Quantitative Methods (GEO 441)
SPSS Mini-Lab: Outliers
Dr. Paul Marr

Please view the mini lecture Outlier Detection before completing this mini lab. It is found at:
http://webspace.ship.edu/pgmarr/Geo441/Lectures/OPT%201%20-%20Outlier%20Detection.pdf

Copy the file **S:\GEO\Marr\Quantitative Methods\SPSS Example Data\Outliers.sav** to your portable media.

- Start SPSS.

- Open Outliers.sav

1. Methods of Outlier Detection in SPSS (1)
    a. Q-Q Plots
        i. Analyze > Descriptive Statistics > Q-Q Plots
        ii. Move the variable **Syr River Inflow** to the *Variables* window.
        iii. Click **Ok**.
            - In the *Estimated Distribution Parameters* table *Location* is the <u>mean</u> and *Scale* is the <u>standard deviation</u>.
            - The *Q-Q plot* show two observation that fall well off the *Expected Normal Value* line, one high and one low.
            - The *Detrended Q-Q plot* show one observation in particular that is nearly 8 standard deviation from the mean.
        iv. Double click on the *Detrended Q-Q plot* to open the *Chart Editor* window.
        v. Click on the *Data Label Mode* button  and then click on the suspect observation in the graph to label it.
        vi. The observation will appear with a circle around it, signifying it has been selected. Right click on the highlighted observation and select *Go to case*.
            - The case will be highlighted in the *Data Editor* window.
            - Note that the Syr River inflow for 1960 was much greater then all other years.
        vii. Close the *Chart Editor* window.
    b. Explore
        i. Analyze > Descriptive Statistics > Explore
        ii. Move the variable **Syr River Inflow** to the *Dependent List* box.
        iii. Under *Display* select **Both**.
        iv. Click on **Plots…** to open the *Plot Options* window.
        v. Under *Boxplots* select **None**, under *Descriptive* uncheck **Stem-and-Leaf**, and check the **Normality plots with tests** box.
        vi. Click on **Statistics…** to open the *Statistics Options* window.
        vii. Check the **Outliers** box and click **Continue**.
        viii. Click **Ok**.
            - Unfortunately, SPSS does not include either a Dixon or a Grubbs test for outliers, however, the data provided in the *Descriptives* and *Extreme Values* tables can be used to perform either test by hand.

- As we can see from the normality tests, these data are non-normal. Both the Grubbs and Dixon tests require normality and would be inappropriate here.
- Non-parametric tests use ranks and are not influenced by outliers. Since these data are non-normal we would use non-parametric tests to analyze them, and outliers will not be a problem.

Notes:

- If you suspect an outlier in your data, first run a normality test. If the data are non-normal and you will be using non-parametric tests, outliers will not be an issue.
- If the data are normal, or nearly so (probabilities very close to 0.05), and you anticipate using parametric tests, then outliers may be problematic.
    - First, determine the circumstances that may have resulted in the outlier. In the case above, the Syr River outflow may have been the result of higher than average precipitation. If so, then the observation is not an outlier but an effect.
    - If precipitation was normal for that year, then further investigation is needed (what other factors would cause increased outflow? Opening a spillway?).
    - If no circumstance that would cause increased outflow is found, then the observation's measurements may be suspect, in which case removing the observation from the data set is warranted, as it is likely either a measurement or coding error.
- If you have reservations about removing a potential outlier(s) from your data set (e.g. small sample size) then the ***BEST*** option—***regardless of whether the data are normally distributed or not***—would be to use <u>non-parametric tests</u>.
    - Non-parametric tests are immune to the negative effects of outliers and if the data are non-normal (especially if the data are non-symmetrical with activity in the tails), non-parametric tests are much more powerful than even the most robust parametric tests.