MA4505 2015 Lecture 4B part 1

## **Testing Assumptions for ANOVA.**

For ANOVA to be a valid analysis, the following assumptions must be valid. Two of these assumptions are as follows:

- Each population from which a sample is taken is assumed to be normal.
- Each sample is randomly selected and independent.

We are interested in assessing the validity of a further three of these assumptions in particular. (For the sake of brevity, we will focus on these particular assumptions).

- The samples have equal variances (homogeneity of variances),
- The residuals are normally distributed,
- The residuals have mean of zero, and have a constant variance.

We will test the validity of these assumptions for the ANOVA model fitted to the Paracetamol data, used in previous labs.

**Bartlett's test for Homogeneity of Variances**:

Equal variances across samples is called homogeneity of variances

Bartlett's test is used to test if multiple samples have equal variances.

Some statistical tests, such as the analysis of variance, assume that variances are equal across groups or samples. The Bartlett test can be used to verify that assumption.

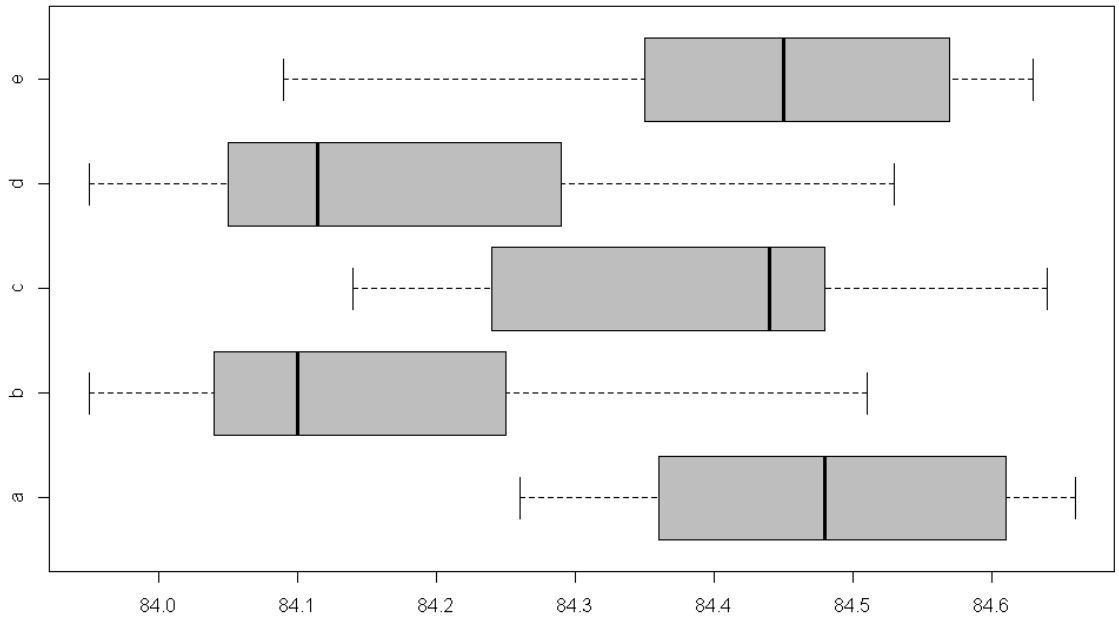
- The null hypothesis is that the samples have equal variance.
- The alternative hypothesis is that at least one sample has a significantly different variance.

To implement the Bartlett test in R, we simply use the command `bartlett.test()`, with the model specification as an argument.

We can also use boxplots to implement a graphical complement to the procedure.

```
boxplot(y~group)

bartlett.test(y~group)
```



While it is clear that there is a significant difference in the medians from each sample. (i.e. Samples b and d have different medians, and hence presumably different means, from the rest of the samples).

Of particular interest now however is the dimension of the boxplot – an indicator of the variance of each sample.

From the boxplots, it is clear that the length of the each is more or less the same. The Bartlett test provides for us a formal hypothesis test to assess the validity of equal variances.

The following R output confirms the validity of the assumption. The very high p-value indicates that we fail to reject the null hypothesis, and proceed with ANOVA procedures.

```
> bartlett.test(y~group)

        Bartlett test of homogeneity of variances

data:  y by group
Bartlett's K-squared = 0.8327, df = 4, p-value = 0.934
```

## Testing the residuals

The residuals are the basic building block for checking the suitability of the working ANOVA model and validating the assumptions on which it is based. In general the residuals can be defined as

***Residuals = Observed Values – Fitted Values***

To test the remaining assumptions, we can use the diagnostic plots provided by the `plot()` command, and also by performing the Shapiro-Wilk test for residuals.

## Shapiro-Wilk Test

The null hypothesis is that the residuals are normally distributed. The high p-value indicates that we should not reject the null hypothesis, and that the residuals are indeed normally distributed.

```
> shapiro.test(Residuals)

        Shapiro-Wilk normality test

data:  Residuals
W = 0.9819, p-value = 0.6348
```

A graphical complement – the QQ plot – should reinforce this. The QQ plots for residuals will be discussed shortly.

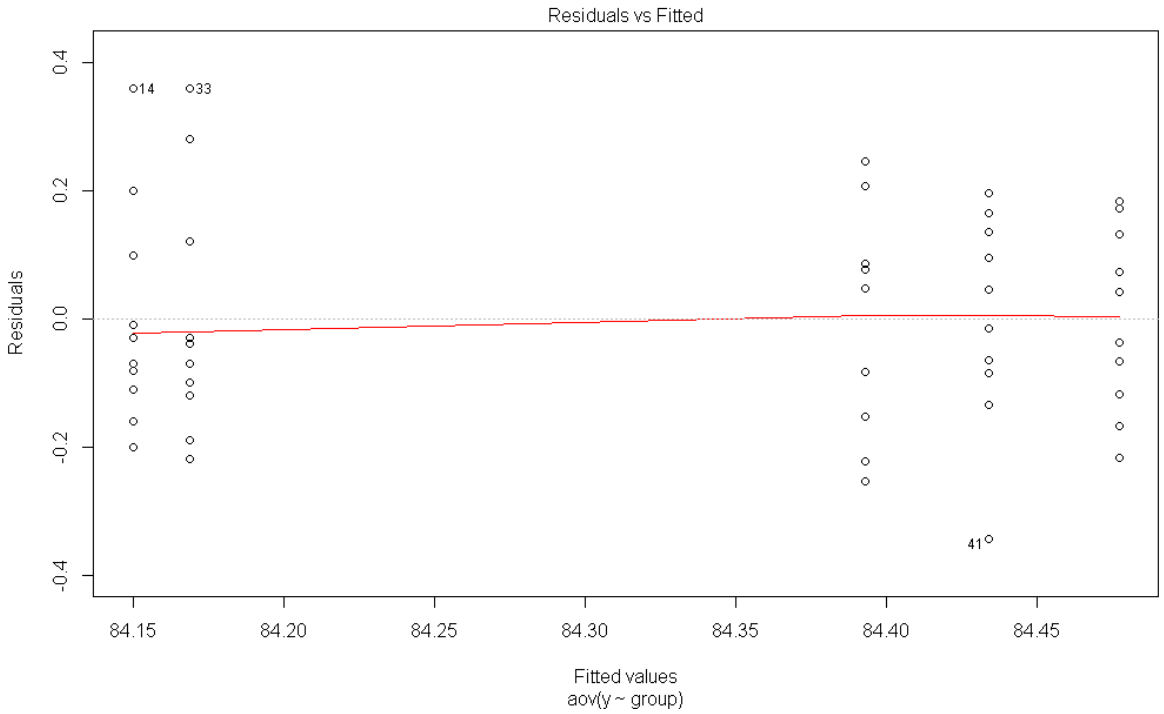
## Diagnostic plots

A series of diagnostic plots can be constructed using the plot command, specifying the name of the fitted model. Four plots will be presented to the screen sequentially.

```
> plot(Model1)
```

The first plot is a "*residual versus fitted value"* plot. Here we are interested in two specific things:

The mean value of the residuals – which should be zero – is indicated by a red trendline. Of key interest is how this trend-line varies through the plot. This indicates how much each cluster of data points influences the calculation of the mean. Ideally the red trend-line should be more or less horizontal at the zero level.




Secondly we are interested in the assumption of constant variance. Disregarding the data points are "stacked", we can see that the range for each cluster is more of less the same.

## Diagnostic plot 2- the QQ plot

This plot is a QQ plot used to test the normality of the residuals?
Id the points follow the diagonal trend-line, then the residuals can be assumed to be normally distributed.

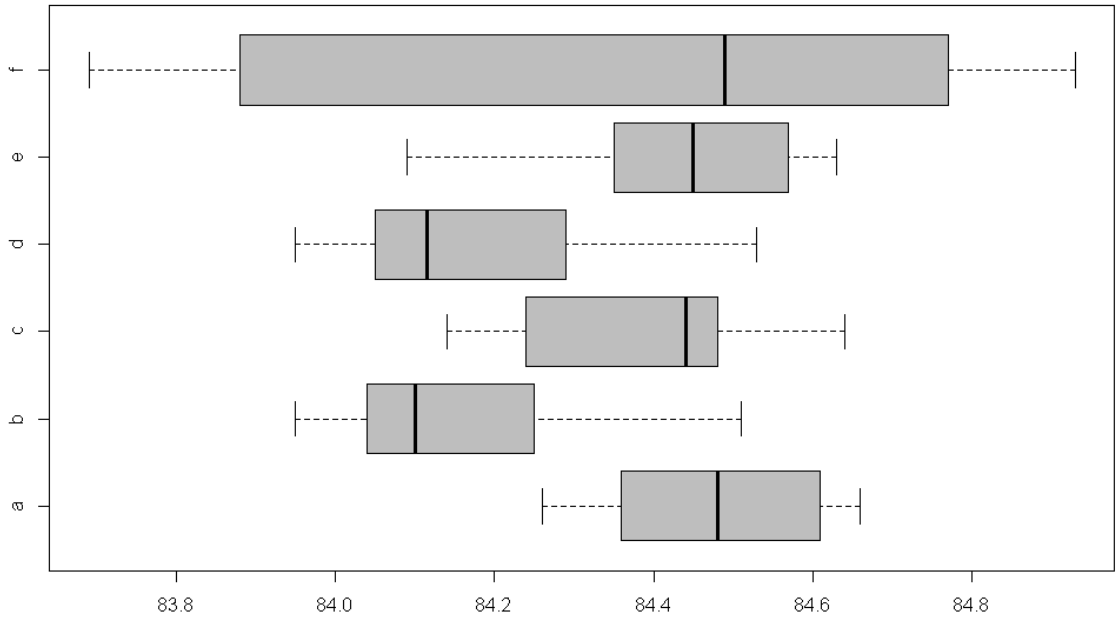
(We will not consider Diagnostic Plots 3 and 4 in this course.)

## A Counter-Example

To provide a counterpoint for our previous example, let us update the paracetamol example. Suppose a sixth sample (***f***) is added to the analysis. We will run the analysis again to consider how this affects the analysis.

A quick examination of the boxplots (below) indicates that the mean of this sample is roughly the same as the main cluster of samples (i.e. ***a,c,e***)

However it is clear that the range of the box-plot for sample ***f*** is much larger than those of any of the other samples.




We perform the Bartlett test again for the updated data.
The low p-value indicates that we should reject the hypothesis of equal variances.

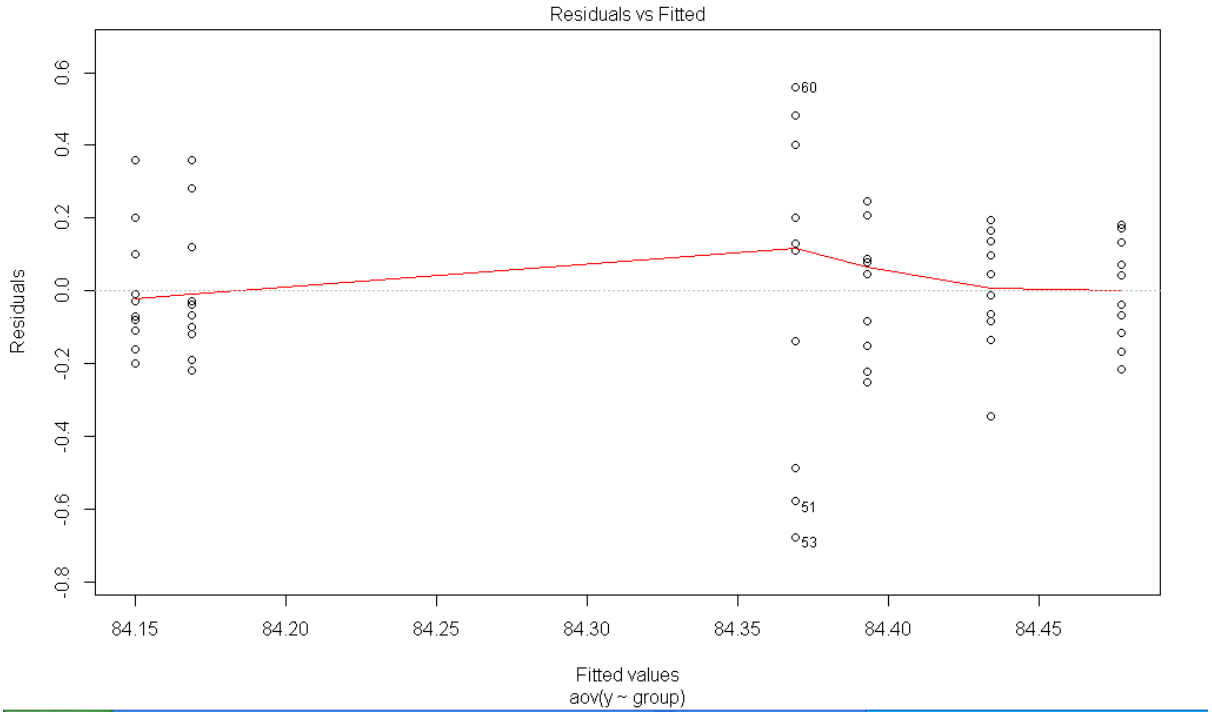
```
> bartlett.test(y~group)

        Bartlett test of homogeneity of variances

data:  y by group
Bartlett's K-squared = 20.0556, df = 5, p-value = 0.00122
```

By examining the diagnostic plots, other concerns about the validity of the model would be raised.

Diagnostic plot 1: Residuals v Fitted Values



Diagnostic Plot 2: QQ plots for Residuals.