**Running a single factor analysis of variance in SPSS**

This example uses data on housing prices. You can find a [description of where to find the data and what the variables represent](https://github.com/pmean/classes/blob/master/biostats-2/data/woodard-data-dictionary.yaml) on my Github site.

First, it is always a good idea to plot the data before you run the regression analysis. Select Graphs | Boxplot from the SPSS menu. This calls up the Boxplot dialog box.

A screenshot of a computer

Description automatically generated

Click on the Simple icon. Make sure that the Summaries for groups of cases option is selected, then click on the Define button.

This calls up the Define Simple Boxplot dialog box.

A screenshot of a computer

Description automatically generated

Add the continuous variable Building$ to the Variable field and the categorical variable No.Baths to the Category Axis box. Then click on the OK button.

You receive a nice table outlining how many observations you have in each group and how many missing values.

A table with numbers and text

Description automatically generated

The default graph is not too bad, but you should always make a few enhancements. In particular,

1. Swap out the extraneous blue color for white (not transparent!)
2. Use dollar signs and commans for building$
3. Change the scale on the Y axis
4. Add tick marks to both axes
5. Put your name and date in the title.

After these enhancements, this is what your graph looks like.

A graph with lines and numbers

Description automatically generated

There is some evidence of heteroscedascity. The three bathroom houses are a lot more variable that houses with fewer bathrooms. There are several outliers, but the only really troublesome one is the three bathroom house worth over $600,000.

It is easiest to examine the expensive houses by sorting from high to low. Select Data | Sort Cases from the menu. This brings up the Sort Cases dialog box.

A screenshot of a computer

Description automatically generated

To

Add Building$ to the Sort by field, change the sort order option to Descending and then click on OK.

The $600,000+ plus house is now at the top of the dataset.

A screenshot of a computer

Description automatically generated

There is nothing unusual about the other values, such as age or the various measures of house size. Perhaps the zip code represents a particularly expensive part of town. You should show this data to the research team and encourage them to look for clues as to why Building$ is so much larger for this house.

To run a single factor analysis of variance, you have two options:

1. Select Analyze | Compare Means | One-Way ANOVA, or
2. Select Analyze | General Linear Model | Univariate

Both have advantages and disadvantages. For simple problems, the former is best and that is the choice illustrated here. Here is the One-Way ANOVA dialog box.

A screenshot of a computer

Description automatically generated

Add the continuous outcome variable (Building$) to the Dependent List field and the categorical predictor (No.Baths) to the Factor field. Although many researchers like effect sizes, I do not. I always make sure to uncheck the Estimate effect size for overall tests box.

To get post hoc tests, click on the Post Hoc button. This brings up the Post Hoc Multiple Comparisons dialog box.

A screenshot of a computer

Description automatically generated

The sample sizes in each category vary quite a bit, so do not use a Tukey post hoc test. Perhaps you should be concerned about the unequal variation, but the Bonferroni option seems like a reasonable choice.

The output includes two tables. The first table is the analysis of variance table.

A table with numbers and letters

Description automatically generated

Notice that the numbers are so large that SPSS needs to use scientific notation. The sum of squares are all multiplied by 10 raised to the 11th power or 100 billion. So the sums of squares are 241 billion, 476 billion and 717 billion. There are n=99 observations total, and k=5 levels for the categorical predictor. The degrees of freedom are k-1=4, n-k-1=94, and n-1=98.

The F-ratio, 11.876, is large and the p-value is small (less than 0.001). So you would reject the null hypothesis and conclude that the average building cost differs among houses with different numbers of bathrooms.

The second table shows multiple comparisons using the Bonferroni correction.

A screenshot of a graph

Description automatically generated

This table is difficult to interpret because of the large number of rows and columns. In this particular case, it helps to start from the bottom. The confidence intervals for the differences of 3 baths versus 2.5, versus 2.0, versus 1.5, and versus 1.0 are all positive, indicating that 3 bath homes have statistically significant larger average building prices than all other bathroom numbers. You can draw largely the same conclusion by looking at the p-values.

Homes with 2.5 baths are not more expensive on average than homes with 2.0 and 1.5 baths, but they are more expensive on average than homes with 1.0 baths.

Homes with 2.0 baths are not significantly more expensive on average than homes with 1.5 and 1.0 baths.

Homes with 1.5 baths are not significantly more expensive on average than homes with 1.0 baths.