Data Screening

**Description:**

Data screening is very important to make sure you’ve met 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.

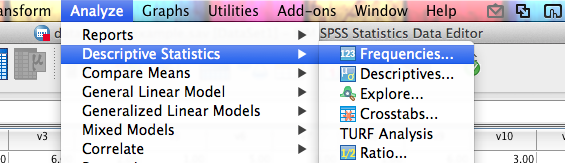
BIG IMPORTANT RULE:

**Hypothesis testing:** Traditionally we use *p* < .05 (so you want values less than .05 because that’s what you are trying to find … statistically significant).

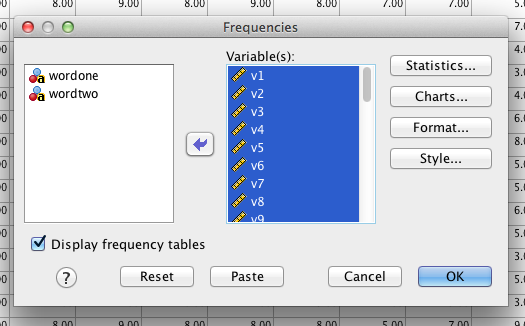
**Data screening:** We want to use *p* < .001 because we want to make sure things are really crazy before we fix/delete/etc. If the scores are less than .1% then it’s really, really different/skewed/what have you, so you would want to fix it.

**Walk through/things to check (in this order):**

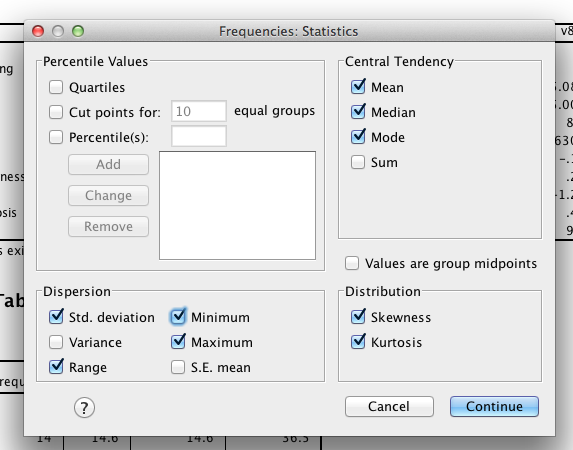
1. Accuracy
   1. Check for typos and problems with the dataset
   2. Frequencies – you can see if there are numbers you aren’t expecting
   3. Ask for Min, Max, Means, SD, Missing Values
   4. Analyze > Descriptive Statistics > Frequencies

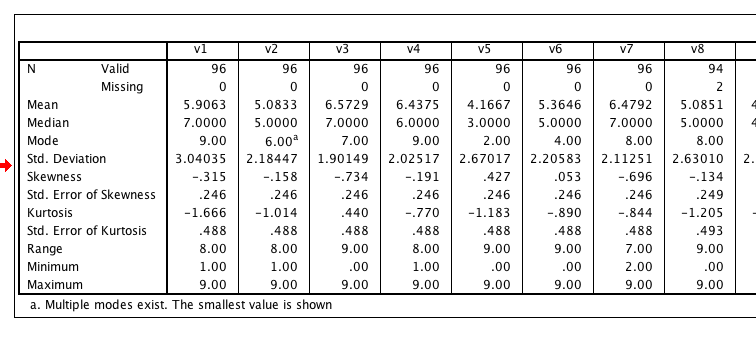


* 1. Move all the variables you need to check into the variables box.



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

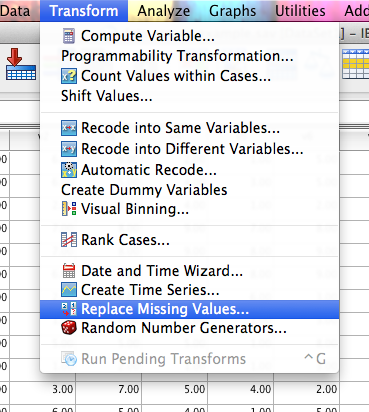




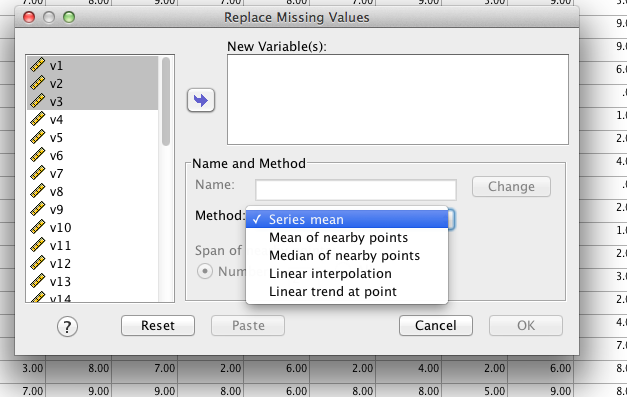
Important information:

* Look at missing data.
* Look at the means for anything you wouldn’t expect … like v5 seems low in comparison to the rest. This information will be based on the study though, you might want that variable to be low.
* Standard deviations are useful to think about … this scale ranges from 0 – 9 so the scores range around the means 2-3 points. You would need to think about how big the range of the data is – larger range variables will have larger standard deviations. So you can look for very small SDs (not enough variance in your data) or for very large SDs (lots of error in your data or maybe outliers).
  + In this example, v1 might be large or v3 might be too small.
* Skewness – how far the data “leans” from normal. Remember that the positive / negative aspect tells you where the extreme scores are (the tail of the distribution).
* Kurtosis – how flat/skinny the distribution is from normal.
  + These values are Z scores, so the cut off score is 3 (which is *p* < .001 for Z scores).
  + You do NOT want scores larger than |3|, which would indicate very extreme skew/kurtosis.
  + Larger sample sizes (*N* > 30) tend to be more normal, and you do not have to worry about skew/kurtosis as much.
  + What to do if they are bad?
    - You *could* use a transform, but that tends to just make things more difficult to interpret.
    - You could use larger samples to help get around the problem.
    - Check out outliers and other assumptions to see those are the problem.
* Minimum / maximum values – great place to look for *accuracy* issues. You want to make sure that values are not out of range (i.e. 10 values for a 0-9 scale).
  + What if you have typos?
    - Fix’em! Or delete that *data point*. Do not delete the whole person, just the wrong data point.
* You can also look at the frequency tables and histograms for values you weren’t expecting or why you got large skew/kurtosis values.

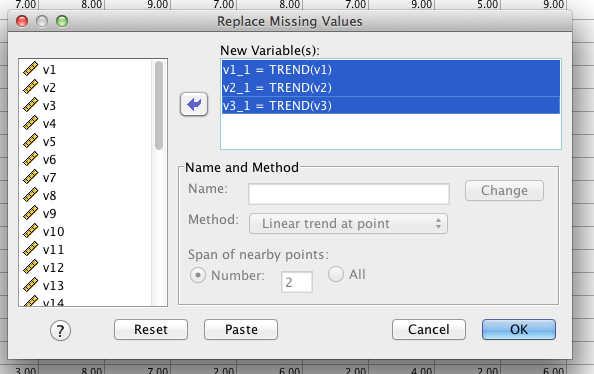
1. Missing data
   1. You can also see the missing data using descriptives > frequencies. Check for “missing” in the frequency column (it will say Missing System) or under Missing in the Statistics box.
   2. MCAR – missing completely at random (you want this)
      1. MCAR – probably caused by skipping 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 replace?
      1. Depends on your sample size – in large datasets <5% is ok.
      2. The 5% rule applies either across (by participant – how much data are they missing?) or down (by variable – how much of that variable is missing).
      3. Small samples = you may need to collect more data.
   5. What should I replace?
      1. Do not replace categorical variables.
      2. Do not replace demographic variables.
      3. Most people replace continuous variables (interval / ratio) … one thing to watch in SPSS, the sorting matters … sometimes it will not replace if it’s at the top (i.e. first row), and sometimes it will replace values that are outside of range (doh).
   6. 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.
   7. 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
   8. 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 – used to be a popular way to fill in data (cuz it was easy, but now most people don’t recommend this solution).
         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. Mean substitution will fill in the missing data based on the column average (grand mean of the column).
      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. Linear trend at point option – replaces data based on the row average and column average together.
      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
   9. How to fill in missing data:
      1. Transform > replace missing values



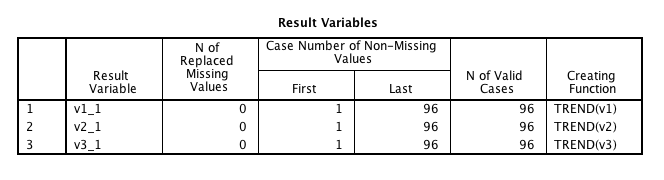
* + 1. Before you move variables over to the right, change the drop down type or you will automatically get series mean.



* + 1. Linear trend at point is a regression type analysis.



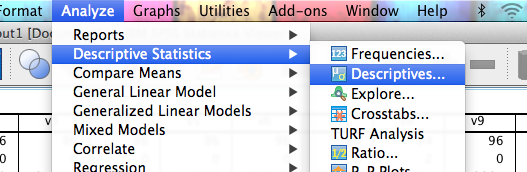
SPSS will then create you new columns with values replaced, and give you this output.

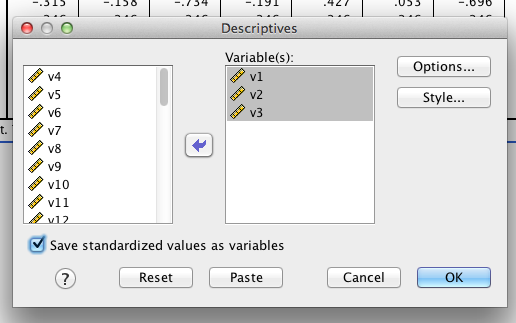


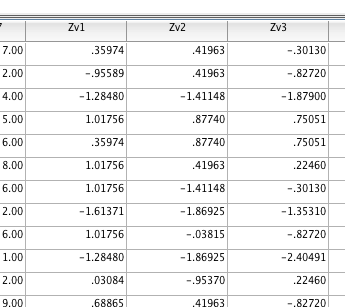
(normally you’d have something in N column, I just didn’t have any in this example).

Here’s the catch. The original columns are in the dataset. If you use those in the analyses that follow, you’ve basically wasted your time replacing because you forgot to use the new fixed columns. DOH. So, I recommend creating a new dataset and deleting the old columns out of the new one. That way you have the original columns in the old dataset, but in the one you are working in you aren’t tempted to use the wrong column.

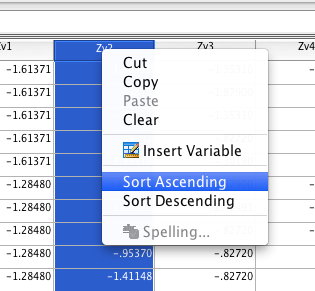
1. Outliers
   1. Case with extreme value on one variable or multiple variables
   2. Univariate
      1. Grouped data – look for outliers in each group, not as a whole
      2. Z-scores – create z-scores for each group and check if they are more than 3 standard deviations (same rule as above Z scores of 3 are *p* < .001).
      3. SPSS descriptives – box plots, histogram, normal probability plots
      4. How to check:
         1. Analyze > descriptive statistics > descriptive
         2. Move all the variables over
         3. Be sure to check “save standardized values as variables”

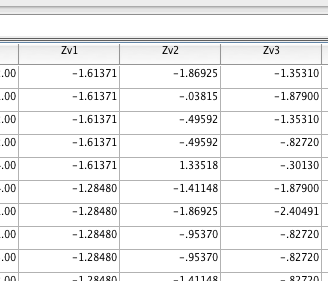






* + 1. It will create these new “z” variables.
    2. Right click > sort ascending
    3. Check for z-scores over +3 or under -3.



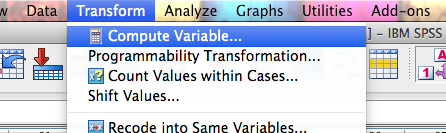


* 1. Multivariate
     1. Mahalanobis distance – distance of a case from the centroid of rest of cases. Centroid is created by plotting the means of the all the variables (like an average of averages), and then seeing how far each person’s scores are from the middle.
     2. How to check:
        1. Create Mahalanobis scores (see below).
        2. Use the chi-square table to find the cut off score (anything past this score is an outlier).
        3. *df* = the number of variables you are testing (important: testing! Not all the variables!)
        4. Use the *p* < .001 value.

REMEMBER THE BIG IMPORTANT RULE.

* + 1. *This is the number one thing to check!*
  1. Regression rules (only check when running regression, we will go over this more later).
     1. Leverage – a score that is far out on line but doesn’t influence regression slopes (measured by leverage values).
     2. Discrepancy – a score that is far away from everyone else that affects the slope
     3. Influence – product of leverage and discrepancy (measured by Cook’s values).
     4. See notes in regression section for specific rules on these – they only apply to regression.
  2. Important note on checking for multivariate outliers (and the rest of everything below minus correlations):
     1. For ANOVA, t-tests, correlation: you will use a *fake* regression analyses – it’s considered fake because it’s not the real analysis, just a way to get the information you need to do data screening.
     2. For regression based tests: you can run the *real* regression analysis to get the same information. The rules are altered slightly, so make sure you make notes in the regression section on what’s different.
     3. Steps are below.
  3. What do I do with them when I find them?
     1. Ask yourself:
        1. Did they do the study correctly?
        2. Are they part of the population you wanted?
     2. Eliminate them
     3. Leave them in
     4. Windsorize: Change the outlier to the last value
  4. How to check for multivariate outliers
     1. First, make a random variable.

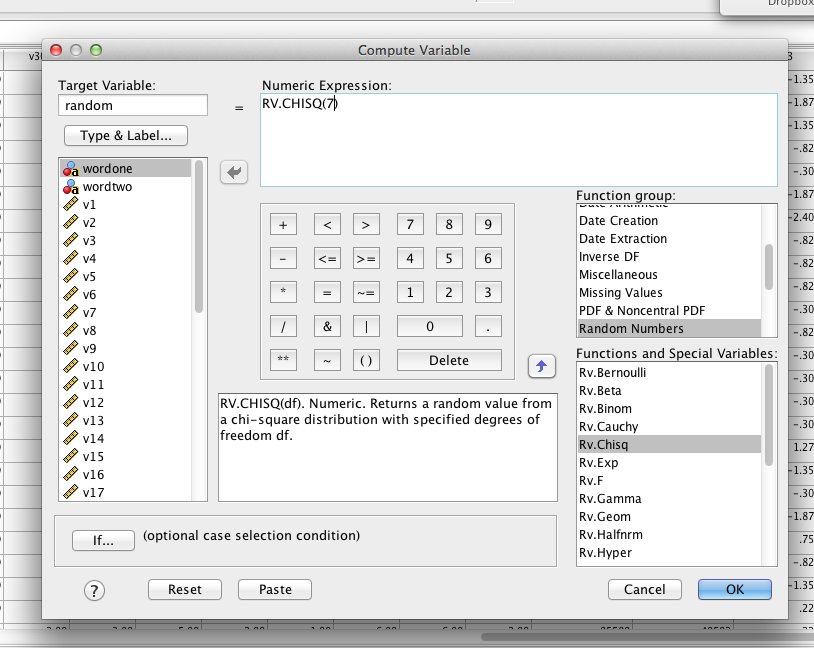
Hit Transform > compute:



Name your random variable:

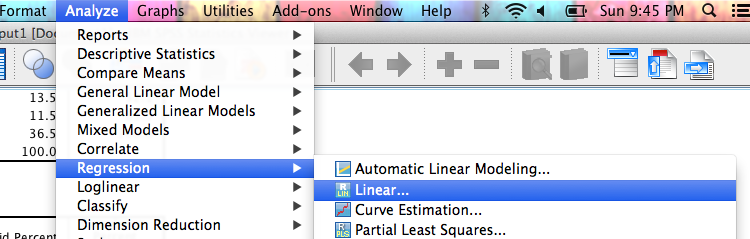
Find a way to compute the random variable (RV.Chisq(pick a number here).

Why chi-square? Because we are going to use the chi-square distribution to determine if those scores are outliers, so we are creating a random variable with the same distribution. In reality, the outlier analyses shouldn’t change based on the random variable, but the other assumption checks will, so you don’t want to use a variable that’s categorical.

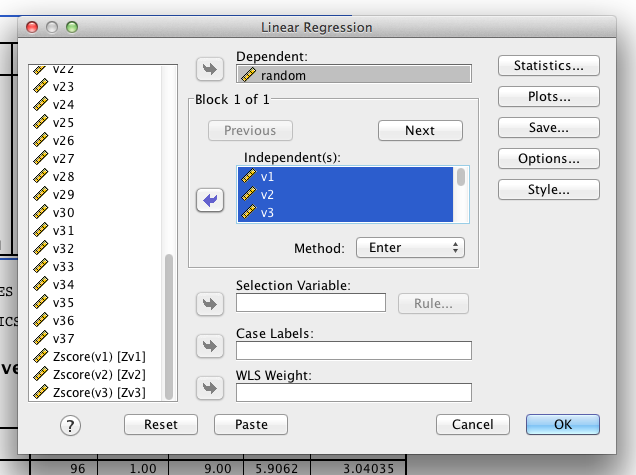


Hit ok. Then use that variable as your DV in your fake regression.

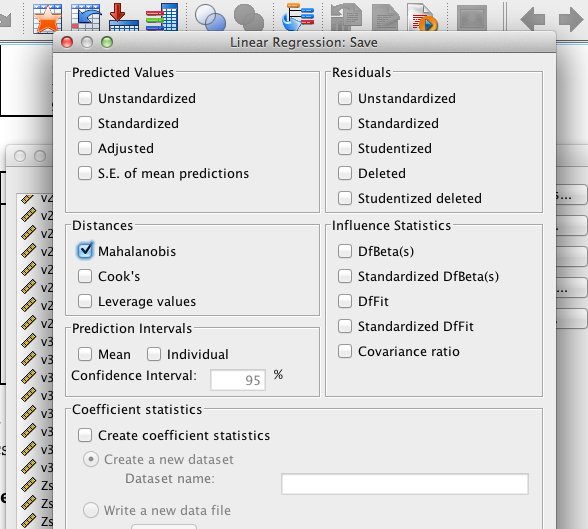
* + 1. Now run a fake regression.
    2. Analyze > regression > linear.



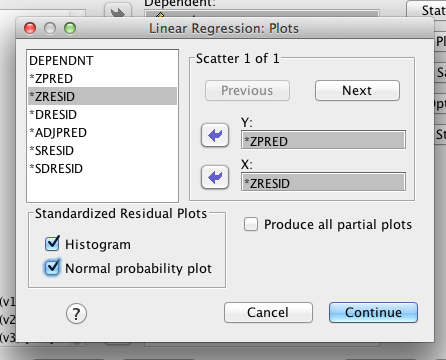
* + 1. Move the new “fake variable” *random* into the dependent box.
    2. Move all the independent variables into the independents box.



* + 1. Click Save.
    2. Click Mahalanobis distance.

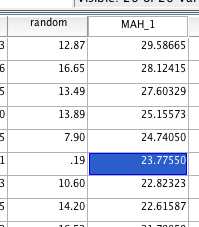


* + 1. Click plots (you will use this for the next analyses below).
       1. Under Y > ZPRED.
       2. Under X > ZRESID.
       3. Click Histogram and Normal Probability Plot.



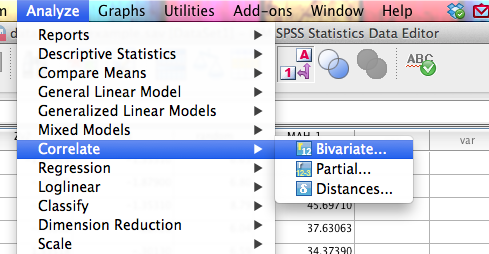
What’s going on here?

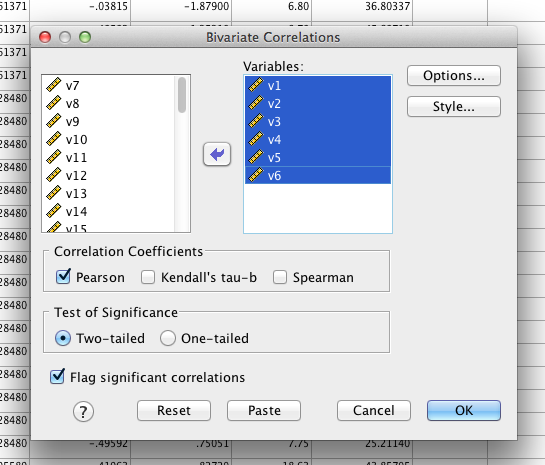
* ZPRED = z predicted value for your fake regression.
* ZRESID = z residual value (aka error variance) for your fake regression.
* We are plotting them against each other. In theory, the residuals should be *randomly distributed* (hence why we created a random variable to test with).
* Therefore, they should look like a bunch of random dots (see below).
* Histogram and Normal PP will give you more charts to help you check for other multivariate assumptions (see below).
  + 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).
    2. Decide if you want to delete or exclude values (can use filters or create yourself a new “no outliers” dataset, etc.).
    3. IF YOU DELETE PEOPLE: you will have to rerun that fake regression for the pictures below to match that you’ve deleted people.
       1. DO NOT delete outliers twice.
    4. IF YOU DO NOT DELETE PEOPLE: then the other output you got will be fine to use below.

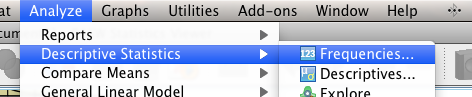
1. Multicollinearity/Singularity (aka check your correlations)
   1. Useful when you are doing a regression analysis or scales with total scores
   2. Inflated correlations – you can have super high correlations when you don’t mean to if you accidently use the same scale twice
   3. Deflated correlations – helps check for restriction of range.
   4. If you get a “Hessian Matrix not definite” error, you’ve used two variables in an analysis that are too similar.
   5. How to:
      1. Analyze > correlate > bivariate
      2. Move variable over, hit ok.

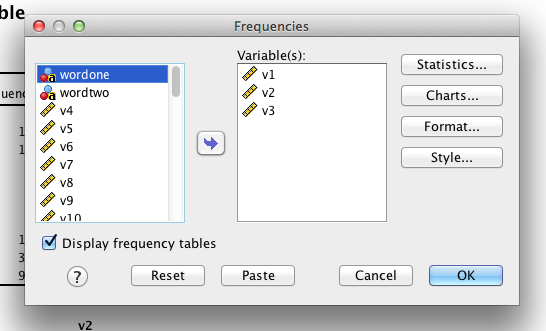


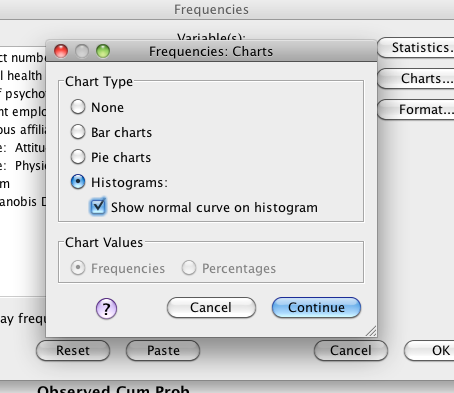


* 1. Check for correlations that are very high (r > .90-.95) or very low (r<.10). Low correlations will be normal, but very high correlations might give you problems in regression analyses. (low correlations are more of a problem for MANOVA, repeated measures ANOVA).

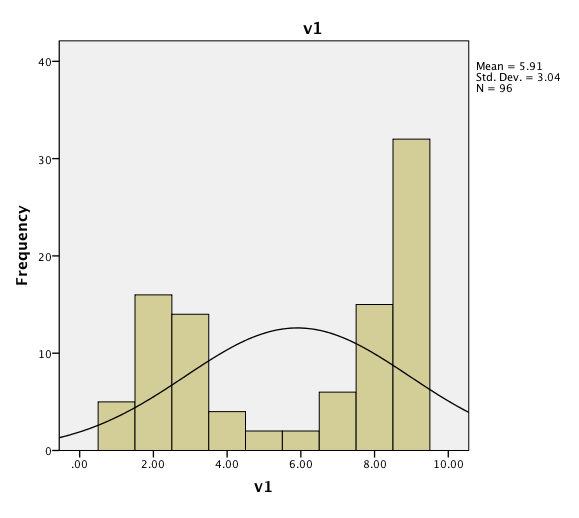
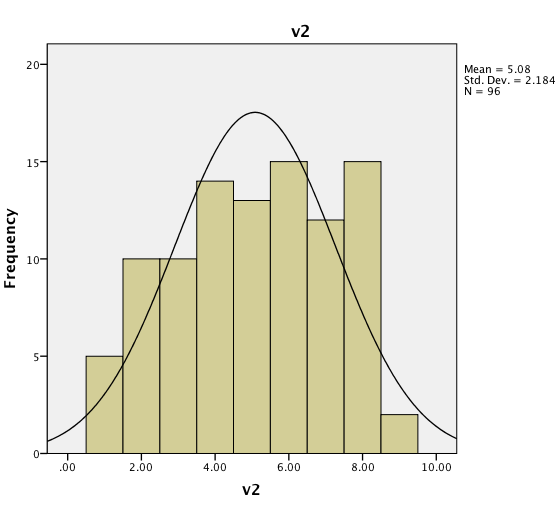
1. Normality
   1. Univariate normality – in reality we’ve already checked this by looking at the frequency analysis you ran earlier when checking for values in and out of range. You don’t want to do anything about them at that point because you aren’t sure if the problem is missing data, outliers, out of range data, etc. You have to fix those issues first. You can go back and check if the variables are normal now, by running that same analysis and looking at skew/kurtosis values.
   2. Analyze > descriptive statistics > frequencies – when you run this analysis hit charts, click histograms > normal curve.



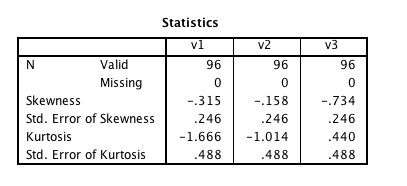




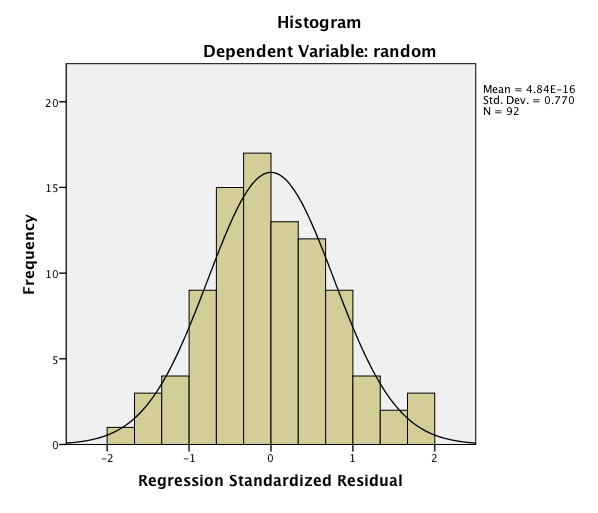
* + 1. Check pretty charts – this is univariate normality.

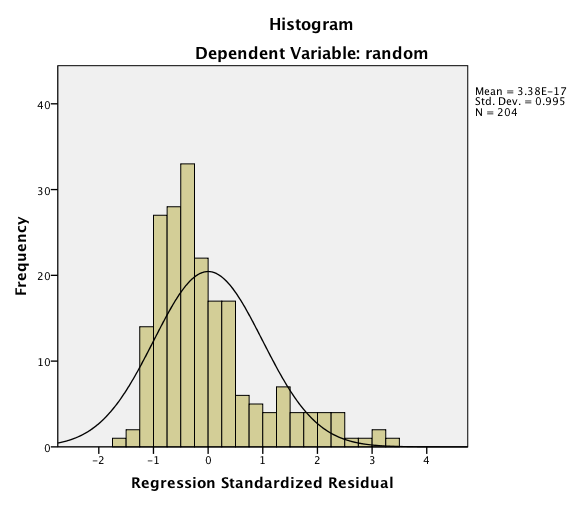
 (this chart is bad). (this chart is good).

* + 1. Check frequency charts for skew and kurtosis values over +/- 3.

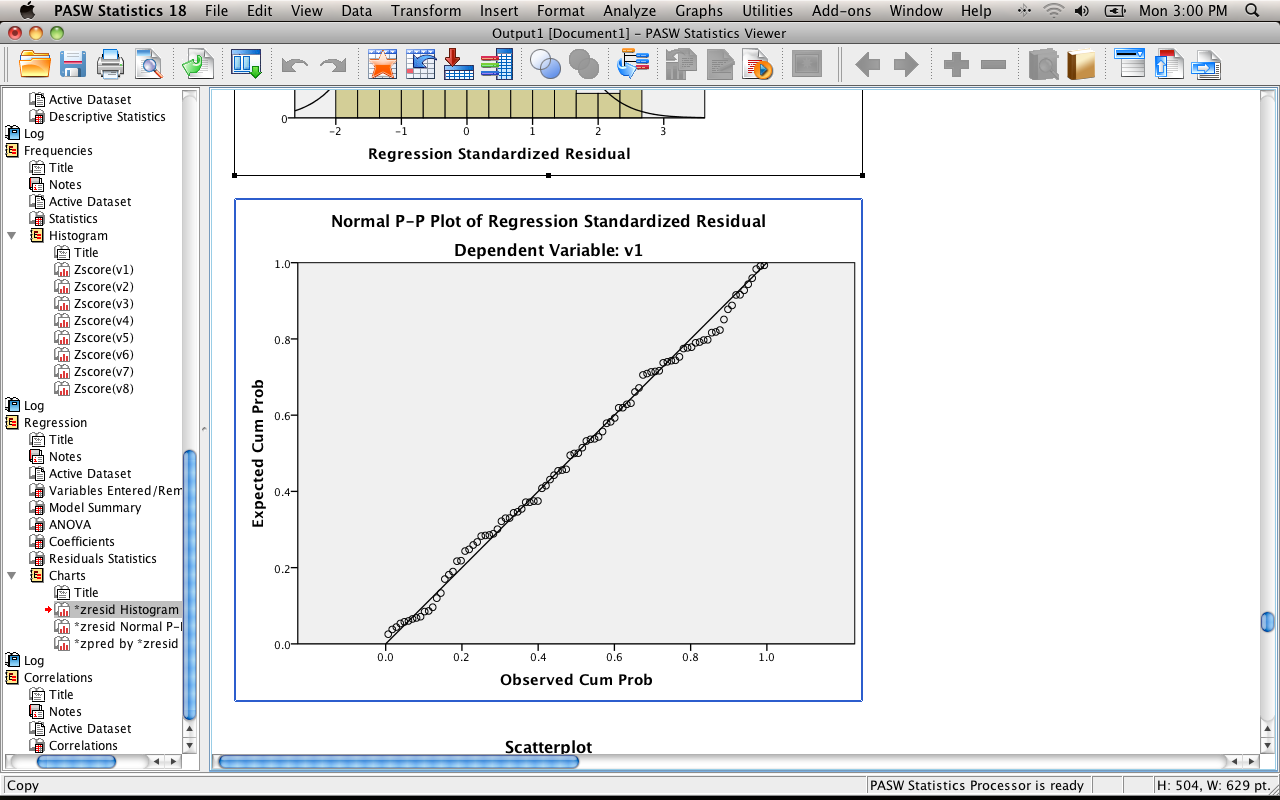


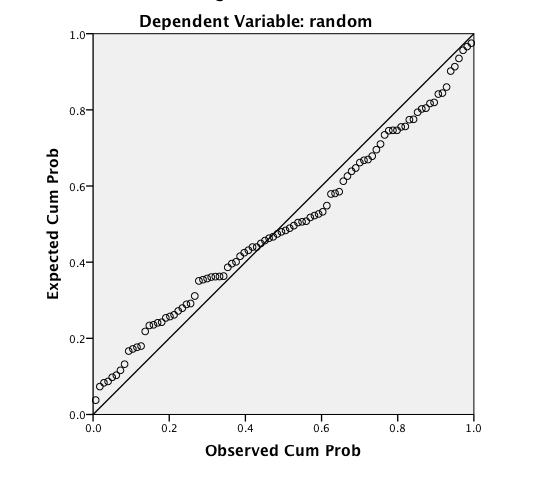
* 1. Multivariate normality – each variable and all linear combinations of variables are normal (more important!)
     1. Given the Central Limit Theorem – with *N* > 30 tests are robust (meaning that you can violate this assumption and get reasonably accurate statistics).
     2. When you run the regression analysis and ask for the normality chart you will get one for the multivariate combination.
     3. What to look for:
        1. See the numbers centered around zero at the bottom?
        2. You want an even spread around zero … so it shouldn’t look like -2 to 0 to +4 … that’s not even.
     4. What to do:
        1. Run a nonparametric test.



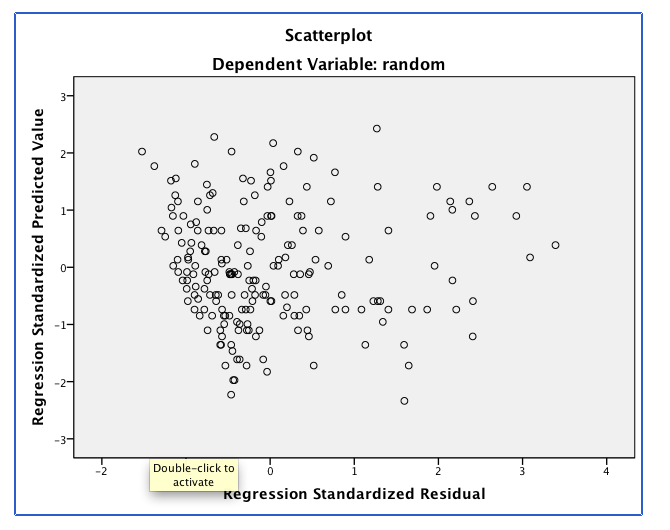


1. Linearity
   1. Assumption that there is a straight line relationship between two variables (or the combination of all the variables)
   2. Univariate: Look at bivariate scatterplots to check for curves (remember we covered how to make scatter plots in the APA section).
   3. Multivariate: Look at the normal pp plot to see if the dots follow the line (using the “Sarah Squint Test”).





1. Homoscedasticity: spread of the variance of a variable is the same across all values of the other variables.
   1. No cheerleaders, ufos, snakes that ate things.
2. Homogeneity
   1. Equal variances – you can look at the descriptive statistics and just look to see if they look roughly the same
   2. No raining
3. Both of these can be checked by looking at the residual scatterplot.
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 solid lines on the picture.
      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 (looks like it’s raining).
      4. If you encounter this problem – check out Levene’s test (for t and ANOVA), look at the SDs for each group to see what might be going on.
   3. Homoscedasticity – is the spread equal all the way across the zero line?
      1. These lines are the dashed lines on the picture.
      2. Look for megaphones or big lumps.
   4. Linearity – you can also see problems with linearity here, it will look like a rainbow or big curve around the center zero line.



**Results**

Prior to analysis, holistic, content, structure, stance, sentence fluency, diction and conventions within English papers were examined through various SPSS programs for accuracy of data entry, missing values, and fit between their distributions and the assumptions of multivariate analysis. The variables were examined for the 322 participants in the study.

The descriptive statistics showed that the means and standard deviations were relatively normal and that the maximum pre-development values for stance, sentence fluency, diction, and convention, and post-development stance, sentence fluency, and diction were incorrect. These values were found to be outside of the used scale and were replaced.

There were seven variables each with one missing value. These were pre-development workshop values for structure, sentence fluency, diction, and convention, and post-development workshop values for structure, stance, sentence fluency, and convention. These missing values were replaced with the linear trend at point.

Two multivariate outliers were found using Mahalanobis distance with p < .001. These outliers were deleted, leaving 320 cases. I ran a bivariate correlation to check for multicollinearity and singularity for the pre-workshop and post-workshop holistic variable. I made no change to this variable because it was not too highly correlated. I checked skewness and kurtosis of the variables and found all of them to be normal, so no changes were made.

The multivariate normality plot showed that results were normal, but slightly skewed to the right. The normal P-P Plot of regression standardized residual scatter plot shows that the variables were linear. The Standardized Regression Scatter plot shows that the results were homogeneic, but were not homoscedastic.