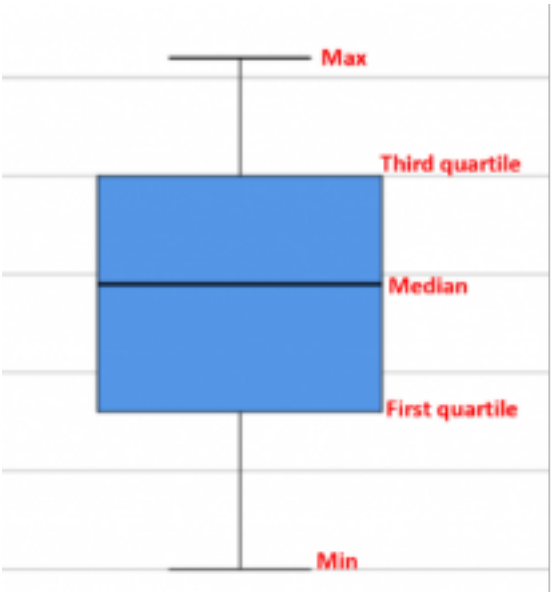
It is relatively common that your data may violate (i.e., fail) one or more of these three assumptions. In each case, there are steps you can take to proceed, and these range from correcting your data in some way, choosing an alternative test, or just carrying on.

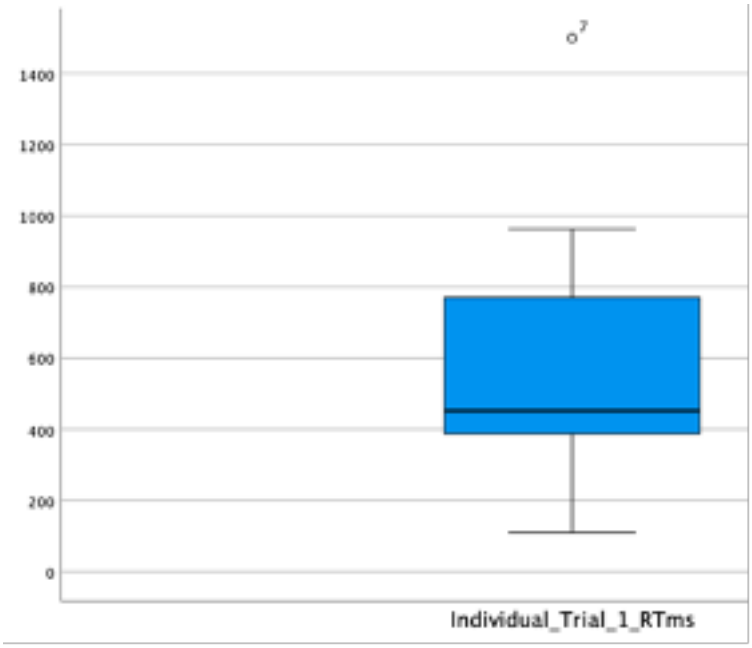
In the overwhelming majority of cases for your Mini-Dissertation, you are encouraged to proceed with the ANOVA in the normal fashion and make it clear which of the assumptions were violated. We wish to be able to assess you on following the procedure for the ANOVA you have learned this year. If in doubt about what to do, please talk to your Lab Tutor.

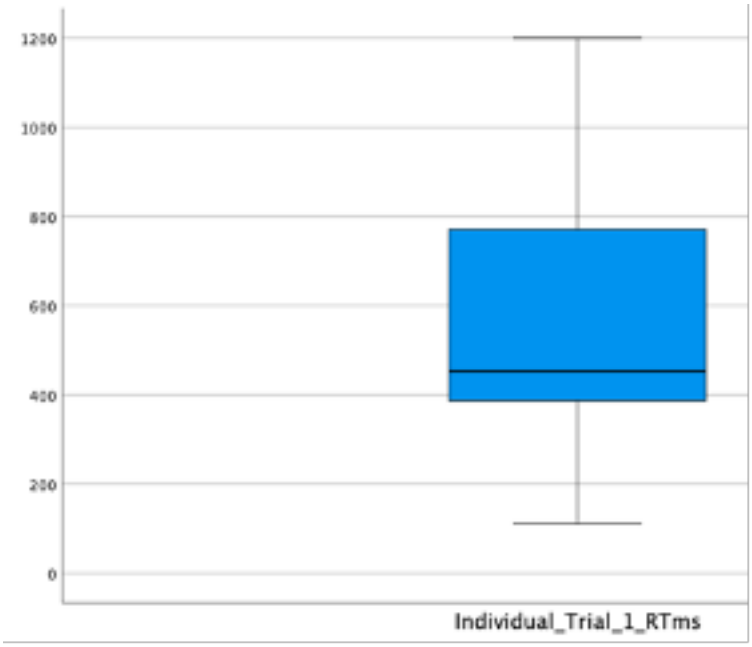
A brief note that you can gloss over if you wish. These assumptions refer to each cell of your design, and also to the residuals, or error, from the ANOVA model, not the actual observed data. This is worth remembering if you choose to do more advanced statistics, but, in essence, you can think about the assumptions being applied to the observed data (i.e. the data you have in your SPSS dataset).

No significant outliers

Testing this assumption can be done by producing a Boxplot. These are pretty cool figures and offer a 'five number summary' of a variable; the median, the first and third quartiles, as well as the minimum and maximum values. But what's really groovy, is that if a figure exceeds the interquartile range by 1.5x it's classed an 'outlier' and excluded. Exceed the interquartile range by 3x and it's termed an 'extreme value' and excluded. Outliers get a little circle and Extreme Values get a star to denote them. See below.





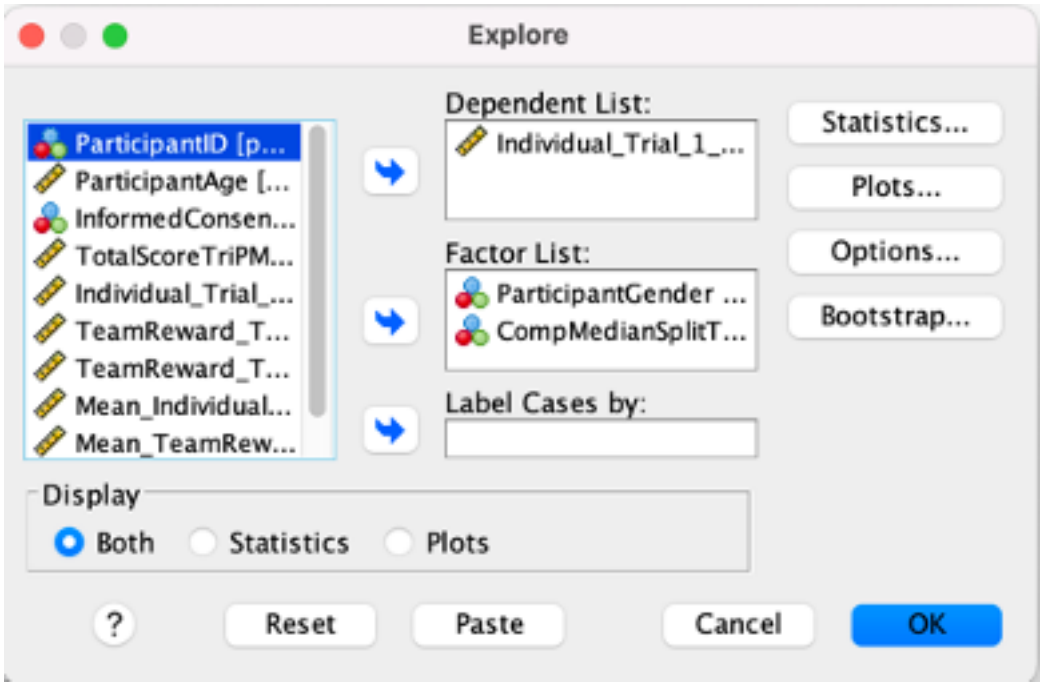


In the first panel, you can see what all the lines in a boxplot refer to. The gap between the First quartile and the median is the lower Inter Quartile Range (sometimes abbreviated to IQR) and if a value falls below the first quartile line by more than 1.5 times this value = outlier. The same applies to values above the median, but the higher IQR is used.

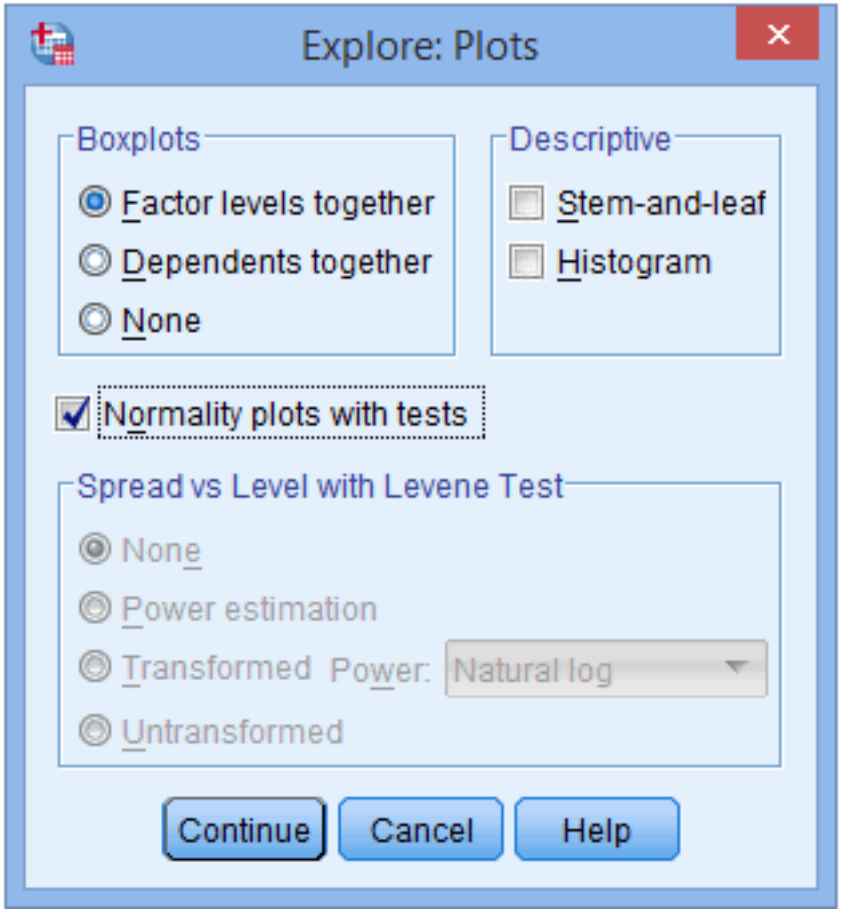
In the second panel, I messed about with one of my variables which originally had a

maximum value of 964 and had no outliers. I popped in a value of 1200, but this wasn't big enough to trigger the outlier warning, as the higher IQR is actually quite big. So I had another go and made the new data value 1500. You'll see that in the third panel. It's the little circle with the number 7 beside it, identifying the row in my data that is an outlier. You see the maximum value bar has dropped back to 964. Sweet. Row 7 is an outlier and you can either exclude that participant, or proceed. In either case, you should VERY CAREFULLY describe your choice. Even if you found no outliers, you should report that you conducted an examination of a boxplot. E.g. "No outliers were observed for the Dependent Variable, as determined by inspection of a boxplot".

To run a boxplot. Go to `Analyse`, `Descriptive Statistics`, and then `Explore`, put your DV in the top box, and your (between-group) IVs together in the `Factor List` box. For a between-groups design, or fully independent design, you will have two Factors (or IVs) and a single dependent variable. If you have a Repeated Measures design, you will have four dependent variables and no Factors, for a mixed design, two dependent variables and one Factor, but the same process ensues.



Click on `Plots` and for the purposes of this, deselect `Stem-and-leaf` and click (and pay attention to) `Normality plots with tests`. You will need that later. Press `continue` and your plots will be produced.



You will have plots for both levels of both of your IVs and you should consider them all in the same way as the single example I showed above.

Normality of distribution

The two-way ANOVA assumes that the data are normally distributed in each cell of your ANOVA design. This can be checked with a visual inspection of a Histogram, but I'm not confident doing that by eye, or you could look at the skewness and kurtosis values (look them up). But the easiest way is the way you have already done. Yes. You clicked for the Normality plots with tests earlier under Explore, and this has produced some tests and a couple of plots above your boxplots.

You will see that a Shapiro-Wilk test has been run for each of the two levels of the independent variable. In my toy example from last week, I had gender with 3 levels, and Psychopathy with a computed median split resulting in High Psychopathy and Low Psychopathy groups. The green highlight shows how to identify the independent variable that

is being tested for normality of the dependent variable (Individual Reward Trials in Milliseconds).

If you look at the Sig. column located under the Shapiro-Wilk column, you will find the significance value for this test for each group of the independent variable.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	CompMedianSplitTriPM	Statistic	df	Sig.	Statistic	df	Sig.
Individual_Trial_1_RTms	HighPsychopathy	.217	9	.200*	.926	9	.447
	LowPsychopathy	.239	6	.200*	.944	6	.692

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	ParticipantGender	Statistic	df	Sig.	Statistic	df	Sig.
Individual_Trial_1_RTms	Female	.220	8	.200*	.942	8	.628
	Male	.331	5	.078	.740	5	.024
	PreferNotToSay	.260	2	.			

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

For the purposes of this test, anything OVER p = .05 is good, correctly noted as (p > .05). So in the toy example above, and in APA format:

> Reaction times were normally distributed for both High Psychopathy group and Low Psychopathy groups, as assessed by Shapiro-Wilk's test (High Psychopathy W(9)= .926, p = .447, Low Psychopathy W(6)= .944, p = .692) however reaction times were normally distributed for females but not males (Females W(8)= .942, p = .628, Males W(5)= .740, p = .024).

Homogeneity (or Equality) of Variances

The final assumption we need to assess is the idea that the dependent variable is of roughly equal variance or spread in each cell of the design. The assumption of homogeneity of variances is tested using Levene's test of equality of variances, which is found in the Levene's Test of Equality of Error Variances table, as shown below.

As you can see, this just came in ABOVE our threshold at p = .052, meaning the assumption is met (just).

Levene's Test of Equality of Error Variances[a,b]

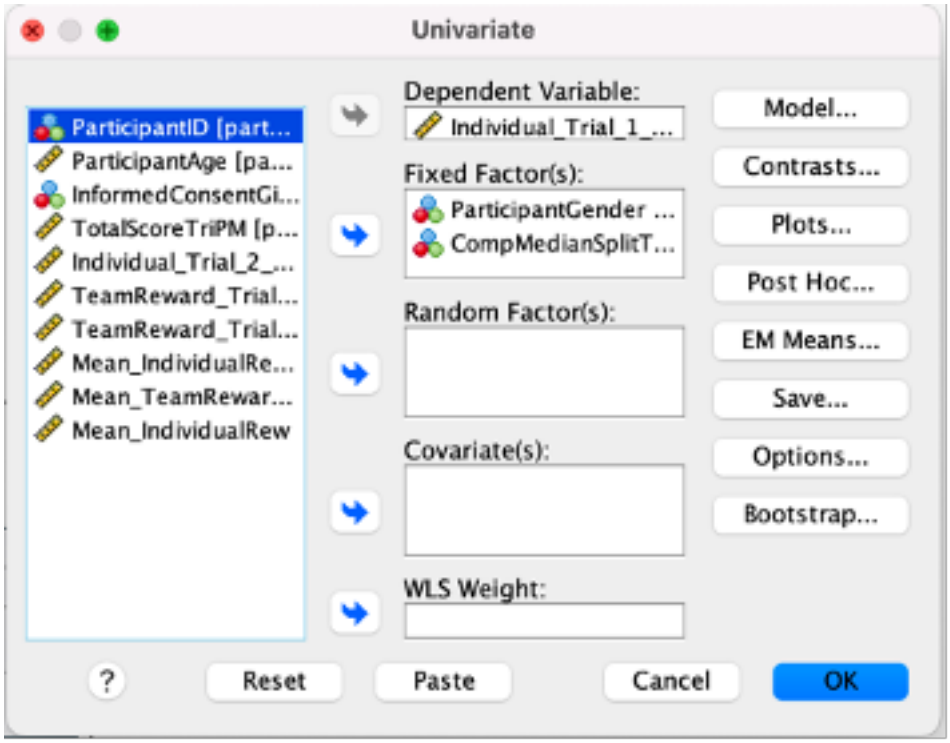
		Levene Statistic	df1	df2	Sig.
Individual_Trial_1_RTms	Based on Mean	3.798	3	9	.052
	Based on Median	3.262	3	9	.073
	Based on Median and with adjusted df	3.262	3	5.948	.102
	Based on trimmed mean	3.671	3	9	.056

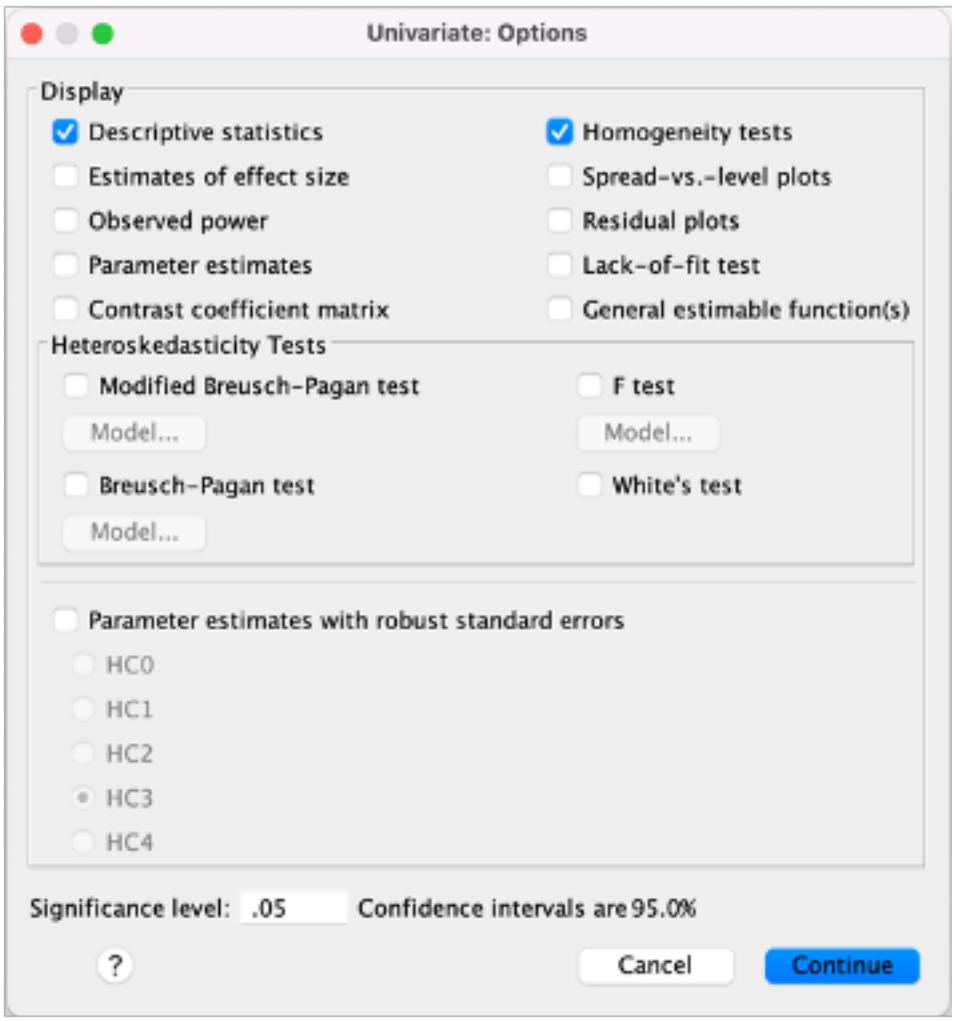
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: Individual_Trial_1_RTms

b. Design: Intercept + part_gender + psychopathy_medlansplit + part_gender * psychopathy_medlansplit

This test can be performed as part of the ANOVA itself, and Homogeneity tests is what you need to select to have it included in your ANOVA output.





Remember, the same rules apply for Levene's test as the Shapiro-Wilk test, we do not want to find a significant result ($p < .05$).

FOR MIXED DESIGNS

A further assumption of the mixed ANOVA is that there are similar covariance matrices (don't worry about it until you start your MSc). You can test for this with Box's test of equality of covariance matrices, which is presented in the Box's test of equality of covariance matrices table, as shown below:

Box's Test of Equality of Covariance Matrices[a]

Box's M	7.082
F	.528
df1	12
df2	8548.615
Sig.	.898

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + group Within Subjects Design: time

But it is obtained by selecting for Homogeneity tests as with the previous examples, but you need to be on the look out for it to report it! And you won't find the Levene test, so it's easy to get confused in the case of a mixed design.