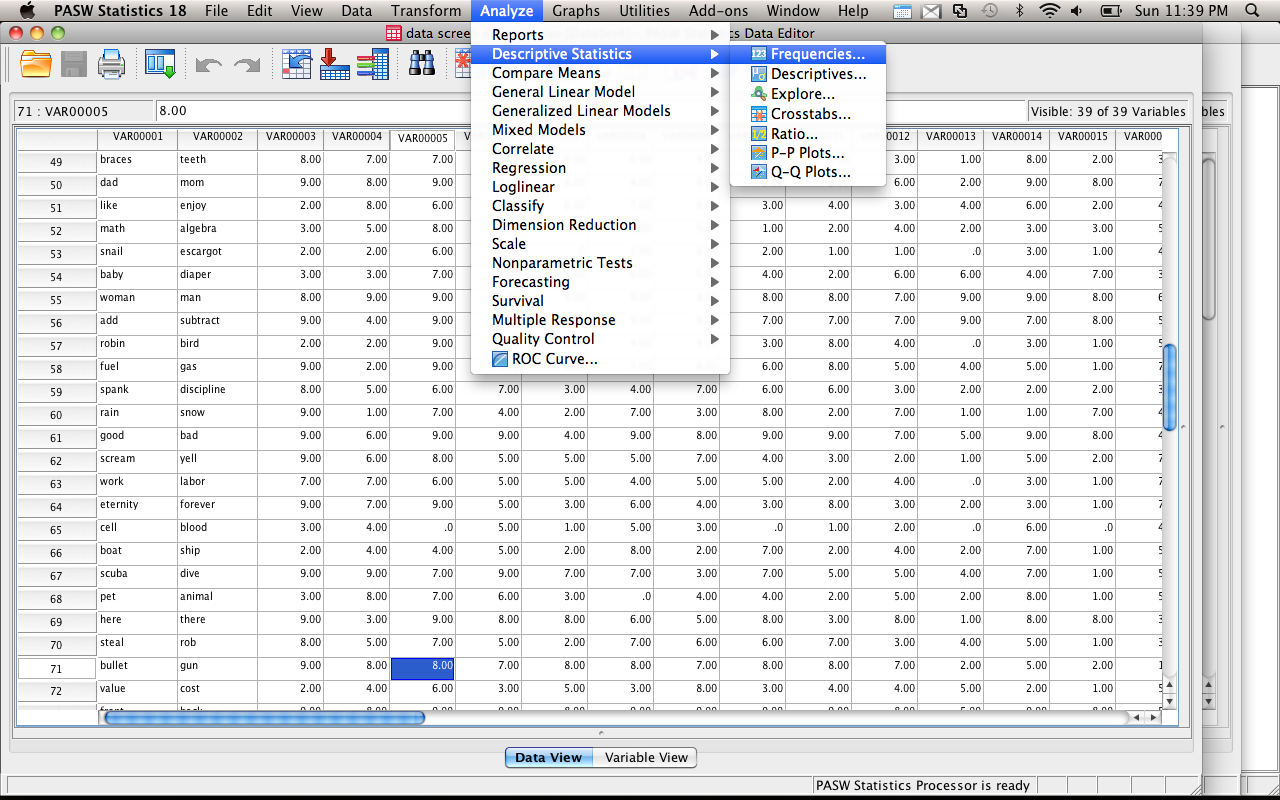
Data Screening

**Description:**

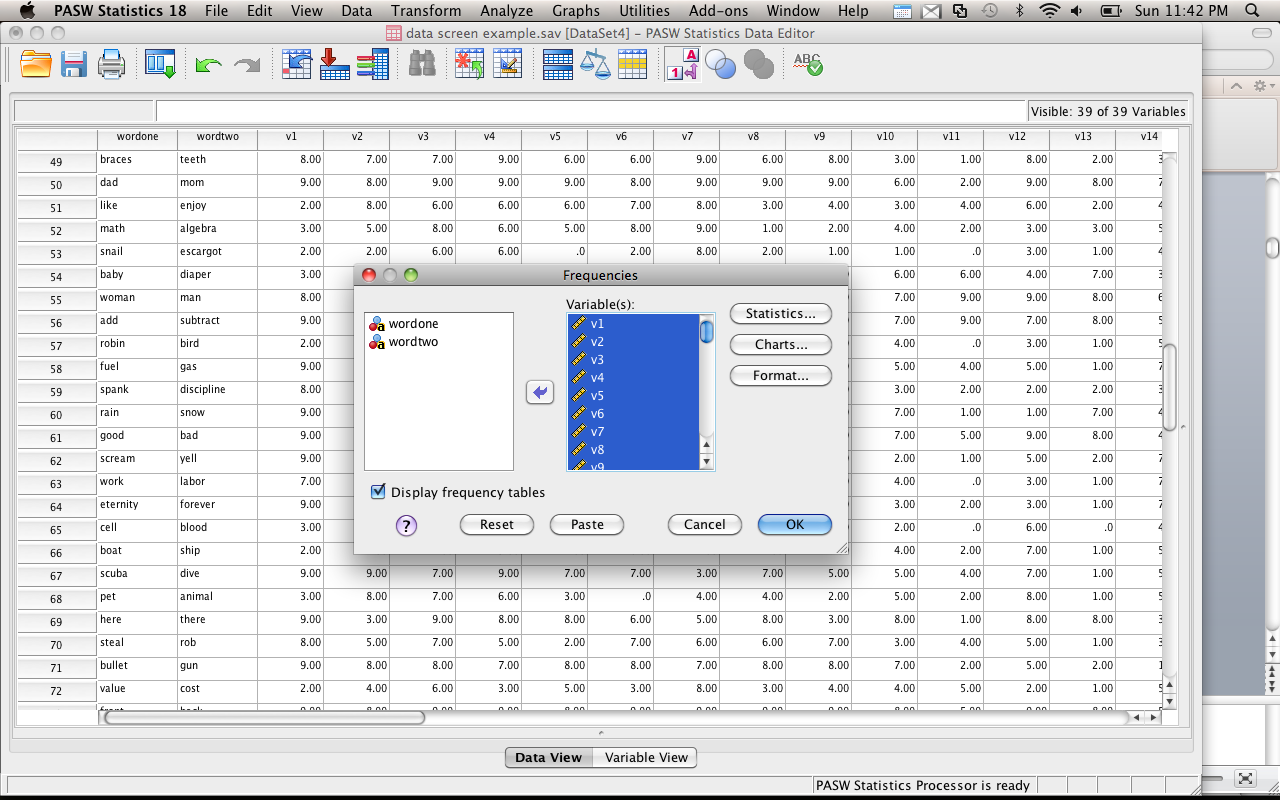
Data screening is very important to make sure you’ve meet all your assumptions, outliers, and error problems. Each type of analysis will have different types of data screening. This guide lists all the types, and check out the individual analysis for the important ones.

**Walk through/things to check:**

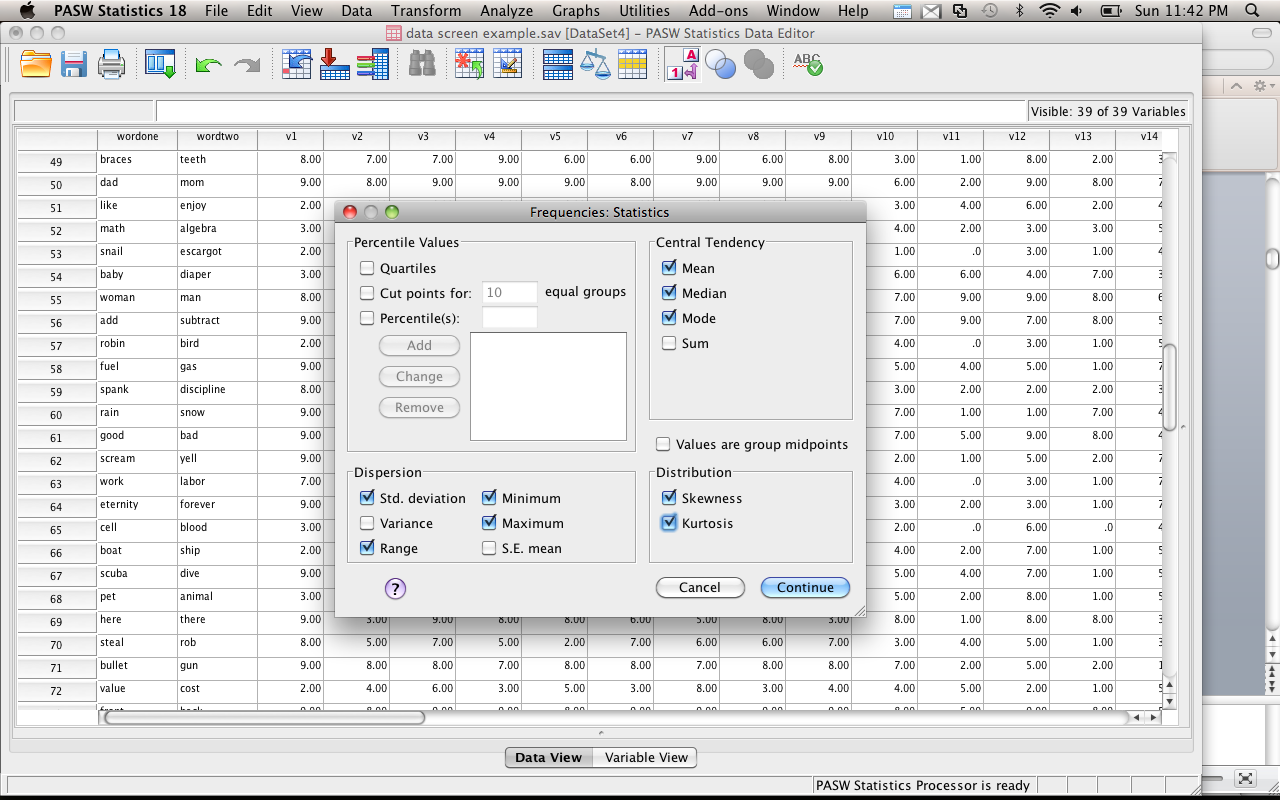
1. Accuracy
   1. Check for typos and problems with the dataset
   2. Frequencies – you can see if there are numbers you aren’t expecting
   3. Ask for Min, Max, Means, SD, Missing Values
   4. Analyze > Descriptive Statistics > Frequencies



* 1. Move all the variables you need to check into the variables box.

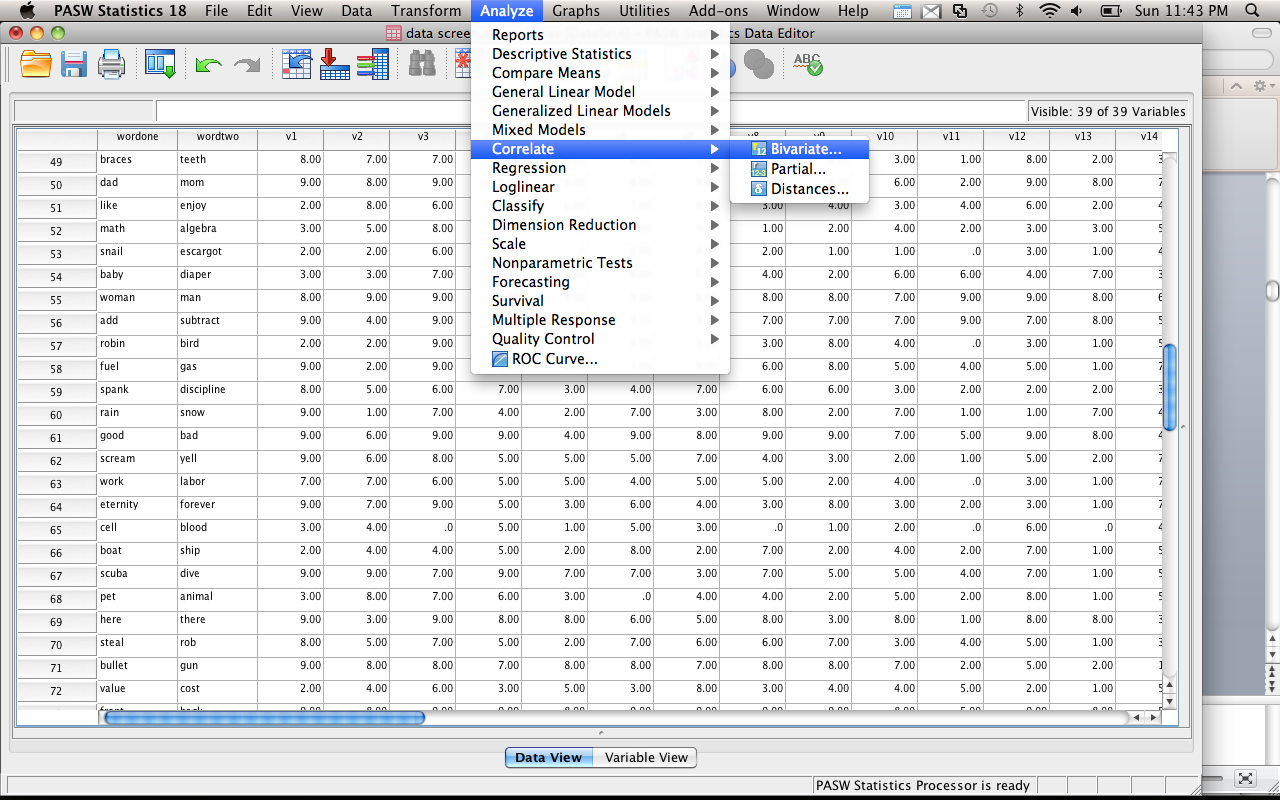


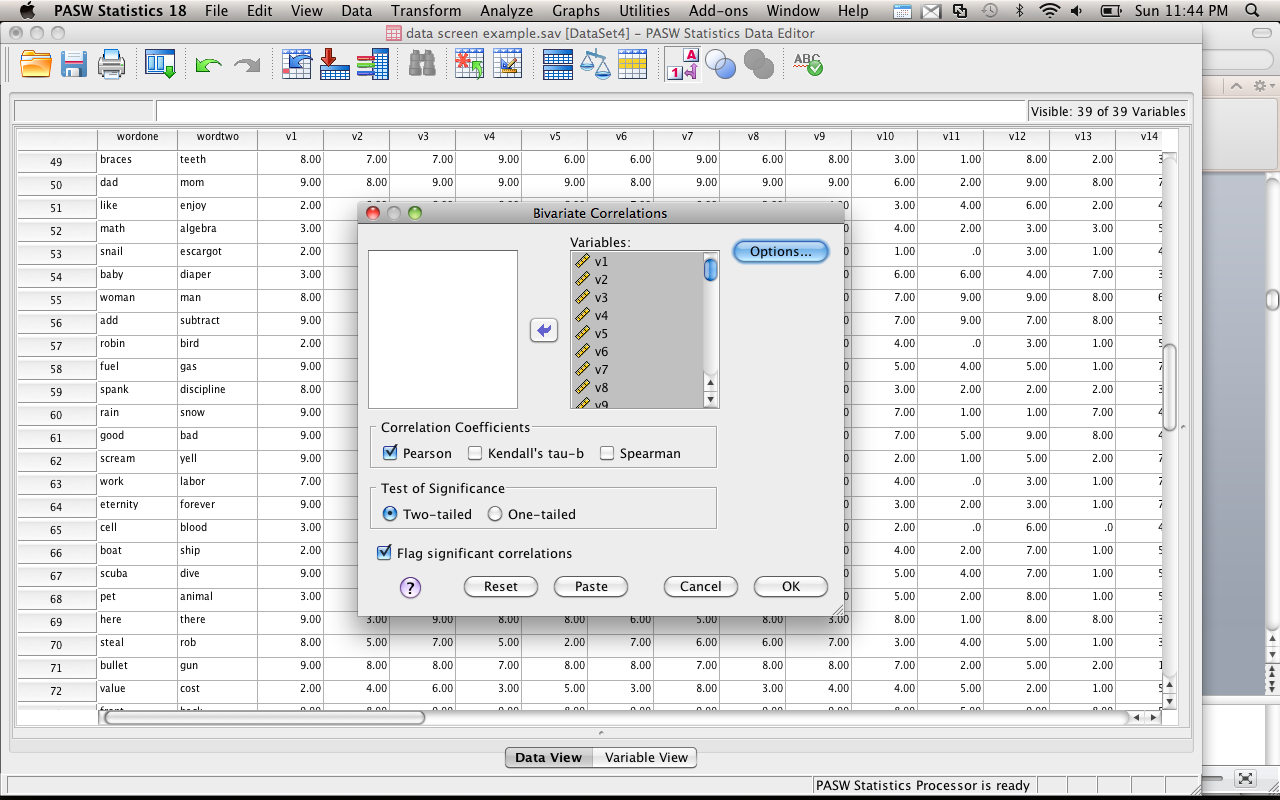
* 1. Click statistics and check all the appropriate boxes.



| **Statistics** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | v1 | v2 | v3 | | v4 | v5 | |
| N | Valid | 96 | 96 | 96 | | 96 | 96 | |
| Missing | 0 | 0 | 0 | | 0 | 0 | |
| Mean | | 5.9063 | 5.0833 | 6.5729 | | 6.4375 | 4.1667 | |
| Median | | 7.0000 | 5.0000 | 7.0000 | | 6.0000 | 3.0000 | |
| Mode | | 9.00 | 6.00a | 7.00 | | 9.00 | 2.00 | |
| Std. Deviation | | 3.04035 | 2.18447 | 1.90149 | | 2.02517 | 2.67017 | |
| Skewness | | -.315 | -.158 | -.734 | | -.191 | .427 | |
| Std. Error of Skewness | | .246 | .246 | .246 | | .246 | .246 | |
| Kurtosis | | -1.666 | -1.014 | .440 | | -.770 | -1.183 | |
| Std. Error of Kurtosis | | .488 | .488 | .488 | | .488 | .488 | |
| Range | | 8.00 | 8.00 | 9.00 | | 8.00 | 9.00 | |
| Minimum | | 1.00 | 1.00 | .00 | | 1.00 | .00 | |
| Maximum | | 9.00 | 9.00 | 9.00 | | 9.00 | 9.00 | |
| a. Multiple modes exist. The smallest value is shown | | | | | | | | |
| **v1** | | | | | | | | | |
|  | | Frequency | Percent | | Valid Percent | | | Cumulative Percent | |
| Valid | 1.00 | 5 | 5.2 | | 5.2 | | | 5.2 | |
| 2.00 | 16 | 16.7 | | 16.7 | | | 21.9 | |
| 3.00 | 14 | 14.6 | | 14.6 | | | 36.5 | |
| 4.00 | 4 | 4.2 | | 4.2 | | | 40.6 | |
| 5.00 | 2 | 2.1 | | 2.1 | | | 42.7 | |
| 6.00 | 2 | 2.1 | | 2.1 | | | 44.8 | |
| 7.00 | 6 | 6.3 | | 6.3 | | | 51.0 | |
| 8.00 | 15 | 15.6 | | 15.6 | | | 66.7 | |
| 9.00 | 32 | 33.3 | | 33.3 | | | 100.0 | |
| Total | 96 | 100.0 | | 100.0 | | |  | |

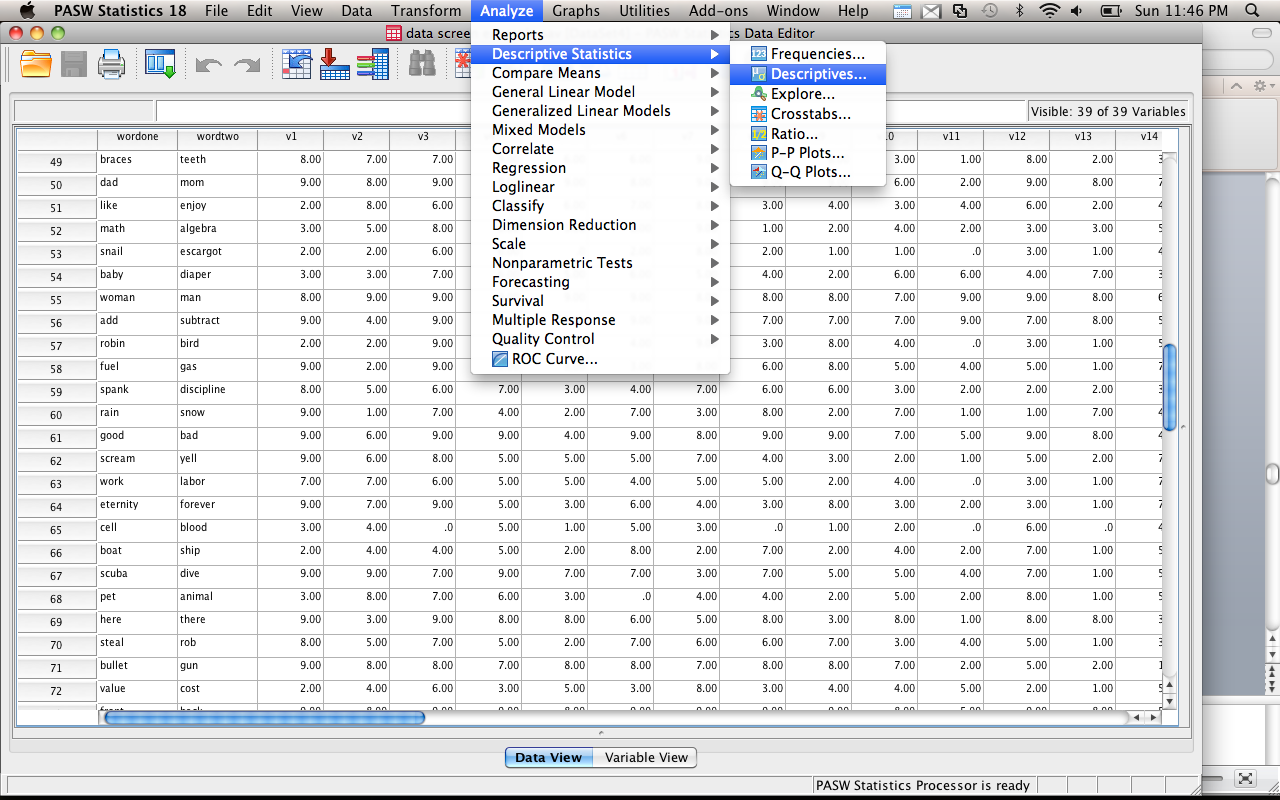
1. Missing data
   1. You can also see the missing data using this analysis. Check for “missing” in the frequency column (it will say Missing System) or under Missing in the Statistics box.
   2. MCAR – missing completely at random (you want this)
      1. MCAR – probably caused by skipping a question or missing a trial.
   3. MNAR – missing not at random (eek!)
      1. MNAR – may be the question that’s causing a problem.
      2. For instance, what if you surveyed campus about alcohol abuse? What does it mean if everyone skips the same question?
   4. How much can I have?
      1. Depends on your sample size – in large datasets <5% is ok.
      2. Small samples = you may need to collect more data.
   5. How do I check if it’s going to be a big deal?
      1. Frequencies – you can see which variables have the missing data.
      2. Sample test – you can code people into two groups. Test the people with missing data against those who don’t have missing data.
      3. Regular analysis – you can also try dropping the people with missing data and see if you get the same results as your regular analysis with the missing data.
   6. Deleting people / variables
      1. You can exclude people “pairwise” or “listwise”
      2. Pairwise – only excludes people when they have missing values for that analysis
      3. Listwise – excludes them for all analyses
      4. Variables – if it’s just an extraneous variable (like GPA) you can just delete the variable
   7. What if you don’t want to delete people (using special people or can’t get others)?
      1. Prior knowledge – if there is an obvious value for missing data
         1. Such as the median income when people don’t list it
         2. You have been working in the field for a while
         3. Small number of missing cases
      2. Mean substitution – fairly popular way to enter missing data
         1. Conservative – doesn’t change the mean values used to find significant differences
         2. Does change the variance, which may cause significance tests to change with a lot of missing data
         3. SPSS will do this substitution with the grand mean
      3. Regression – uses the data given and estimates the missing values
         1. This analysis is becoming more popular since a computer will do it for you.
         2. More theoretically driven than mean substitution
         3. Reduces variance
      4. Expected maximization – now considered the best at replacing missing data
         1. Creates an expected values set for each missing point
         2. Using matrix algebra, the program estimates the probably of each value and picks the highest one
      5. Multiple Imputation – for dichotomous variables, uses log regression similar to regular regression to predict which category a case should go into
      6. Use it as data – create a dummy variable and compare people who have missing values and those who don’t.
         1. May give you interesting results
2. Correlations
   1. Useful when you are doing a regression analysis or scales with total scores
   2. Inflated correlations – you can have super high correlations when you don’t mean to if you accidently use the same scale twice
   3. Deflated correlations – helps check for restriction of range.
   4. Analyze > correlate > bivariate
   5. Move variable over, hit ok.

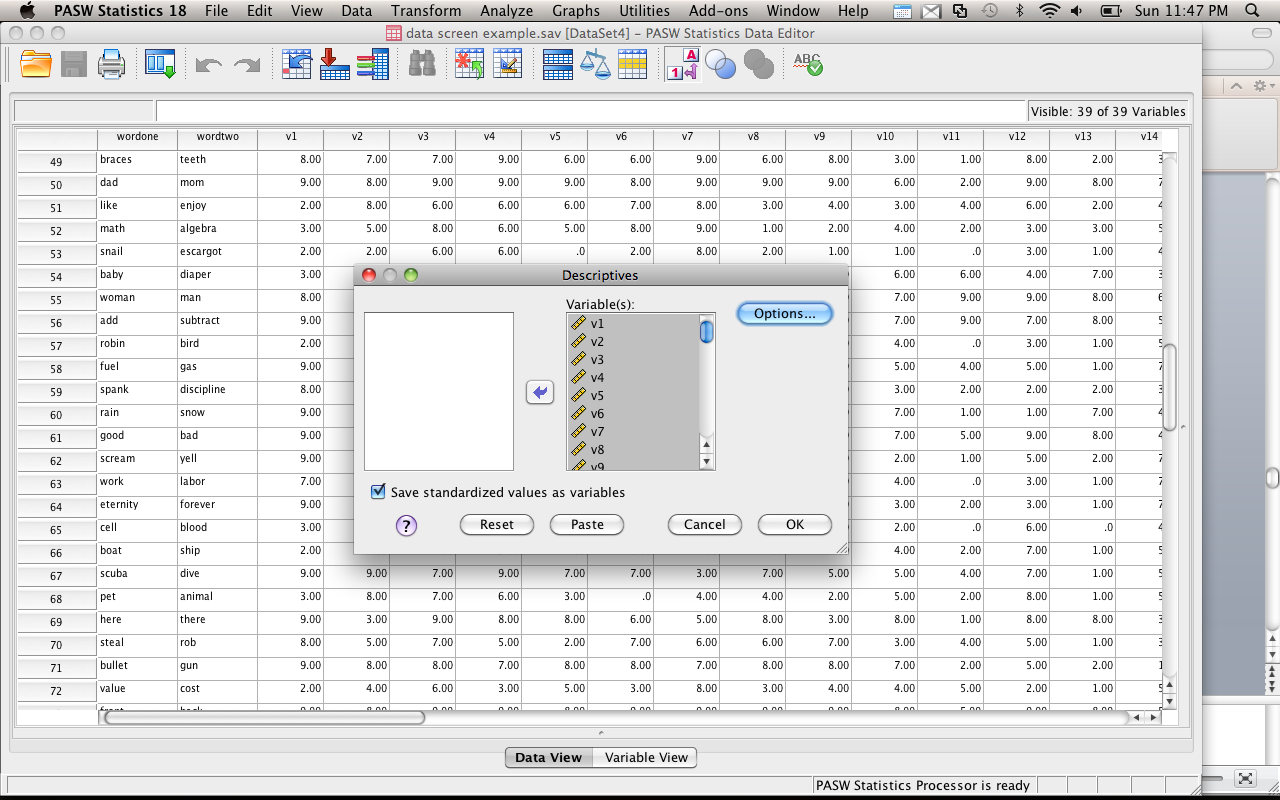


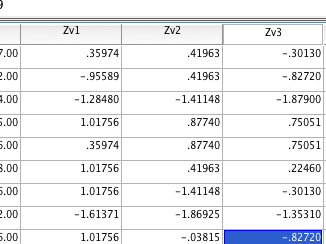


* 1. Check for correlations that are very high (r > .90) or very low (r<.10). Low correlations will be normal, but very high correlations might give you problems in regression analyses.

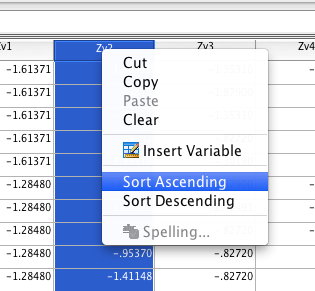
1. Outliers
   1. Case with extreme value on one variable or multiple variables
   2. Univariate
      1. Frequency analysis
      2. Grouped data – look for outliers in each group, not as a whole
      3. Z-scores – create z-scores for each group and check if they are more than 3 standard deviations
      4. SPSS descriptives – box plots, histogram, normal probability plots
      5. How to check:
         1. Analyze > descriptive statistics > descriptive
         2. Move all the variables over
         3. Be sure to check “save standardized values as variables”

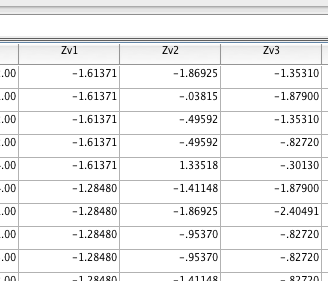






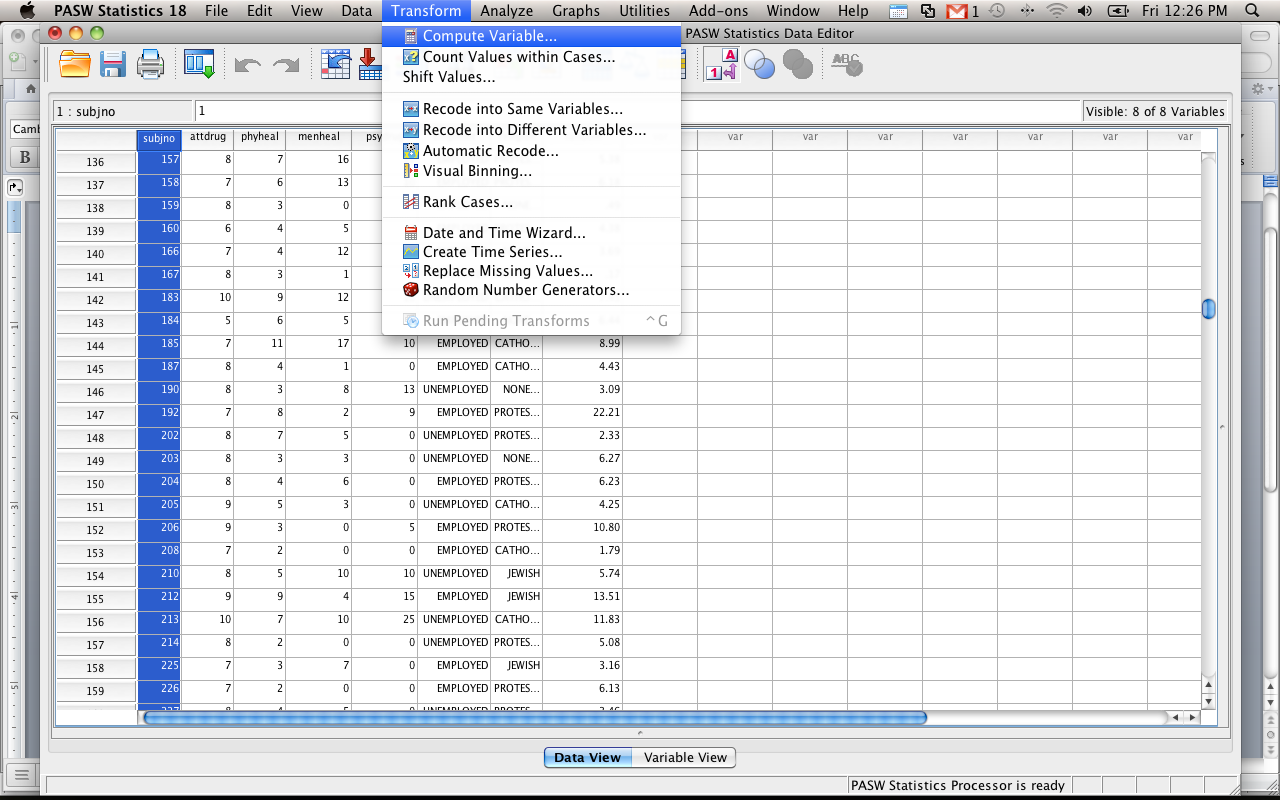
* + 1. It will create these new “z” variables.
    2. Right click > sort ascending
    3. Check for z-scores over +3 or under -3.





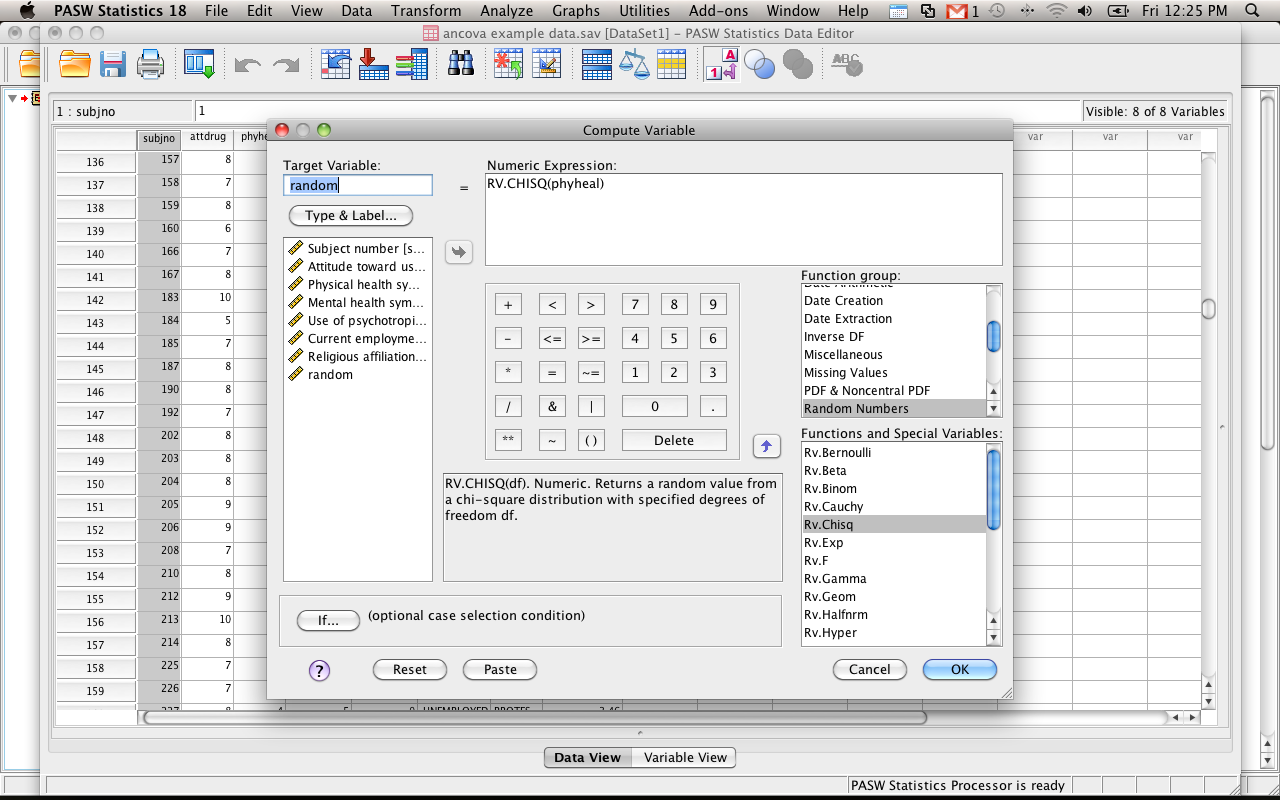
* 1. Multivariate
     1. Mahalanobis distance – distance of a case from the centriod of rest of cases.
     2. Centriod is created by plotting 3D means of all the means
     3. Rule of thumb: chi-square cue off for degrees of freedom and use p<.001
        1. Use table linked online.
     4. *This is the number one thing to check!*
  2. Regression rules
     1. Leverage – far out on line but doesn’t influence regression slopes
     2. Discrepancy – how much something will affect the slope (how different it is from the line)
     3. Influence – product of leverage and discrepancy
     4. See notes in regression section for specific rules on these – they only apply to regression.
  3. How to check for these:
     1. Create a fake regression and ask for leverage, Mahalanobis, and cooks (for influences)
  4. What do I do with them?
     1. Did they do the study correctly?
     2. Are they part of the population you wanted?
     3. Eliminate them
     4. Transforms
     5. Change the outlier to the last value
  5. How to check for multivariate outliers
     1. First, make a random variable.

Hit Transform > compute:



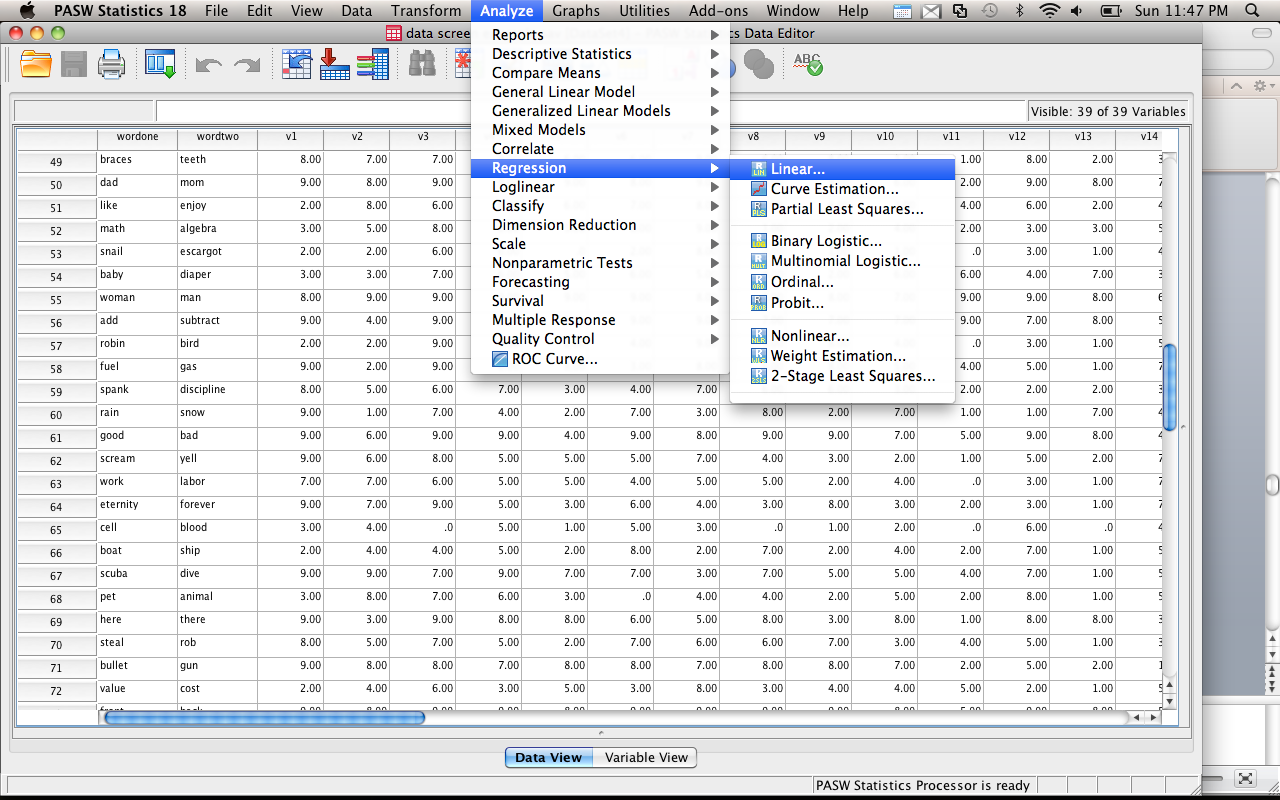
Name your random variable:

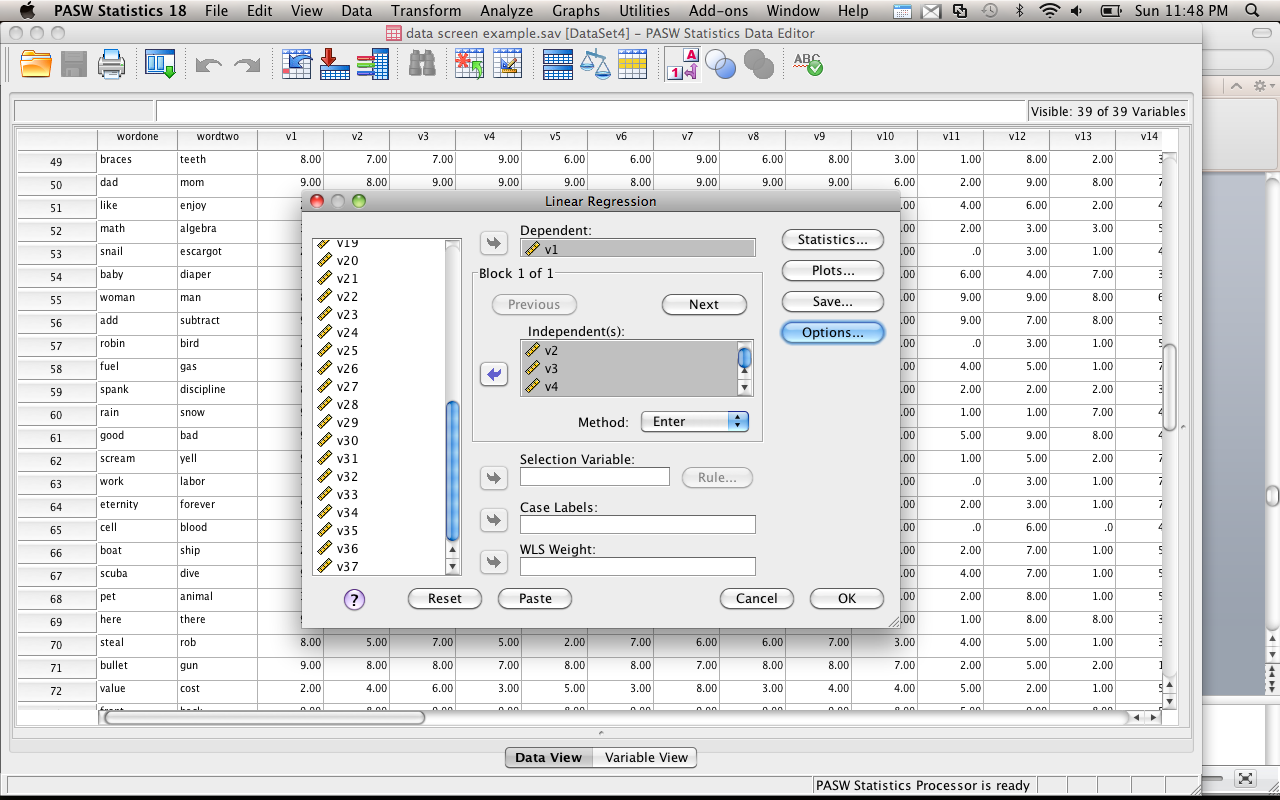
Find a way to compute the random variable (RV.Chisq(pick a number here).

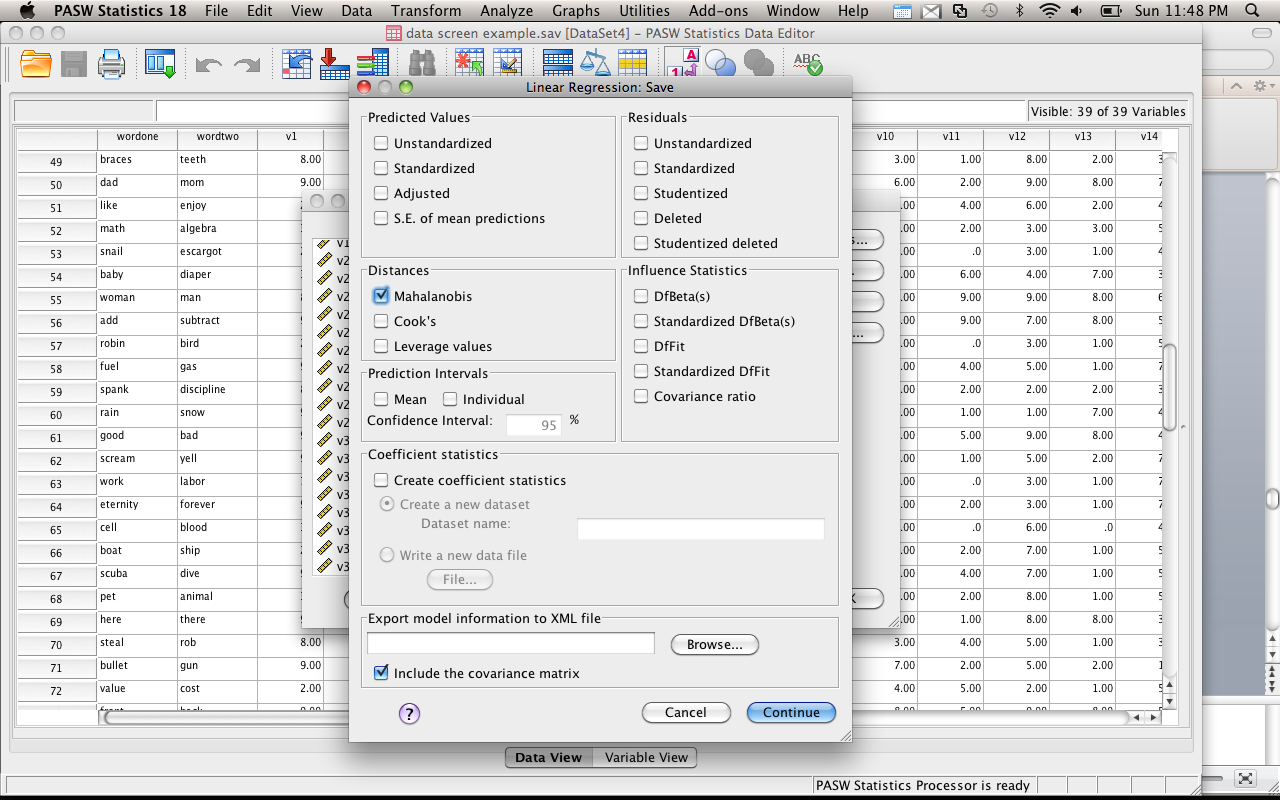


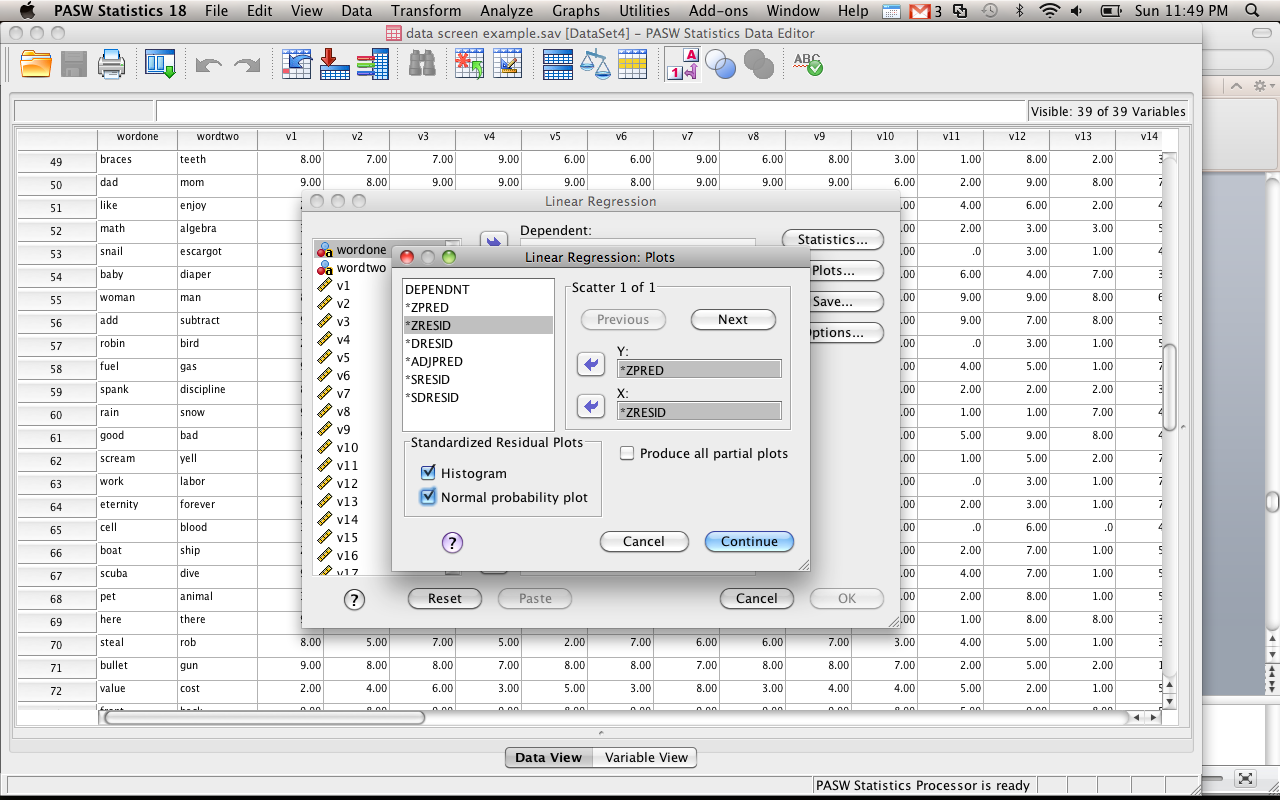
Hit ok. Then use that variable as your DV in your fake regression.

* + 1. Now run a fake regression.
    2. Analyze > regression > linear.
    3. Move the new “fake variable” into the dependent box (not quite what’s pictured here).
    4. Move all the independent variables into the independents box.
    5. Click options.
    6. Click Mahalanobis distance.
    7. Click plots (you will use this for the next analyses below).
    8. Under Y > ZPRED.
    9. Under X > ZRESID.
    10. Click Histogram and Normal Probability Plot.

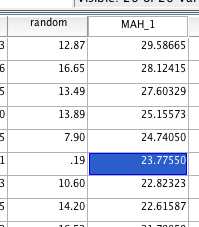






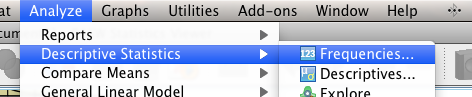


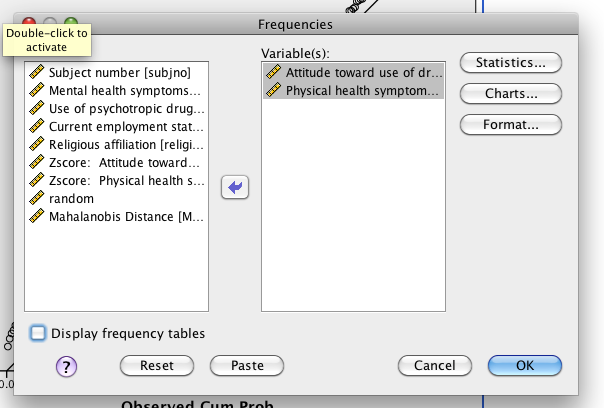
* + 1. You will get a new column with Mahalanobis distance.

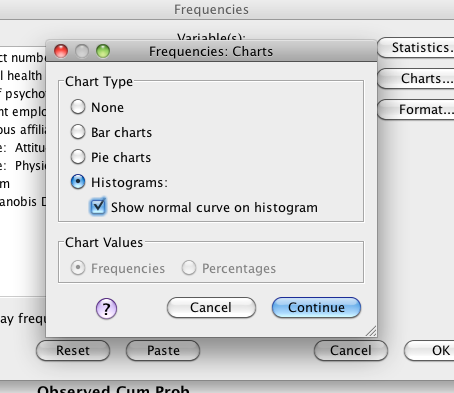


* + 1. Figure out the cut off score for Mahalanobis. You will want to use a Chi-Square table for cut off scores. Degrees of Freedom are the number of variables (here we have 11 variables for example). You will use the p<.001 value for cut off scores. This example has a cut off score of 31.26. Therefore, none of the people in this dataset are outliers (because the highest score is 29.58).

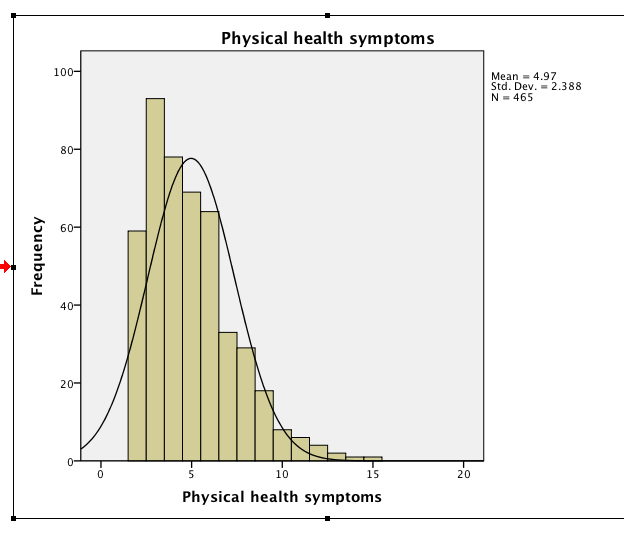
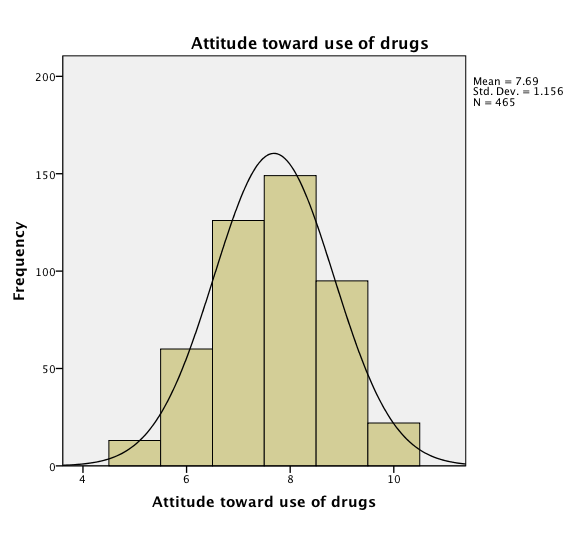
1. Normality
   1. Multivariate normality – each variable and all linear combinations of variables are normal
      1. Given the Central Limit Theorem – with 20dfs tests are robust
      2. P Plots – you can look for values not on the normal line (you got this chart with the fake regression
   2. Univariate normality
      1. Skewness – symmetry of a distribution (see above in the frequency chart)
         1. Skewed – mean not in the middle
      2. Kurtosis – peakedness of a distribution (see above in the frequency chart)
         1. Tall and skinny or fat and short
      3. Analyze > descriptive statistics > frequencies – when you run this analysis hit charts, click histograms > normal curve.

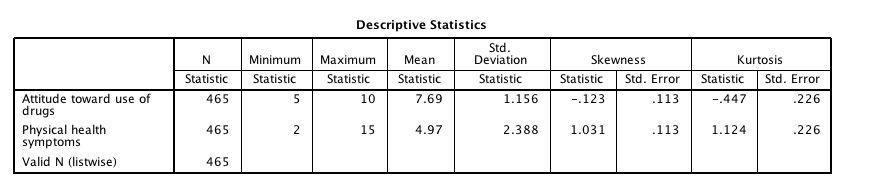


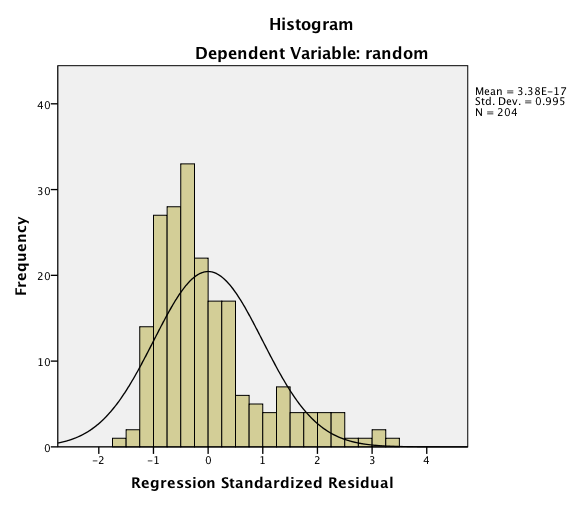




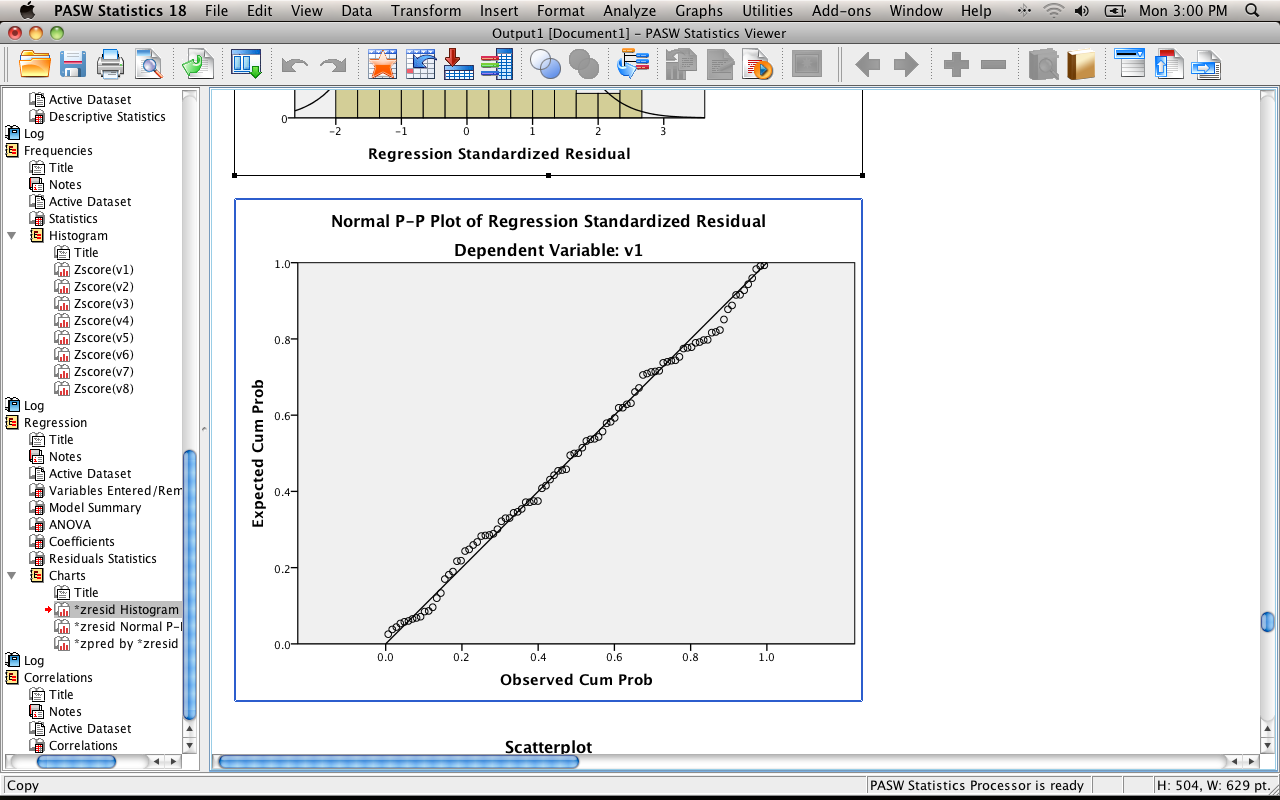
* + 1. Check pretty charts – this is univariate normality.



* + 1. Check frequency charts for skew and kurtosis values over +/- 3.
    2. When you run the regression analysis and ask for the normality chart you will get one for the multivariate combination.



1. Linearity
   1. Assumption that there is a straight line relationship between two variables (or the combination of all the variables)
   2. Look at bivariate scatterplots to check for curves
   3. Look at the p-plots for rainbows! Or this plot to make sure the dots follow the line (see below).



1. Homoscedasticity
   1. Spread of the variance of a variable is the same across all values of the other variable
   2. The test for this is about as good as Levine’s = sucks.
   3. Best way to check is by looking at scatterplots.
2. Homogeneity
   1. Equal variances – you can look at the descriptive statistics and just look to see if they look roughly the same
   2. Or you can check out the residual plot
   3. No raining
3. All of these can be checked with the p-plot or by checking out scatter plots.
4. What to look for:
   1. Draw a line at 0.
   2. Homogeneity – is the spread above that line the same as below that line?
      1. These lines are complete lines.
      2. So it goes from 0 to 3 above and 0 to -3 below.
      3. You *do not* want a very large spread on one side and a small spread on the other side.
      4. If you encounter this problem – check Levene’s when you run the data.
   3. Homoscedasticity – is the spread equal all the way across the zero line?
      1. These lines are the dashed lines.
      2. Look for megaphones.
   4. Linearity – no rainbows on this graph.

