

## Using SPSS to Screen Data<sup>©</sup>

---

Download the file Screen2210.sav from my SPSS data page at <http://core.ecu.edu/psyc/wuenschk/SPSS/SPSS-Data.htm> and bring it into SPSS. The 'subjects' in this data file are automobiles. Among the variables on which you have data are:

- ID -- the identification number assigned to this subject.
- MPG -- the vehicle's mileage (gasoline consumption, miles per gallon).
- REPAIR -- the cost of repairs done on the automobile during the last year.
- SPEED -- the speed at which a vibration detector first crossed a threshold value (indicating that the vehicle's ride was becoming uncomfortable) when tested on the track.
- LIKERT4 -- the owner's response on a 4-option Likert-type question. The stem was "Overall, I am satisfied with this automobile." The response options were: (1) Strongly disagree, (2) disagree, (3) agree, and (4) strongly agree.
- GENDER -- of the owner, (1) for the one sex, (2) for the other.

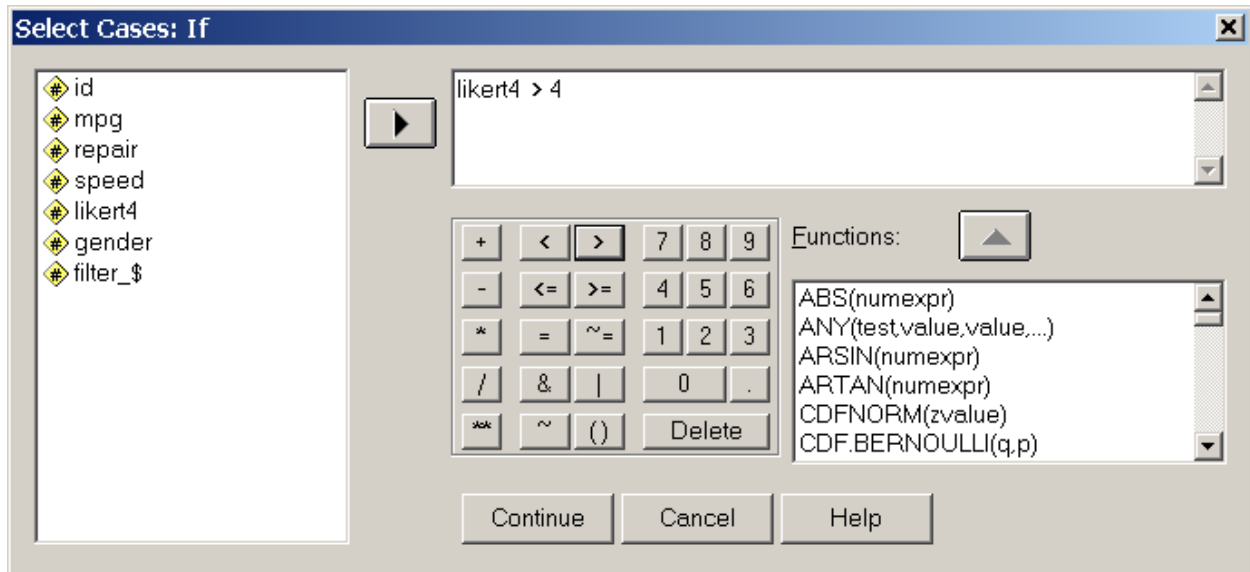
We want to screen these data for outliers and out-of range values. Since we intend to analyze the continuous variable with techniques that involve a normality assumption, we also want to determine if any of the continuous variables are distinctly non-normal in their distribution, and, if so, we want to try to find a transformation that will make them more nearly normal.

Let us first get some descriptive statistics on every variable except the ID number. Click Analyze, Descriptive Statistics, Descriptives. Scoot the five variables into the Variables box. Click Options and select Mean, Std. deviation, Minimum, Maximum, Kurtosis, and Skewness. Click Continue, OK.

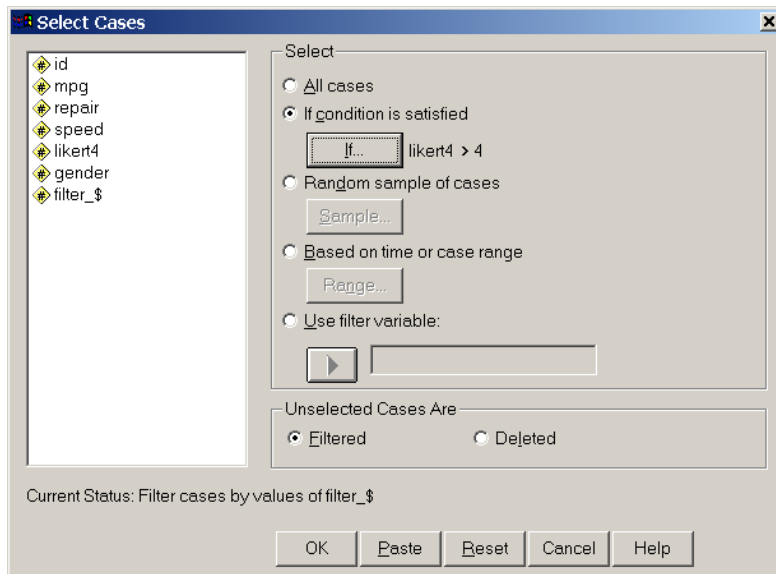
Look at the output. The variable Likert4 has values that range from 1 to 5, but there should be no values greater than 4. We shall need to determine which subjects have bad data on this variable. Do you see any evidence of bad data on another variable?

Now look at the Skewness and Kurtosis statistics for the variables MPG, Repair, and Speed. For MPG the skewness and kurtosis values are close enough to 0 that I would not be uncomfortable using them in an analysis that assumes that the data came from a normally distributed population. Repair and Speed are troublesome, however. I generally worry about a variable whose skewness exceeds an absolute value of 1, and I am not very comfortable with one whose skewness exceeds an absolute value of .7 or .8. High values of kurtosis also get my attention, since they often indicate that there are outliers in the distribution.

Let us now find the subjects who have bad data on the Likert4 variable. Go to the Data View and click Data, Select Cases. Select “If condition is satisfied” and “Filtered” for Unselected Cases, and then click on the “If” button. In the resulting “Select cases if” box, enter “likert4 > 4,” like this:

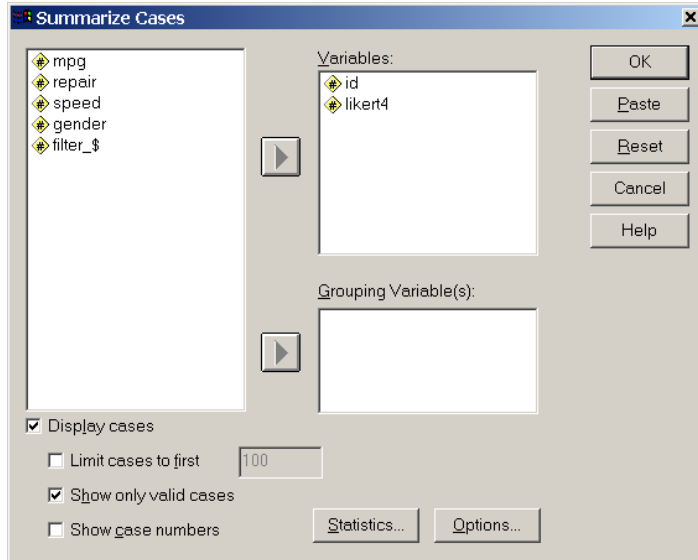


Click Continue. The Select Cases window should now look like this:



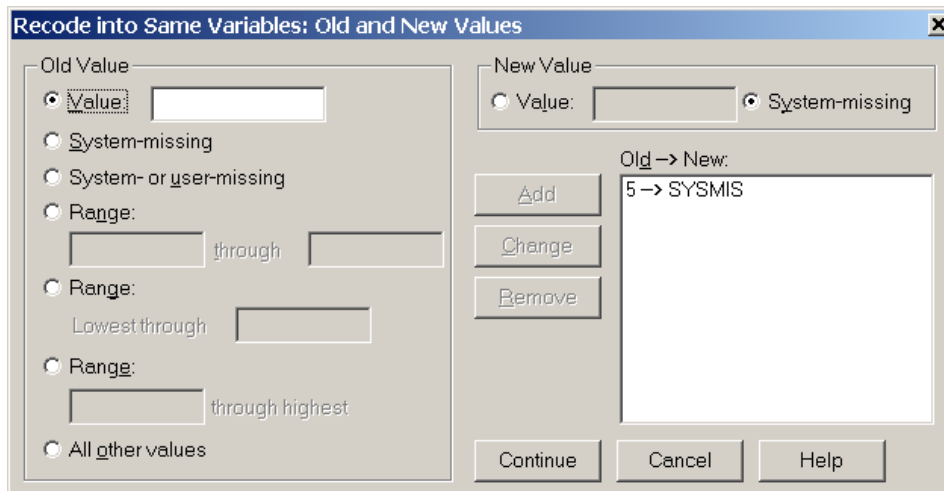
Click OK. Now look back at the data. You will see that there is a new variable, filter\_\$, with values of 1 for those cases where Likert4 > 4 and 0 for other cases. You will also see a slash through the case number of each case that has been filtered out.

Now let us get a listing of all the cases that have been selected and what value(s) they have on the Likert4 variable. Click Analyze, Reports, Case Summaries. Scoot ID and Likert4 into the variables box. Check only “Display cases” and “Show only valid cases.” The window should look like this:



Click OK. Look at the output. You will see listed there the ID numbers for five subjects who have out-of-range values for the Likert4 variable. We should now go find the original data sheets for these subjects and see what their actual responses were. If their actual responses have been incorrectly entered into the data file, we need to correct them. If their responses have been correctly entered, then we need to decide what to do with a response that is out of range. In some cases we might decide to recode them to a valid value -- for example, suppose that the survey had mostly questions with five response options, so that our subject got used to coding the 'E' or '5' response when their choice was the last option, but that for this item there were only four response options. Maybe those people who selected 5 really intended to select 4. Maybe we should recode all the scores of 5 to 4 on this variable.

Click Data, Select Cases, select "All cases," and click OK. On the Data View, click on Filter\_\$ and hit the delete key. Click Transform, Recode, Into Same Variables. Scoot Likert4 into the variables box. Click "Old and New Values." Under "Old Value" select "Value" and enter the number 5. Under "New Value" select "System-missing." Click "Add" and the window should look like this:



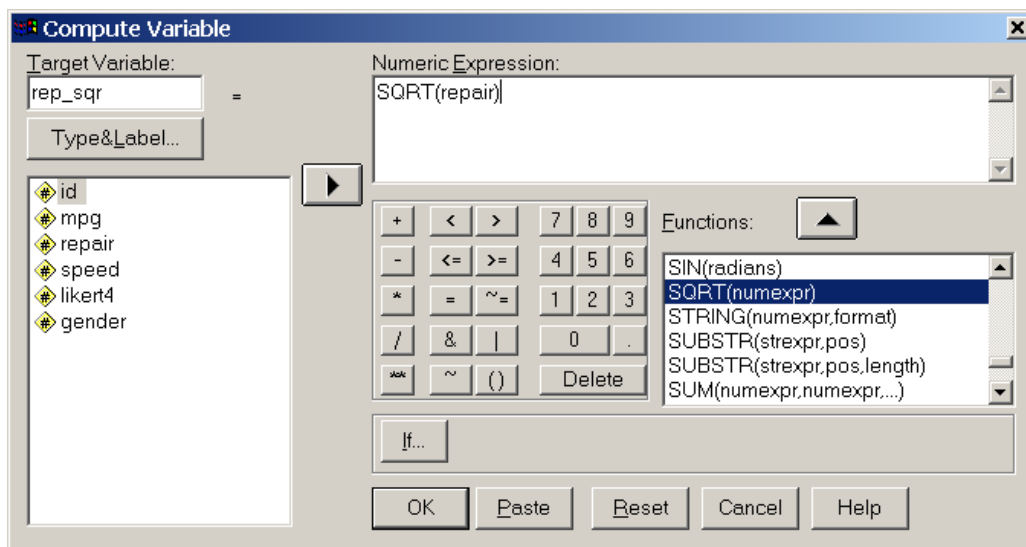
If you had decided to change the scores of 5 to scores of 4 instead of to missing values, you would select, under “New Value,” “Value,” enter the number 4, and then click “Add.” Go ahead and click Continue, OK to finish the recoding. If you look at the data you will now find that all of the scores of 5 have been set to a missing value.

Can you find the ID numbers of subjects who have out-of-range values on other variables in this data set?

Now let us use box and whisker plots to see if there are any outliers that deserve investigation. Click Analyze, Descriptive Statistics, Explore. Scoot MPG, Repair, and Speed, into the “Dependent List” and ID into the “Label cases by” box. Under “Display” select “Plots.” Click OK. Look at the output. The box plot for Repair shows one outliers, ID number 46. If you go back and check the data file you will find that car 46 had \$1,061 in repairs last year. While that is not an unbelievable value, you probably should investigate it just to be sure it is correct. The box plot for Speed shows six outliers, one of which is an extreme outlier (plotted with a star). That extreme outlier is ID number 33, an automobile that started vibrating at only 12 miles per hour, according to our data file.

If you cannot read the ID numbers for some of the outliers, you can always just use the Select Cases and Case Summaries procedures to get a list of ID numbers of cases with outliers. Do you remember from your statistics course how to find the “fences” that serve as the boundaries between outliers and adjacent values? If not, you should read my document [Exploratory Data Analysis \(EDA\)](#).

Finally, let us attend to the two variables which were unacceptably skewed. First, let us try to find a transformation which will reduce the skewness in the Repair variable. Click Transform, Compute. Type “rep\_sqr” in the “Target Variable” box and enter “SQRT(repair)” in the “Numeric Expression” box. The window should now look like this:



Click OK. If you look back at the data, you will see that the Rep\_Sqr transformed variable has been added. The square root transformation is often useful for reducing positive skewness. If the original variable has any negative values, you must remember

first to add a constant to all scores to avoid trying to take the square root of a negative number.

Let us also try an even stronger transformation for positive skewness, a logarithmic transformation. Click Transform, Compute. Type “rep\_log” in the “Target Variable” box and enter “LG10(repair)” in the “Numeric Expression” box and click OK. Also try a super powerful skewness-reducing transformation, the negative reciprocal. Click Transform, Compute. Type “rep\_nr” in the “Target Variable” box and enter “-1000/repair” in the “Numeric Expression” box and click OK.

Now we are ready to see what effect these transformations had on skewness and kurtosis. Compute skewness and kurtosis on the three transformed variables. You will find that the square root transformation reduced skewness nicely but that the other two transformations resulted in distributions that are unacceptably skewed in the negative direction.

We should try transforming the speed variable too. Recall that it is negatively skewed. We shall first reflect the variable by subtracting every score from a constant that is one greater than the highest score. Click Transform, Compute. Type “sp\_ref” in the “Target Variable” box and enter “91 - speed” in the “Numeric Expression” box and click OK. Compute the skewness of Sp\_Ref and you will find that it has exactly the same amount of skewness as did Speed but in a positive rather than a negative direction. Now try a square root and a log transformation on the Sp\_Ref variable. You will find that the log transformation does a good job of reducing the skewness. You could now use the log transformed reflected speed scores in an analysis that assumes normal distributions. When interpreting the results of that analysis you would have to remember that on your reflected speed variable low scores now represent high speeds and high scores represent low speeds. That can be confusing.