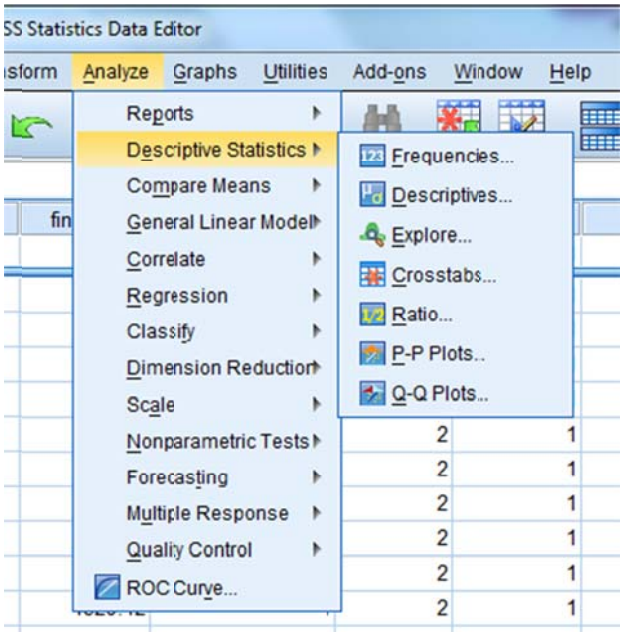
Introduction to statistics in SPSS

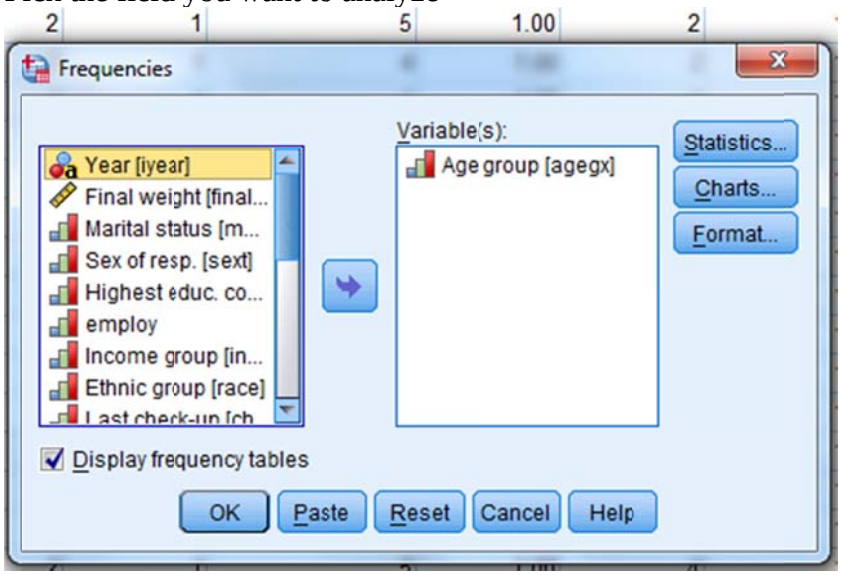
Jennifer LaFleur, VVOJ 2013

Frequencies in SPSS

In SPSS you can do several things with categorical data. First, frequencies will tell you how many items are in each category. To do that, click ANALYZE | DESCRIPTIVE STATISTICS |FREQUENCIES



Pick the field you want to analyze



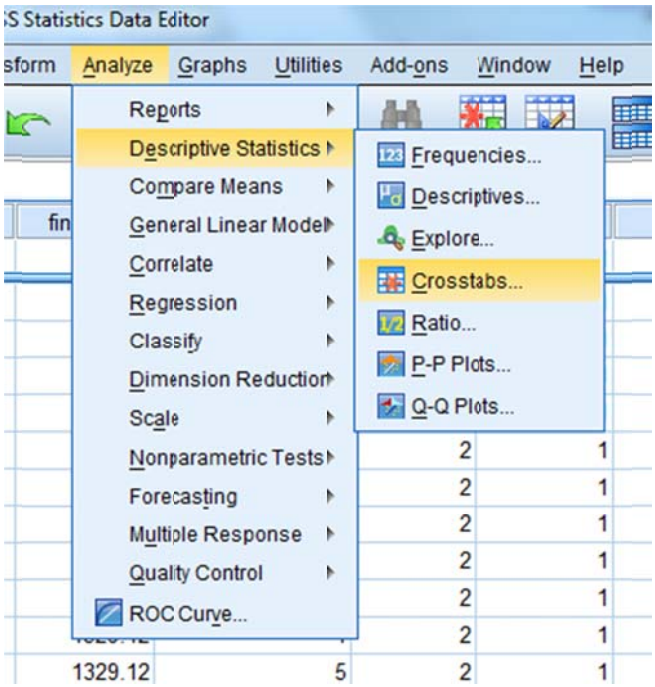
SPSS will give you a table with your results that includes the count, the percent of total and the cumulative percent of total:

Age group

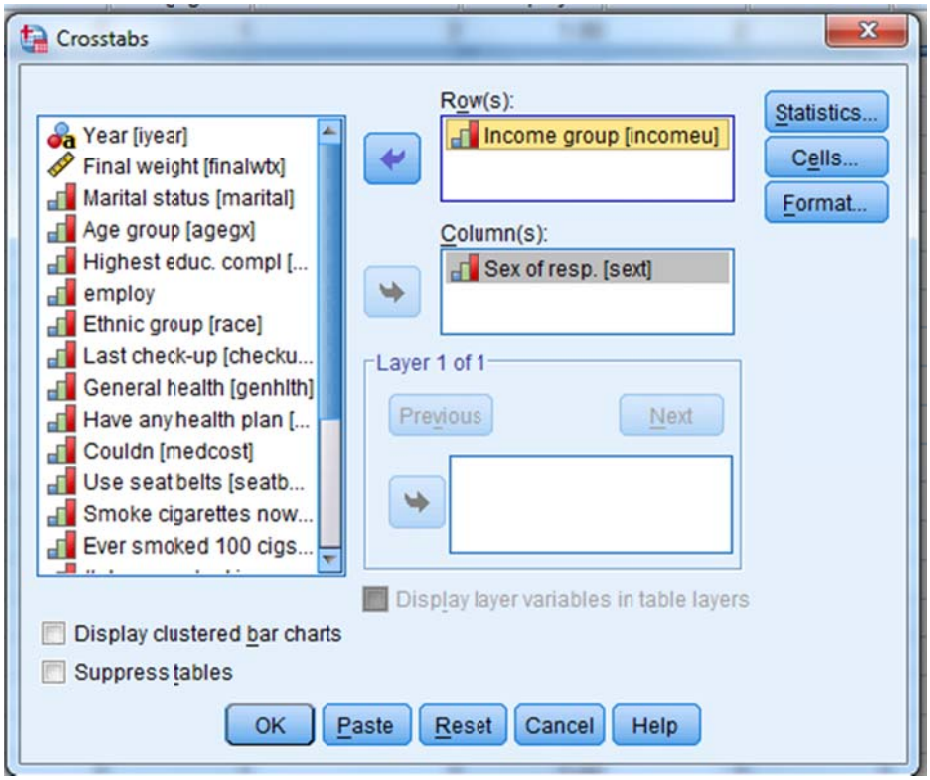
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18-34	1341	29.3	29.3	29.3
	35-64	1657	36.2	36.2	65.5
	65 & up	1577	34.5	34.5	100.0
	Total	4575	100.0	100.0	

Another useful tool is the crosstab

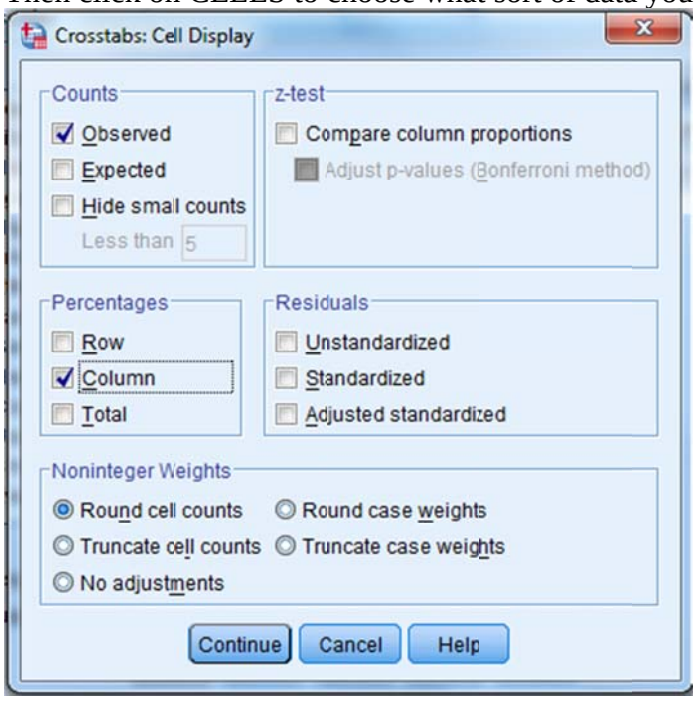
Go to ANALYZE | DESCRIPTIVE STATISTICS | CROSSTABS



Pick your row and column variables



Then click on CELLS to choose what sort of data you want to see.



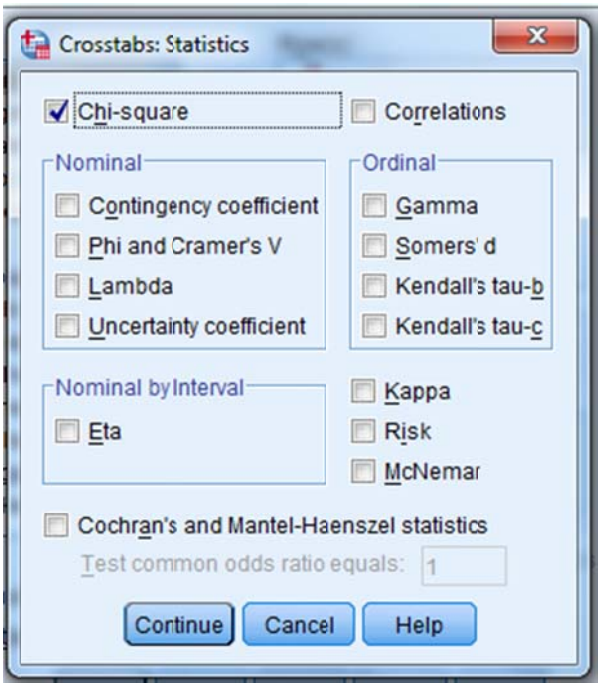
If I choose OBSERVED and COLUMN PERCENTAGE, I will get a typical crosstab table.

It gives me a table breaking down the two categorical variables

Income group * Sex of resp. Crosstabulation

			Sex of resp.		
			Male	Female	Total
Income group	<$10k	Count	111	423	534
		% within Sex of resp.	6.3%	15.0%	11.7%
	10k-14,999	Count	161	317	478
		% within Sex of resp.	9.1%	11.3%	10.4%
	15-19,999	Count	158	317	475
		% within Sex of resp.	9.0%	11.3%	10.4%
	20-24,999	Count	223	274	497
		% within Sex of resp.	12.6%	9.7%	10.9%
	25-34,999	Count	309	374	683
		% within Sex of resp.	17.5%	13.3%	14.9%
	35-50k	Count	291	357	648
		% within Sex of resp.	16.5%	12.7%	14.2%
	Unknown	Count	156	198	354
		% within Sex of resp.	8.8%	7.0%	7.7%
	=> 50k	Count	245	302	547

The Chi Square statistical test will tell you whether the difference between the two column cateogires is significant. Choose that under the statistics option in the crosstabs dialog box:



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	137.579[a]	8	.000
Likelihood Ratio	143.870	8	.000
Linear-by-Linear Association	42.989	1	.000
N of Valid Cases	4575		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 136.42.

If the last column is <.05, then there is significant difference. You may have to explore your data further to know which categories are significant.

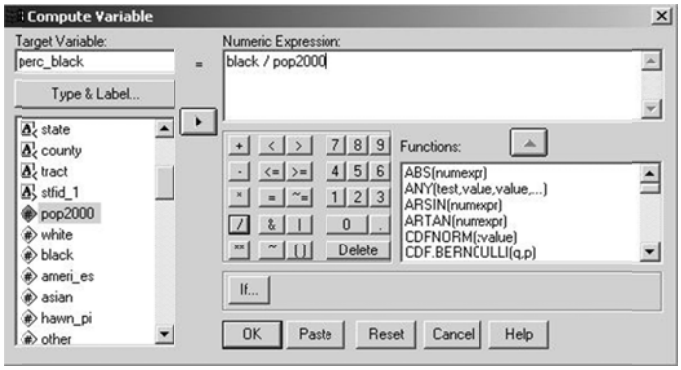
Calculating values in SPSS

Just like with your database manager, in SPSS, you will need to calculate new fields.

Let's do this with a crime database from the city of Dallas.

We have demographic data and crime data. We might want to calculate percentages for our race fields. Let's start with BLACK.

Click on TRANSFORM | COMPUTE. SPSS will prompt you for the name of your new fields and the fields you want to use for your calculation.



Be sure to type a variable name that does not exist in your database or SPSS will overwrite that field. Here we will assign the new field PERC_BLACK to be BLACK/POP2000. If you want to assign a specific type and label – click that button. Otherwise, you new variable will be assigned the same type as the fields you use to calculate it.

If you want to perform a calculation only on certain records, you can click the IF box and SPSS will let you filter your data before performing you computation.

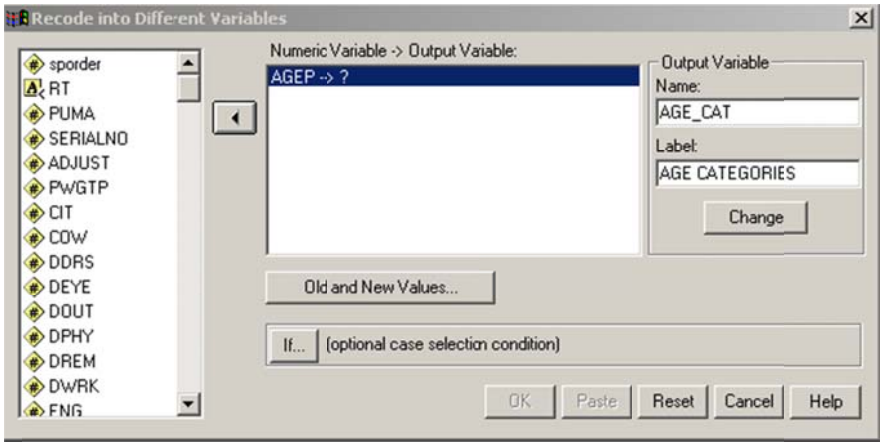
Recoding values to categories

Just like in a database manager, you can edit your data to show categories or to show codes or to show ranges of values. In SPSS, you do all that work under TRANSFORM | RECODE.

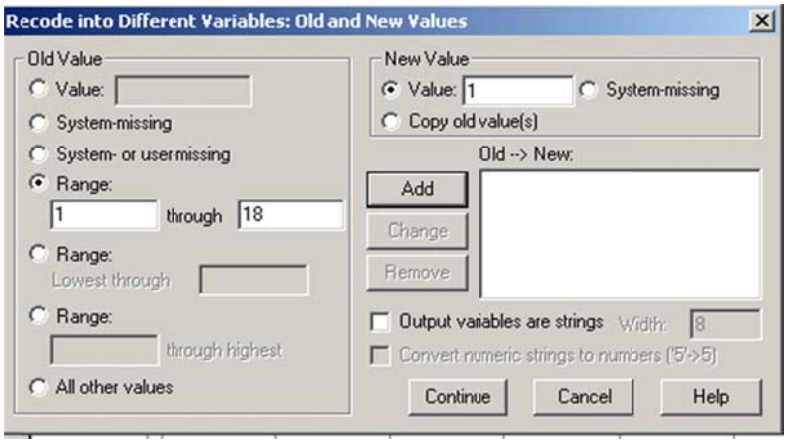
Let's say, for example, that rather that you'd like to run a crosstab using age data from the American Community Survey. The data you get in the raw file actually contains every single age, not age categories. Using the recode function, you can crunch all those ages into a few categories.

With your ACSPUMS03 files open, go to TRANSFORM | RECODE | INTO DIFFERENT VARIABLES. Note that you'll almost always want to put it into a different variable so you don't overwrite your original field.

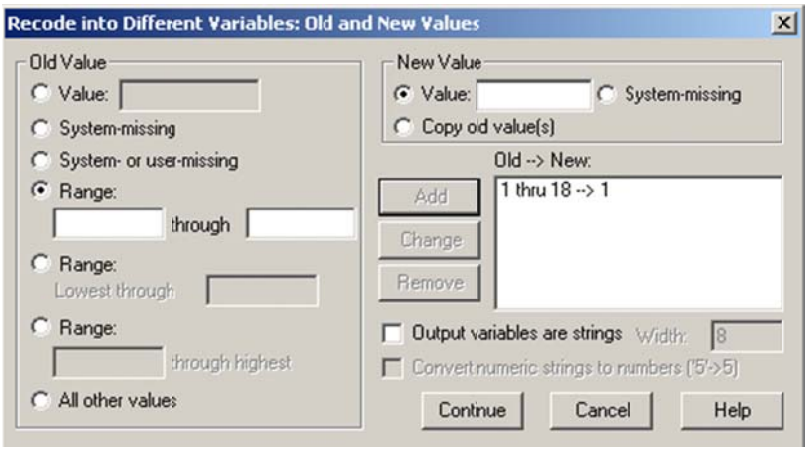
Choose your age field (AGEP) and move it into the main value box. Then under OUTPUT VARIABLES give the new field a name and a label. The name is the field name. The label can be a fancier name you use in your output.



Next, click on OLD AND NEW VALUES.



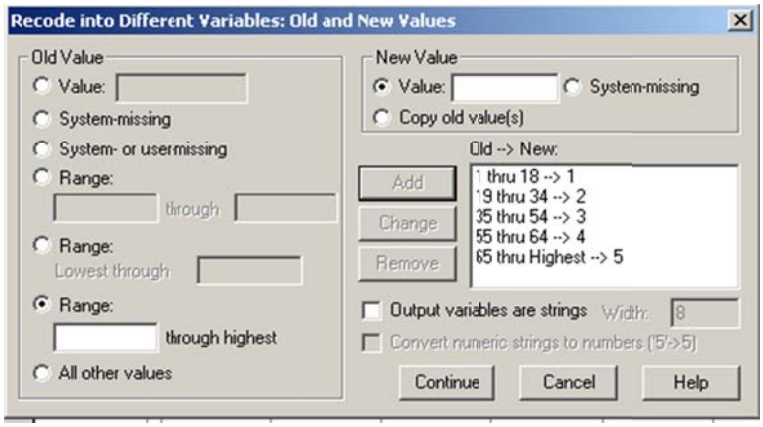
Here's where you change the ranges into categories. Let's say we want ages 1 through 18 to be our first category. Click on bubble next to RANGE and type 1 through 18. Under NEW VALUE, type 1. Then click the ADD button.



Let's continue this until we have all our ranges using this breakdown:

2 19 through 34
3 35 through 54
4 55 through 64
5 65 and older

To get that last category, click on the option that has an empty box through HIGHEST. This is what it should all look like:



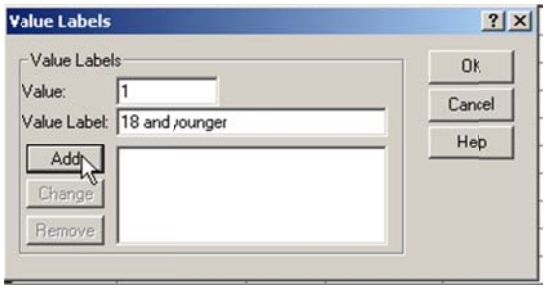
Next, click CONTINUE to go back to the main window.

Then click CHANGE under your output variables. Then click OK.

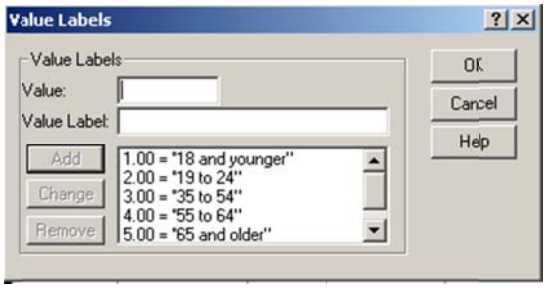
If you look at your data, you should have a new column called AGE_CAT. But wait! We're not done yet. If we ran anything, we wouldn't know what the fields meant. Now we have to create labels.

To do that, go to the VARIABLE view of your data and go to the AGE_CAT variable. Click on the button in the VALUES column.

You'll get the value label dialog box.



Type in 1 for our first category and type in "18 and younger" for the label. Then click ADD. Do this for all the categories. Then click OK.



If you go back to your data and then choose VIEW | VALUE LABELS, you'll see your labels instead of your codes. You can change that back by going to VIEW and unchecking VALUE LABELS.

Now when you run a frequency table, using AGE_CAT, you'll have something a little more useable.

Practice:

Recode the field called MSP – Marital Status, using these codes:

1 Now married, spouse present
2 Now married, spouse absent
3 Widowed
4 Divorced
5 Separated
6 Never married

Time-saving tip for labeling: Syntax

Syntax files let you make some adjustments to tasks without having to go through all the clicking again and again.

For example, here's the syntax to label educational attainment in our file:

```
value labels SCHL
01 'None'
02 'Nursery school to grade 4'
03 'Grade 5 or grade 6'
04 'Grade 7 or grade 8'
05 'Grade 9'
06 'Grade 10'
07 'Grade 11'
08 'Grade 12 no diploma'
09 'High school graduate'
10 'Some college but no degree'
11 'Vo/Tech/Bus school degree'
12 'Associate degree in college'
13 'Bachelor's degree'
14 'Master's degree'
15 'Professional school degree'
16 'Doctorate degree'
```

You can save these files to use later. To run them go to RUN | ALL.

Let's create a syntax file to label the field called CIT (citizenship):

Go to FILE | NEW | SYNTAX

Type the following into the box:

```
value labels CIT
1 'Born in the U.S'
2 'Born in Puerto Rico, Guam, the U.S. Virgin Islands, or the Northern Marianas'
3 'Born abroad of American parents'
4 'U.S. citizen by naturalization'
5 'Not a citizen of the U.S.'
```

Save your file as CIT_LABELS. Then go to RUN | ALL.

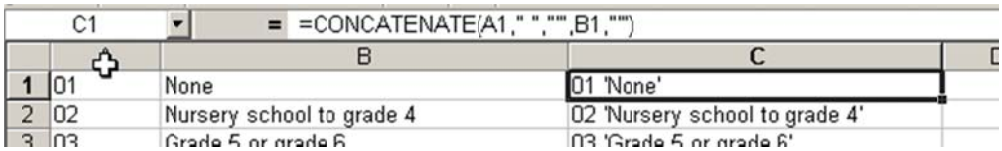
Some government agencies that output data in SPSS format also include syntax files, which will save you lots of time getting your data organized.

When you have a lot of labels to add and you have an electronic version of your records layout, you can use Excel and syntax to make it go a little faster.

Our record layout was in an Excel spreadsheet and had the following columns for educational attainment:

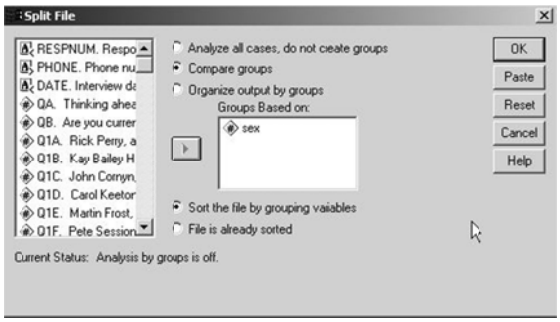
	A	B
1	01	None
2	02	Nursery school to grade 4
3	03	Grade 5 or grade 6
4	04	Grade 7 or grade 8
5	05	Grade 9
6	06	Grade 10
7	07	Grade 11
8	08	Grade 12 no diploma
9	09	High school graduate
10	10	Some college but no degree
11	11	Vo/Tech/Bus school degree
12	12	Associate degree in college
13	13	Bachelor's degree
14	14	Master's degree
15	15	Professional school degree
16	16	Doctorate degree

Using the concatenate function, you can get Excel to do most of the work:



Some great basic data tools in SPSS

The **SPLIT** function: This function lets you run the same analysis on groups of variables. Get to this function by clicking on DATA | SPLIT FILE.

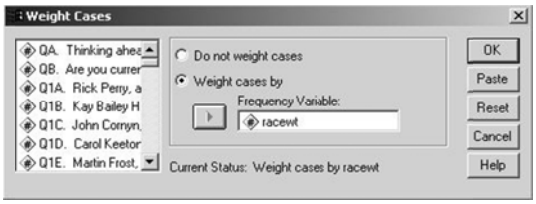


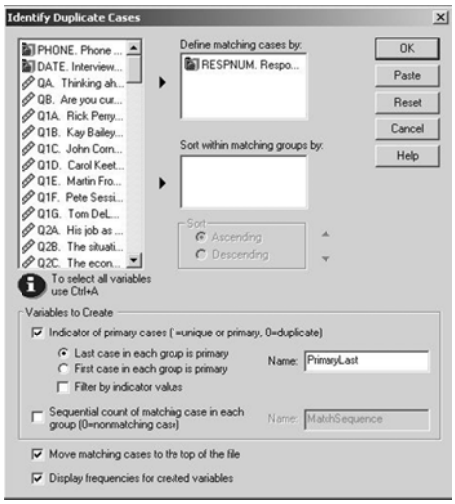
Compare groups puts both groups into one table. Organize output by groups gives you two separate outputs.

CROSSTABS: Think of these as grouping by multiple variables. You can do this in your database manager – but it's easier and more manageable in SPSS. Get there by: ANALYZE | DESCRIPTIVE STATISTICS | CROSSTABS. (More on this one later.)

Using SPSS syntax: You may get data sets that already include syntax (IPUMS data is one example.) And THAT saves a lot of work. You can also create your syntax files from your record layout. A syntax file is a separate file of commands. It is saved as a *.sps. Once you open it – you can run all or part of the program.

WEIGHTING data: You can do this in your database manager as well, but it's easier to do in SPSS. Get there by DATA | WEIGHT CASES. If you do much work with PUMS data, you'll love this function.





IDENTIFY DUPLICATE CASES: Yes, you can do this in Access, but SPSS will help you out. It identifies duplicate cases based on a field you identify and then will do things with this cases such as move them to the top of your file or sort them. Get there by DATA | IDENTIFY DUPLICATE CASES.

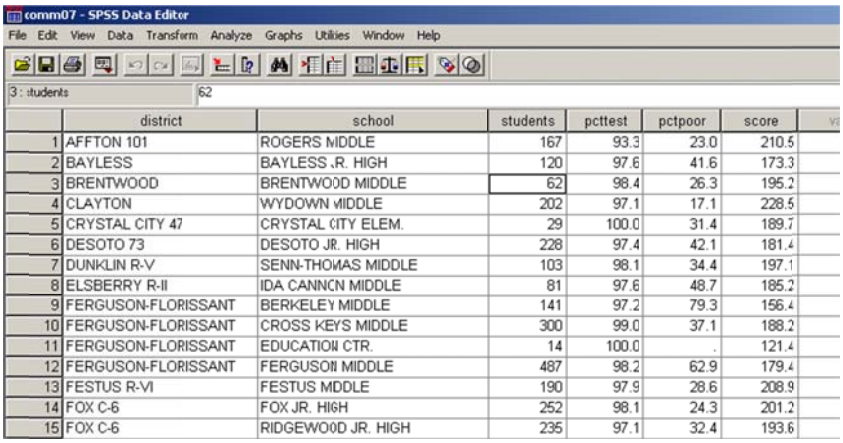
RANK: The SPSS RANK function adds a new field to your data with a rank based on another field. SPSS lets you do things with your RANK such as telling it ahead of time how to deal with ties. You can give it different types of ranks such as n-tiles. Get there by TRANSFORM | RANK CASES.

Introduction to Linear Regression in SPSS

Step-by-step example: School test scores

Let's focus on the relationship between school test scores and family income. This exercise uses SPSS 11.5 to analyze scores of St. Louis-area schools on a Missouri state test: seventh-grade communication arts (a fancy education term for reading and writing).

Open the file called comm07.sav and you should see this:


comm07 - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

3 : students 62

	district	school	students	pcttest	pctpoor	score
1	AFFTON 101	ROGERS MIDDLE	167	93.3	23.0	210.5
2	BAYLESS	BAYLESS JR. HIGH	120	97.6	41.6	173.3
3	BRENTWOOD	BRENTWOOD MIDDLE	62	98.4	26.3	195.2
4	CLAYTON	WYDOWN MIDDLE	202	97.1	17.1	228.5
5	CRYSTAL CITY 47	CRYSTAL CITY ELEM.	29	100.0	31.4	189.7
6	DESOTO 73	DESOTO JR. HIGH	228	97.4	42.1	181.4
7	DUNKLIN R-V	SENN-THOMAS MIDDLE	103	98.1	34.4	197.1
8	ELSBERRY R-II	IDA CANNON MIDDLE	81	97.6	48.7	185.2
9	FERGUSON-FLORISSANT	BERKELEY MIDDLE	141	97.2	79.3	156.4
10	FERGUSON-FLORISSANT	CROSS KEYS MIDDLE	300	99.0	37.1	188.2
11	FERGUSON-FLORISSANT	EDUCATION CTR.	14	100.0	.	121.4
12	FERGUSON-FLORISSANT	FERGUSON MIDDLE	487	98.2	62.9	179.4
13	FESTUS R-VI	FESTUS MDDLE	190	97.9	28.6	208.9
14	FOX C-6	FOX JR. HIGH	252	98.1	24.3	201.2
15	FOX C-6	RIDGEWOOD JR. HIGH	235	97.1	32.4	193.6

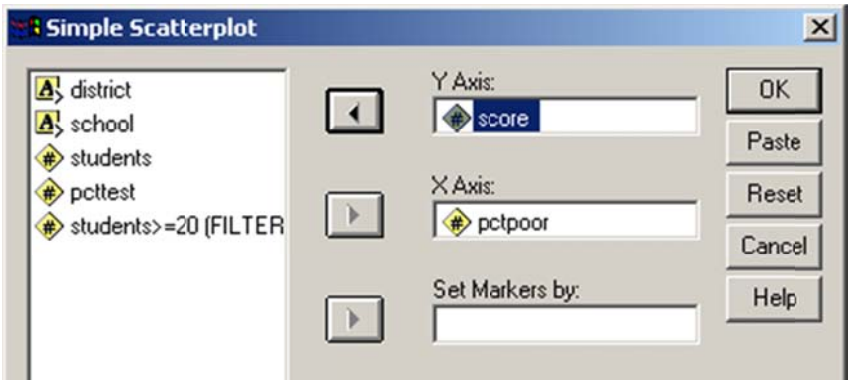
Each row contains several variables: the name of a school and its district; the number of students who took the test (*students*); the percent of students tested (*pcttest*); the percent of children from low-income families (*pctpoor*); and the average test score (*score*), which ranges from a low of 100 to a high of 300.

We suspect that family income is linked to test scores: As the percent of poor children in a school rises, the average test score will decline. We call the test score (*score*) the *dependent* variable and family income (*pctpoor*) the *independent* variable. That's because we believe scores *depend* partly on income.

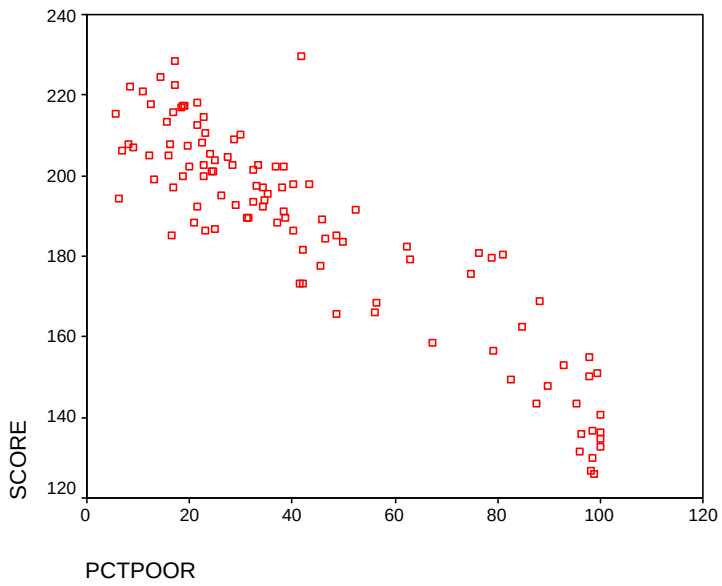
First, spend time examining each of the variables in your analysis by running descriptive statistics. Once you've done that, do a scatter plot to look at the data together.

We need to plot the dependent variable, *score*, on the vertical (Y) axis and the independent variable, *pctpoor*, on the horizontal (X) axis.

To do that, go to the data editor window (the one with your table) and select **Graphs**, then **Scatter**. Make sure the **Simple** box is highlighted, then click the box that says **Define**. You want to show *score* on the Y axis and *pctpoor* on the X axis, like so:



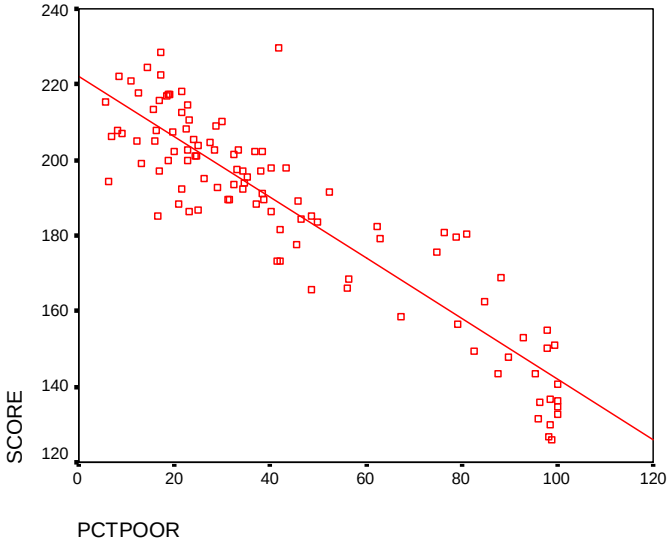
Click **OK** and the scatterplot appears in the Output window.



Each point represents a school. Its location on the graph indicates its poverty rate (on the X axis) and average test score (on the Y axis). See how the points follow a pattern: As the percent poor increases, the average test score declines. So, we're on the right track thinking family income is related to test scores.

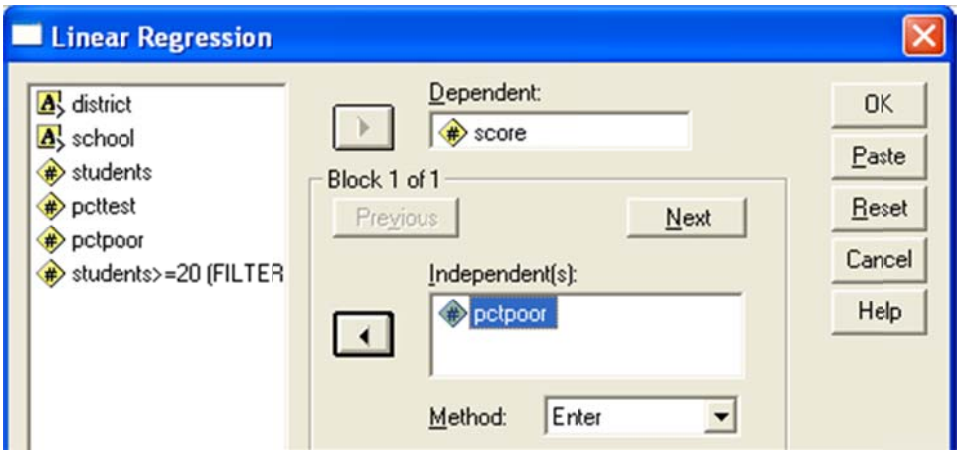
We can draw a line to represent the pattern. Click on the scatterplot in the Output viewer and right-click on the mouse. Choose **SPSS Chart Object**, **Open**. The chart editor pops up. In that editor, select Chart, Options. Under **Fit Line**, check the box next to **Total**. Click **OK**.

Voila, SPSS draws a line that follows the pattern of the points. This is called the regression line. It is the line that best fits all the points on the graph. Still, we don't know how accurate the line is – that is, how well does it describe the relationship between family income and test scores?

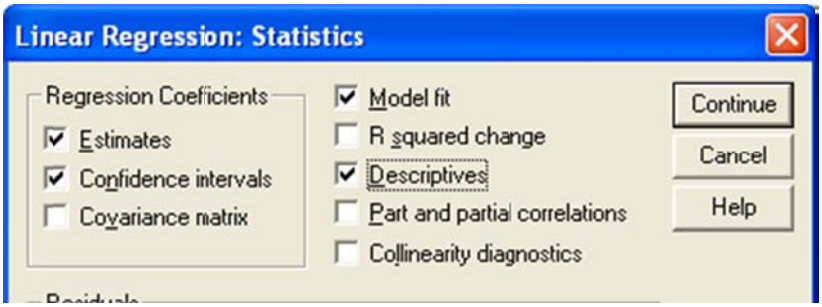


Ready to regress

To answer that question, we get to the fun part: Running the actual regression. To do that, go to **Analyze**, **Regression**, **Linear**. In the dialog box that pops up, select *score* as the dependent variable and *pctpoor* as the independent variable.



Next, click the **Statistics** box. Make sure the **Estimates** and **Model fit** boxes are checked, then also check **Confidence intervals**.



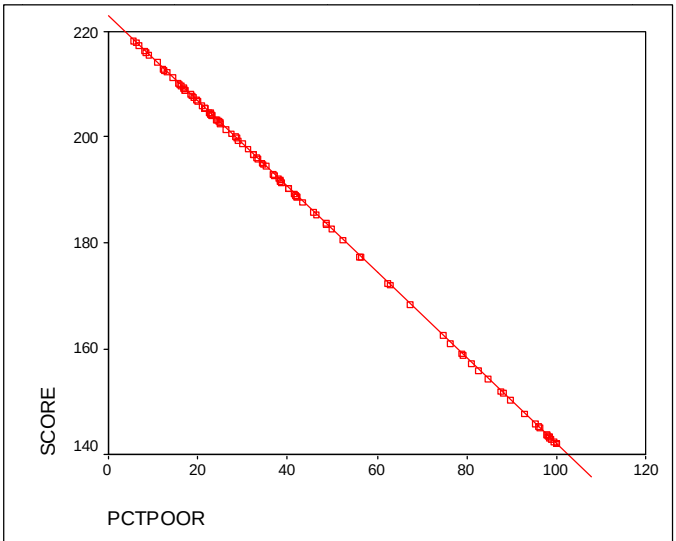
Click **Continue**, then **OK**. Look at table in the output called Model Summary:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.911(a)	.830	.829	10.8330

a Predictors: (Constant), PCTPOOR

Focus on the Adjusted R Square, .829 in this case. What the heck does this mean, and why should you care? The R-square measures the accuracy of the regression line and ranges from 0 to 1. An R-square of 1 means that the line fits perfectly. If poverty predicted test scores perfectly, the scatterplot would look like this:



At the other extreme, an R-square of 0 means the line fits terribly. The points would be randomly scattered, with no discernible pattern. So the Adjusted R-square of .829 shows a strong relationship. The best way to understand it – and to explain it to readers – is to say that 83 percent of the variation in test scores depends on family income. That means only 17 percent is explained by other factors. Our theory seems to be right on.

Next we want to know just how much income affects test scores. To determine that, look at the Coefficients table:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	222.294	1.884		117.995	.000	218.558	226.030
	PCTPOOR	-.802	.036	-.911	-22.556	.000	-.873	-.732

a Dependent Variable: SCORE

Look at the first number listed for PCTPOOR, -.802. This is the slope of your regression line. Huh? Think back to algebra when you learned the formula for a line:

$$y = mx + b$$

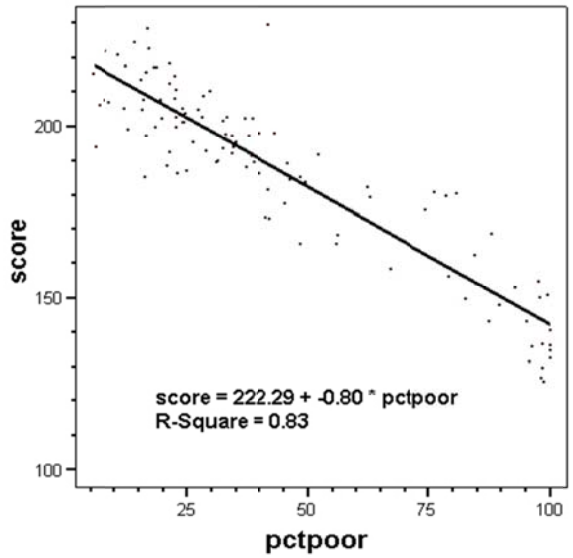
y is the value along the vertical axis – in our case, the predicted test score

x is the value along the horizontal axis – in our case, the low-income rate

m is the slope of the line. It measures how much y changes in relation to x. The steeper the line, the stronger the effect. In our example, for every one-point increase in the poverty rate, we expect a school's test score to decline 0.8 points.

b is the y-intercept. It represents the value of y when x is 0. So a school with zero poverty should have a

We can plug the values from our Coefficients table into that equation, y = mx + b, to get the specific equation to predict school test scores based on poverty.



We're not done yet, though. We need to make sure our model is significant – that this relationship between poverty and test scores does not exist by chance. Look at the ANOVA (analysis of variance) table.

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	59704.532	1	59704.532	508.759	.000(a)
	Residual	12204.731	104	117.353		
	Total	71909.263	105			

a Predictors: (Constant), PCTPOOR
b Dependent Variable: SCORE

The ANOVA table tells us if the relationship between test scores and income could have happened by chance. Because the significance (**Sig**.) is less than .05, we can say that our model is significant – that poverty does help predict test scores.

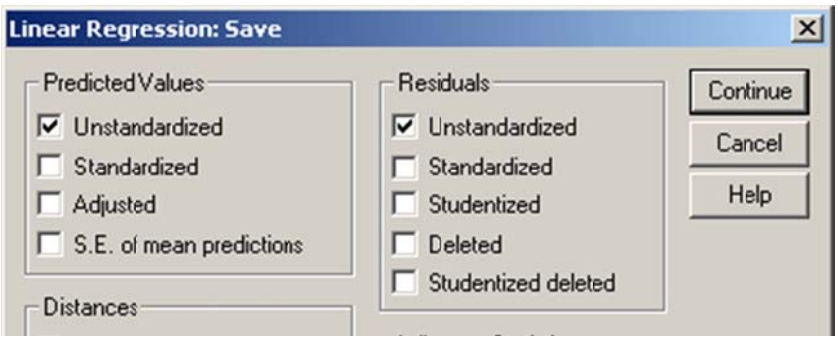
We also need to make sure poverty predicts test scores as strongly as we suspect. Our model says that a one-point increase in the poverty rate should result in test scores declining eight-tenths of a point. Go back to the Coefficients table.

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	222.294	1.884		117.995	.000	218.558	226.030
	PCTPOOR	-.802	.036	-.911	-22.556	.000	-.873	-.732

Look at the Significance (Sig.) and the 95% Confidence Interval. The significance level is below .05, which means our results were in all likelihood not due to chance. The Confidence Interval gives us the range we'd expect if we took repeated samples of test scores. We can be 95 percent certain that the true effect of poverty on test scores is between -.87 and -.73. That is, for every one-point increase in a school's poverty rate, we expect its test score to decline between .87 and .73 points.

Predicting individual scores

Remember, the beauty of regression is that we can predict test scores for each school based on its poverty rate. To do that, go to **Analyze**, **Regression**, **Linear**. Click on **Statistics** and select **Casewise Diagnostics** in addition to the boxes already checked. Click on the **Save** box. Under both **Predicted Values** and **Residuals**, check **Unstandardized**.



Click **Continue**, then **OK**. Go to the Data Editor to view your school data. You'll see two new fields at the end, *pre_1* and *res_1*. SPSS used the regression line formula to predict each school's score. The residual is the difference between the actual and predicted score.

comm07.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

9 : pctpoor 79.288

	district	school	students	pcttest	pctpoor	score	filter_$	pre_1	res_1
1	AFFTON 101	ROGERS MIDDLE	167	93.3	23.0	210.5	1	203.81353	6.68647
2	BAYLESS	BAYLESS JR. HIGH	120	97.6	41.6	173.3	1	188.93184	-15.63184
3	BRENTWOOD	BRENTWOOD MIDDLE	62	98.4	26.3	195.2	1	201.19670	-5.99670
4	CLAYTON	WYDOWN MIDDLE	202	97.1	17.1	228.5	1	208.55128	19.94872
5	CRYSTAL CITY 47	CRYSTAL CITY ELEM.	29	100.0	31.4	189.7	1	197.07444	-7.37444
6	DESOTO 73	DESOTO JR. HIGH	226	97.4	42.1	181.4	1	188.51857	-7.11857
7	DUNKLIN R-V	SENN-THOMAS MIDDLE	103	98.1	34.4	197.1	1	194.72643	2.37357

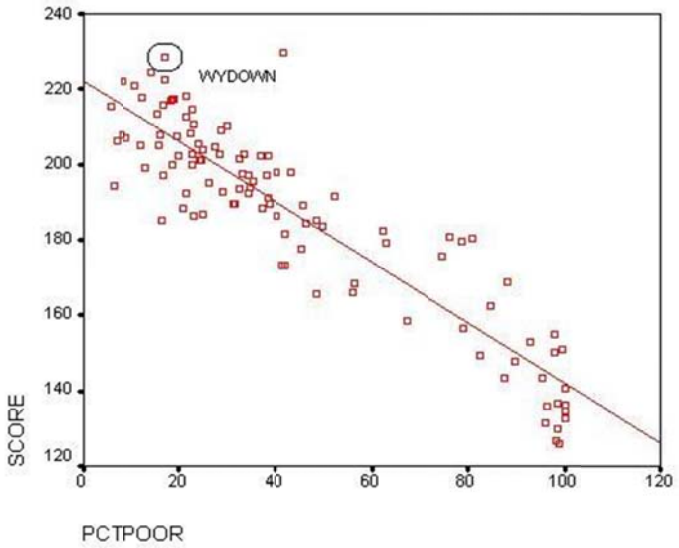
Take the fourth school on the list, Wydown Middle in the Clayton School District. It had an average score of 228.5 and a poverty rate of 17.1 percent. So, go back to our formula. We want to know our predicted score:

Predicted score = 222.3 + (-.802 * Pctpoor)
Predicted score = 222.3 + (-.802 * 17.1)
Predicted score = 222.3 – 13.7
Predicted score = 208.6

But Wydown actually scored 228.5. So it did better than expected:

Actual score – Predicted score = Residual
228.5 – 208.6 = 19.9

Here's Wydown circled on the scatterplot. Note how it sets well above the line, showing it did better than expected.



We can even rate schools based on how much better or worse they did than expected. People can set different criteria, but one is called the **Standard Deviation of the Residual**, which appears in the Residuals Statistics box generated in your last regression.