Chapter 4 – Data Screening

Data screening will really depend on the type of analysis you are doing because each analysis has different assumptions.

Issues in data screening:

1. Accuracy of Data file – check for typos! How do you do this with a whole big data set?
   1. Frequencies – so you can see if there are any numbers that shouldn’t appear in your dataset.
   2. What to check for:
      1. Min – is that the lowest score you can get?
      2. Max – is that the highest score you can get?
      3. Means – is that a number that makes sense?
      4. SD – is that within the range?
      5. Missing values – do you see holes when you shouldn’t?
2. Correlations – especially important for research using scales with composite scores
   1. Inflated correlations – or correlations that are super high. You need to make sure you aren’t going to use a total score and an item that was included in a total score in a regression type analysis. That will cause multicollinearity = bad.
   2. Deflated correlations – checking for restriction of range – best for regression analyses. You do not want to have only a small number of people in a category or have only collected 10-15 year olds when you want all kids. You could create null correlations by only measuring a small range of people.
3. Missing data – possibly the most important problem. What do you do when people don’t answer questions you’ve asked them? What do you do when someone leaves the study half way through?
   1. Examine the pattern of missing data –
      1. Randomly scattered missing data is ok. Means there’s just some errors or people just randomly left questions blank or some other goof that is life.
         1. MCAR – missing completely at random
      2. Non-random spread of missing data – let’s say you are doing a study on alcohol attitudes around campus here. What if suddenly everyone skipped the question on how much they drink? What does that mean? If you eliminate those people you are leaving out all sorts of important information.
         1. MNAR – missing not at random
      3. Sample size – if you have a large data set – 5% or less is fixable with missing data input. A lot of data missing from a small set = bad maybe collect more data.
      4. How do I check for patterns and if this is going to be a big deal?
         1. First look where the data is missing – you should see it in N or missing when you do frequencies.
         2. Code people into two groups – people with missing data, people without. Run a simple procedure (ANOVA, T-TEST) on one of the other DVS to see if those groups are different. No? Move along. Yes? Crap.
         3. Or – try the analysis with the missing data people and try it without…did you get the same results? Then move along. If not, try a missing values analysis.
   2. MVA in SPSS page 64
      1. This analysis will test if the missing values are at random or are correlated to some particular variable. You can test at whatever percent you desire (say 2% of missing data). You want a non-sig value for the MCAR test (similar to Levine’s or Mauchley’s) so that you can see that the probability of missing values non random would be greater than 5%.
   3. Deleting people with missing values or deleting variables
      1. If you only have a few missing pieces of data, it’s probably ok to just delete them – SPSS can do this for you automatically “pairwise and listwise”
      2. If it’s one variable that seems to be missing and it’s not crucial (like GPA), just delete it and move on.
      3. What if you have a lot of people though?
   4. Estimation methods for missing data
      1. Prior knowledge – using the median for income when people don’t want to list it. This happens when there are only a small number of missing cases and you have been working in a field for a while (or you can use established norms).
      2. Mean substitution – most popular way to input missing data. It’s fairly conservative – you aren’t going to change a significance test because you are using the mean value (so you aren’t changing the means, which is what is tested in a sig test). Does reduce the variance – so with a lot of cases may change the outcome of a test.
         1. SPSS will do this for you but it replaces the grand mean of that variable as the number.
      3. Regression – uses the data with complete information to predict the value in the missing slot and inserts that value. Becoming more popular since some computer programs will do this for you automatically SPSS MVA and is a little more theoretically logical than mean substitution. However, again it reduces variance and can be spurious if the variables using to predict are not good.
      4. Expected maximization – this procedure is a little more complex and a little better at missing data. It creates an expected curve for the data presented. Then (using matrix algebra), the program estimates the probability of each value and chooses the value for this person based on those probabilities. Use SEM programs or SPSS MVA.
      5. Multiple imputation – used for dichotomous variables – uses log regression to estimate the likelihood of group membership and picks the one with the higher likelihood. Creates several equations and tests several times to see which value is best (hence the multiple part). NORM free on the internet, SAS
   5. Missing data as data – create a dummy variable for those with data and those without. Maybe they are different on some other value that might be interesting.
   6. Repeat analysis with and without data – always recommended (especially for reviewer number 2).
   7. What’s best? Depends on the data set
      1. Delete cases – very few missing at random
      2. Mean substitution and prior knowledge – going by the wayside…only for small numbers of missing data
      3. EM – probably easiest and respected.
      4. Multiple imputation – most respected (although personally I haven’t seen it).
4. Outliers
   1. Definition – case with extreme value on one variable or multiple variables.(picture)
   2. Why?
      1. Data input error
      2. Missing values as “9999”
      3. Someone not from the population you intended to sample
      4. They are from the population but it has really long tails and very extreme values (more variance) than you expected
   3. Finding them
      1. Univariate – (only one variable weird value)
         1. Easy to spot when you do a frequencies analysis
         2. If data is grouped – search for outliers in each group separately (split file)
         3. Create z-scores and check if they are more than 3sds from the group
         4. spss explore or descriptives, can create normal curve charts as well
            1. box plots, histograms, normal probability plots
      2. Multivariate (weird value on multiple variables
         1. Mahalanobis distance – distance of a case from the centroid of rest of cases. Centroid is created by plotting the 3d means of all the means.
         2. Rule of thumb: chi-square cut off for df and then use p<.001
         3. Regression rules – leverage, discrepancy and influence
            1. Leverage – pg 75 pictures – far out on a line, but don’t influence the slope of a regression
            2. Discrepancy – is how much something will affect the slope – how different it is from the line
            3. Influence – is the product of leverage and discrepancy
            4. How to: create a fake regression and ask for leverage, mahalanobis and cooks (for influence)
   4. What do I do with them?
      1. Are these people part of the population you intended to sample from? Did they do the study correctly? If not, just eliminate them.
      2. Transforms – you can transform a variable that is extremely non-normal or has big tails (such as reaction times, use a log transform)
      3. Change to the score to the last furthest out score
5. Normality, Linearity, Homoscedasticity
   1. Multivariate normality – each variable and all linear combinations of variables are normal
      1. With the central limit theorem – 20dfs f-test is robust to normality
      2. Skewness – symmetry of a distribution – skewed = mean not in middle
      3. Kurtosis – peakedness of distribution – tall and skinny or fat and short?
      4. Page 80 picture
      5. SPSS frequencies – will give you statistical tests on skew and kurtosis
      6. PPlots – scores are ranked and sorted – and a z score “normal value is calculated”. Then the normal value is plotted against the real value. You can also use this to check for outliers, you don’t want anything over 3sds. You want things to fall on a normal line.
   2. Linearity – assumption that there is a straight line relationship between two variables (or the combination of all the variables)
      1. Check for linearity by looking at bivariate scatter plots (make sure they aren’t curved or x power function)
   3. Homoscedasticity – the spread of the variance of a variable is the same across all values of the other variable.
      1. (when you have grouped data – often called homogeneity). Levine’s = suck. Box’s M = suck.
      2. You’ll want to check for this and transform if it is a problem.
      3. You can do that by looking at the bivariate scatter plots (or a residual plot).
6. Data transforms
   1. Page 87 picture – page 89 syntax
   2. Most common types – square root, log, inverse
   3. When you transform – you have to make sure you explain your output as a transform. If you square root reaction time, you cannot talk about the new mean as the “reaction time” it’s not the square root of the reaction time.
7. Multicollinearity and Singularity
   1. Mostly used when doing regression or regression like analyses (PCA, EFA).
   2. Multicollinearity – when two variables used in an analysis are too highly correlated (over r-.9 normally).
      1. Why do you care? Well every statistical analysis measures the variance right? When two variables are that highly correlated, somebody is hogging the variance from the other one and the stats programs change their mind all the time who gets the variance. Also, they basically say the same information – just use one of them.
   3. Singularity – redundant information. For example, when you use both the total and all the numbers to make the total in a regression.
      1. Why? Again, you don’t need both pieces of information. Also, most stats programs will crash if you do this. It will give you a “matrix is singular” error.
   4. If you are running regression like analyses – you can get tolerance values. Tolerance is 1-R2, so you want bigger values (nothing less than .10). You can also just check the R2 values – you want nothing over .9.
8. Checklist
   1. Accuracy
      1. Out of range
      2. Means and sds
      3. Univariate outliers
   2. Missing data
   3. Linearity and Homoscedasticity
   4. Non-normal data and outliers
      1. Skew and kurtosis, p plots
      2. Transforms
   5. Multivariate outliers
      1. Eliminate or not
   6. Multicollinearity and singularity