# Using SPSS

Econ 29000
Kevin R Foster, CCNY
Spring 2011

PASW and SPSS are the same program, the publisher just changed the title from SPSS to PASW. In the computer lab you'll find it as SPSS.

# Why learn SPSS?

Table COMP: Comparison of Data Analysis Packages

| Name | Advantages | Disadvantages | Open source? | Typical users |
|---|---|---|---|---|
| R | Library support; visualization | Steep learning curve | Yes | Statistics |
| Matlab | Elegant matrix support; visualization | Expensive; incomplete statistics support | No | Engineering |
| SciPy/NumPy/ Matplotlib | Python (general-purpose programming language) | Less mature | Yes | Engineering |
| Excel | Easy; visual; flexible | Large datasets; weak numeric support | No | Business |
| SAS | Large datasets | Very baroque; hardest to learn | No | Business |
| Stata (and SPSS) | Easy statistical analysis | Less programmatic than R/Matlab/Py | No | Science (bio and social) |

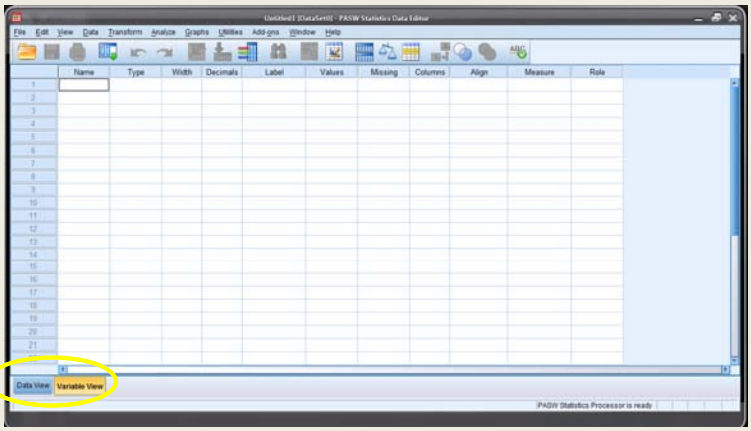
from: http://assets.doloreslabs.com/blog/oconnor_biewald_beautiful_data_final_nonlayout_20090803_20090327.pdf

# The Absolute Beginning

- Start up SPSS from the Start menu (under "Programs")
  - Note that sometimes double-clicking an associated file won't work
- You'll see this sort of screen – click "Cancel"


# Load a SPSS Dataset

- When SPSS starts, you will be in the "SPSS Data Editor" which looks like this.
- Click on "File" then choose "Open" then "Data…" [**not** "File/Open Database" – that's different]. For example, the ATUS dataset.

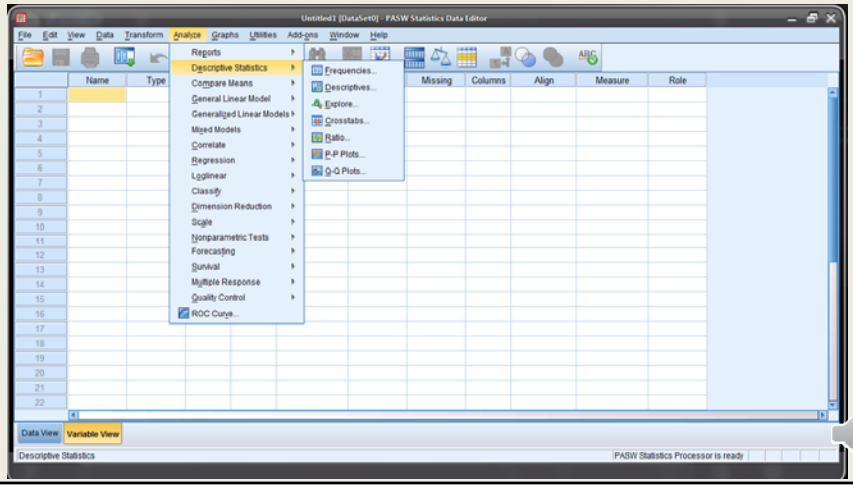


# Save your Work!

- Choose "File" then "Save" and save it to the desktop or somewhere else where you will remember it.
- **Save regularly!**
- You might "Save As…" a new name, so that if you screw up you can revert to an older version.
- Back it up – email or USB drive

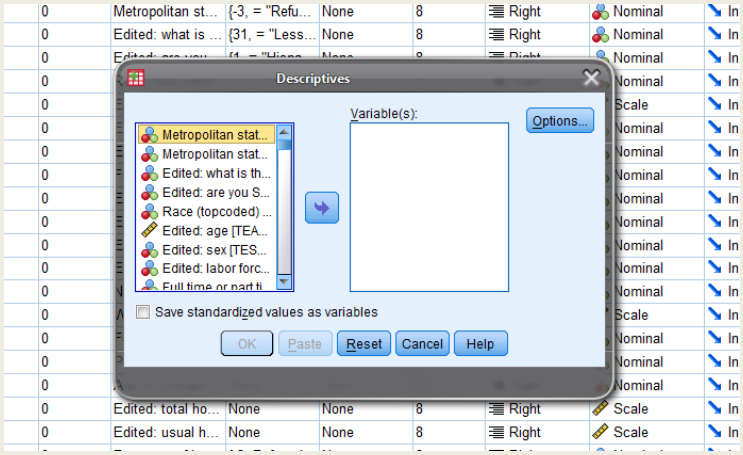
# Getting Basic Statistics

- From either the "Data View" or "Variable View" tab, click "`Analyze`" then "`Descriptive Statistics`" then "`Descriptives`":



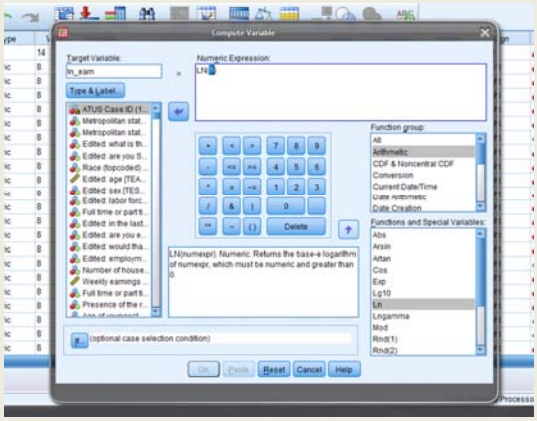
# Getting Basic Statistics

- This will bring up a dialog box asking you which variables you want to get Descriptive Statistics on.




# Create New Variables

- We often create new variables. One common transformation is taking the log. This is a common procedure to cut down the noise and help to examine growth trends. Click on "`Transform`" and then "`Compute`...". This will bring up a dialog box labeled "`Compute Variable`".
- Type in the new variable name (whatever you want, just remember it!) under "`Target Variable`". (You can click '`Type & Label`" if you want to enter more info that can remind yourself later.) For example we'll find the log (natural log) of weekly earnings.



## Create New Variables

- Under "Target Variable" type in the new name, "ln_earn" or whatever and then in "Numeric Expression" you tell it what this new variable is. You can make any complicated or convoluted functions that are necessary for particular analyses; for now find the "Function Group" to click on "Arithmetic" and then in the "Functions and Special Variables" list below find "Ln". Double-click it and see that SPSS puts it up into the "Numeric Expression" box with a (?) in the argument. Double-click on the variable, weekly earnings (TRERNWA), that you want to use and then hit "OK".
- You'll get a bunch of errors where the program complains about trying to find the log of zero, but it still does what you need. For wages, where many people have wage=0, we often use `lnwage = ln(wage + 1)` which eliminates the problem of `ln(0)` that returns an error; for most other people the distinction between `ln(1000)` and `ln(1001)` is tiny. You can go back and re-do your variable if you're feeling a need to be tidy.

## Create New Dummy Variables

- Dummy variables have values of either zero or one; usually a zero corresponds to "No" and a one to "Yes" – they're Boolean
- For example, a dummy variable for "female" has a one if the individual is female and zero if male

# Understanding Complicated Variables

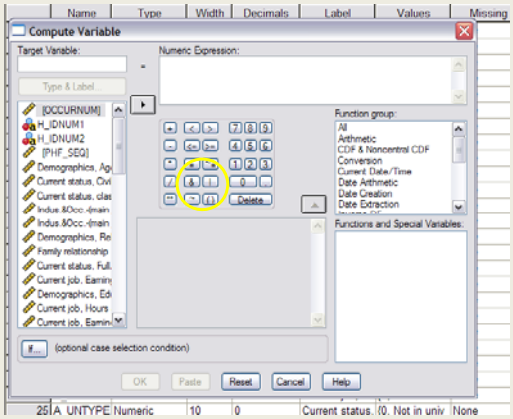
- The "Values" label can be tricky to use.
- For instance in the ATUS, as you look at the "Variable View" of your dataset, one of the first variables in the dataset has the name "PEEDUCA", which is short for "PErson EDUCation Achieved" – the person's education level.
- But the coding is strange: under "Values" you should see a box with "..." in it – click on that to see the whole list of values and what they mean. You'll see that a "39" means that the person graduated high school; a "43" means that they have a Bachelor's degree.
- Without that "Values" information you'd have no way to know that. It also means that you must do a bit of work re-coding variables before you work with the data.
- The variable "TEAGE" (which is the person's age) has numbers like 35, 48, 19 – just what you'd expect. These values have a natural interpretation; you don't need a codebook for this one!
- The variable "TESEX" tells whether the person is male or female – but it doesn't use text, it just lists either the number 1 or 2. We could guess that one of those is male and the other female, but we'd have to go back to "Variable View" to look at "Values" for "TESEX" to find that a 1 indicates a male and a 2 indicates female.

# Create New Dummy Variables

- From"TESEX" create, instead, a dummy variable (that takes a value of just zero or one) called "female" that is equal to one if the person is female and zero if not.
- To do this, click "Transform" then "Compute..." which will bring up a dialog box.
- The "Target Variable" is the new variable you are creating; for this case, type in "female".
- The "Numeric Expression" allows considerable freedom in transforming variables. For this case, we will only need a logical expression: "TESEX = 2". You can either type in the variable name, "TESEX", or find the variable name in the list on the left of the dialog box and click the arrow to insert the name.
- If you go to "Data View" and scroll over (new variables are all the way on the right) you can check that it looks right.

# Create New Dummy Variables

to create a variable "Hispanic" you'd type "`Hispanic`" into the Target Variable, then click the "`( )`" button (see the yellow circle in the screenshot below) to get a parenthesis, double-click the variable that codes ethnicity so as to get `PEHSPNON` in the "`Numeric Expression`" and then add "`=1`" to finish, so getting a relationship that Hispanic is defined as: (`PEHSPNON = 1`). SPSS understands that whenever that relation is true, it will put in a 1; where false it will put in a 0.




# Create New Dummy Variables

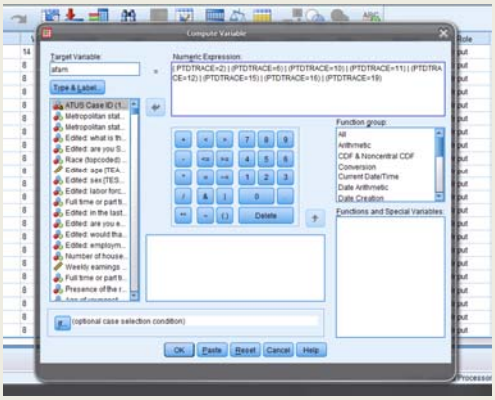
- Now that the Census asks people for detailed race info (could be African-American only or Asian or bi-racial or tri-racial in various combinations – see the online note on ATUS for more details), researchers might aggregate together everyone who replies that they are all or part African-American. So maybe create a new variable AfricanAmerican as `(PTDTRACE = 2) | (PTDTRACE = 6) |(PTDTRACE = 10) |(PTDTRACE = 11) |(PTDTRACE = 12) |(PTDTRACE = 15) |(PTDTRACE = 16) |(PTDTRACE = 19)`
- The line up and down, `|`, represents the logical "OR"; the tilde, ~, is logical "NOT".
- If you wanted to create a variable for those who report themselves as African-American and Hispanic, you'd create the expression `(AfricanAmerican = 1) & (PEHSPNON = 1)`

# Dummy Variables for Race

- Create dummy variables for "white", "African-American", "American Indian/ Inuit/ Hawaiian/ Pacific Islander", and "Asian." We'll lump together the people who give multiple identities with those who give a single one (this is standard in much empirical work, although it is evolving rapidly).

# Dummy Variables for Race

- So "`Tranform/Compute...`" and label "`Target Variable`" as "`white`" with "`Numeric Expression`" "`PTDTRACE=1`". Then "`afam`" is "`( PTDTRACE=2) | (PTDTRACE=6) | (PTDTRACE=10) | (PTDTRACE=11) | (PTDTRACE=12) | (PTDTRACE=15) | (PTDTRACE=16) | (PTDTRACE=19)`" – note the parentheses and the "or" symbol.
- "`Asian`" is "`( PTDTRACE=4) | (PTDTRACE=8)`". "`Amindian`" is "`( PTDTRACE=3) | (PTDTRACE=5) | (PTDTRACE=7) | (PTDTRACE=9) | (PTDTRACE=13) | (PTDTRACE=14) | (PTDTRACE=17) | (PTDTRACE=18) | (PTDTRACE=20) | (PTDTRACE=21)`". (Many of these codings of multiple races could be argued – you can make changes if you wish.)

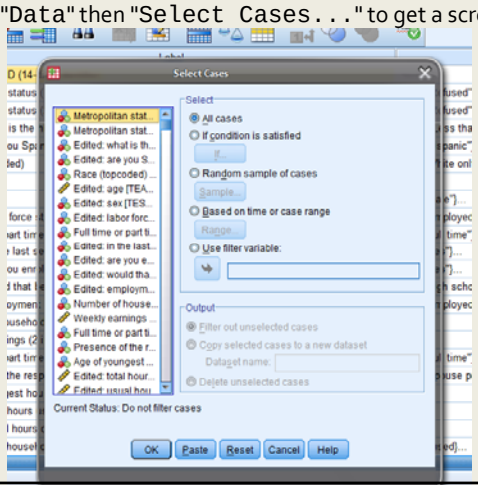


# Dummy Variables for Education

- Create a dummy for no high school "ed_nohs", for high school but no further "ed_hs", for some college "ed_scol", for a bachelor's degree "ed_coll", and for more than a 4-year degree "ed_adv".
- "Transform/Compute…", set "Target Variable" as "ed_nohs" and "Numeric Expression" as " PEEDUCA <39". Then "ed_hs" is " PEEDUCA =39"; "ed_scol" is "(PEEDUCA >39)&(PEEDUCA <43)"; "ed_coll" is "PEEDUCA =43"; "ed_adv" is "PEEDUCA >43".
- run "Descriptive Statistics" to make sure everything looks right – your dummy variables should have min=0 and max=1, for example!

# Data Sub-Sets

- Often we want to compare groups of people within the dataset to each other, for example looking at whether men or women spend more time with their family or watching TV or whatever. Comparisons are often more useful than just raw numbers because comparisons allow us to begin to judge which differences are substantial.
- Do this with "Data" then "Select Cases..." to get a screen like this:




# Data Sub-Sets

- Usually we select cases "`If condition is satisfied`" so choose that, then click on "`If...`"
- This brings up a dialog box that looks like the "Compute Variable" box from above. If we have already created a dummy variable that has values of only zeroes and ones then you can just put that into the "Select Cases" box. If you want a more complicated set then you can build it up using the logical notation that we discussed above. So suppose you want to look at just the subgroup of women between the ages of 18-35. Then we would enter "`(TESEX = 2) & (TEAGE > 18) & (TEAGE <= 35)`". Click "Continue". Make sure the output is "Filter out unselected cases" (you don't usually want to permanently delete the unselected cases!). Then all of your subsequent analyses will be done for just that subgroup.
- Often an analysis will be more concerned with whether a particular item is done rather than how long – for example, when looking at working, whether a person has a second job (so time spent working second job is greater than zero) is probably more important than just how long they spent working at this second job. So often the "`if...`" statement will be of the form, "`X > 0`" for whatever variable, X, you're considering.

# Data Sub-Sets

- Alternately, you can use "`Analyze \ Descriptive Statistics \ Explore...`" which asks you to put in the "`Dependent List`" which are the variables whose means you want to find, and then the "`Factor List`" which defines categories, by which subgroups are being compared.
- So, for example, if you wanted to look at the time sleeping, depending on whether there are kids in the house, you could put "`Time Sleeping`" into the "`Dependent List`" and then "`Presence of Household Children`" into the "`Factor List`".
- SPSS also has "`Analyze \ Compare Means`"

## Do It!

The only way to learn SPSS is to start doing it. Trust me! You'll need to use SPSS for many homework assignments, for your final project, and even on the exams. Start learning it now.