# Jumpstart Statistics: Data entry and data management issues

Steve Simon, P.Mean Consulting

Informatic Medicine and Personalized Health, UMKC

www.pmean.com

www.facebook.com/pmean

www.linkedin.com/in/pmean

# 2. Why am I offering this webinar?

- The Department of Informatic Medicine and Personalized Health at UMKC has just opened a new Research and Statistical Consult Service.
  - www.med.umkc.edu/informatic_medicine/consultation.shtml

# 3. Why am I offering this webinar?

- The RSCS offers advice on research design and data analysis, but the customers are responsible for their own data entry and data analysis. These Jumpstart Statistics webinars offer additional support to beginning researchers who may not know how to get started.
  - www.pmean.com/webinars/Jumpstart.html

# 4. The next webinar

Jumpstart Statistics:

Simple descriptive statistics

Friday, March 4, noon-1pm

Details to appear at

www.pmean.com/webinars/jumpstart.html

# 5. Abstract

- This **training class** will give you a general introduction to data management using IBM SPSS software. This class is useful for anyone who needs to enter or analyze research data. **No statistical experience is necessary.**

# 6. Abstract

- There are three steps that will help you get started with data entry for a research project. First, arrange your data in a rectangular format. Second, create a name for each column of data and provide documentation on this column such as units of measurement. Third, create codes for categorical data and for missing values.

# 7. Abstract

- This class will show examples of data entry including the tricky issues associated with data entry of a two by two table and entry of dates.

# 8. Abstract

In this class, you will learn how to:

- document variables in a IBM SPSS data set;
- enter and manipulate dates in IBM SPSS; and
- import data into IBM SPSS from other programs.

# 9. Outline

1. Pop quiz
2. Documenting variables in IBM SPSS
3. Inputting two by two table
4. Inputting dates
5. Importing data
6. Repeat of pop quiz

# 10. Pop quiz

1. IBM SPSS provides documentation about the individual levels of a categorical variable using
   - Format type
   - Missing value codes
   - Variable labels
   - Value labels

# 11. Pop quiz

2. In IBM SPSS, if you subtract one date from another to compute the number of days between two events you will get the following result.
   - An error message
   - A missing value
   - A result that is far too large to be correct
   - A warning message

# 12. Pop quiz

3. In IBM SPSS, you can simplify the data entry for a two by two table by using
   - Automatic recode
   - Crosstabs
   - Restructure wizard
   - Weight cases

# 13. First three steps in data entry

- Every data set is different, but most data entry procedures will start out the same. Here are the first three steps that you should follow to help insure a successful data entry.
  1. Arrange your data in rectangular format.
  2. Create variable names (8 characters or less).
  3. Assign number codes for categorical data and missing values.

# 14. Rectangular format

- Arrange your data in a rectangular format. **The intersection of each row and column should contain a single number**. Don't leave a cell empty if you can possibly avoid it. Don’t try to squeeze two numbers into a single cell.

# 15. Problem with empty cells

- Empty cells are ambiguous (does it represent a missing value or is it a sign that data entry is not yet complete on this patient). Some computer programs (not IBM SPSS) will take an empty cell and convert it to zero, which can lead to disastrous results.

# 16. Don’t squeeze two numbers into one cell

- You might be tempted, for example, to list blood pressure as 120/80 to represent the systolic and diastolic pressures. Don't do this! You will make it difficult for the computer to compute an average blood pressure of any type. Furthermore, some computer software programs might look at the entry 120/80, misinterpret the slash as a division sign, and replace the whole cell with 1.5.

# 17. Transforming to a rectangular format

- Here's an example of data which does not fit into a rectangular format. These data are loosely based on a study of breast feeding in pre-term infants.

Breast feeding status at six months

| No | | | Yes | | | Lost to follow-up | | |
|---|---|---|---|---|---|---|---|---|
| Mom's Age | Marital Status | Birth Weight | Mom's Age | Marital Status | Birth Weight | Mom's Age | Marital Status | Birth Weight |
| 18 | Married | 1.550 | 28 | Single | 2.381 | 28 | Married | 1.685 |
| 33 | Single | 1.990 | | | 1.130 | | | 2.435 |
| 34 | Married | | 26 | Married | 2.060 | | | |
| 36 | Married | 1.640 | | | | | | |

# 18. Transforming to a rectangular format

- Notice that there is a 4 by 3 matrix (4 rows by 3 columns) for the “No” group, a 3 by 3 matrix for the “Yes” group, and a 2 by 3 matrix for the “Lost” group.

Breast feeding status at six months

| No | | | Yes | | | Lost to follow-up | | |
|---|---|---|---|---|---|---|---|---|
| Mom's Age | Marital Status | Birth Weight | Mom's Age | Marital Status | Birth Weight | Mom's Age | Marital Status | Birth Weight |
| 18 | Married | 1.550 | 28 | Single | 2.381 | 28 | Married | 1.68[illegible] |
| 33 | Single | 1.990 | | | 1.130 | | | 2.43[illegible] |
| 34 | Married | | 26 | Married | 2.060 | | | |
| 36 | Married | 1.640 | | | | | | |

# 19. Transforming to a rectangular format

- If you stack these matrices one beneath the other rather than side by side, you will get closer to a rectangular format.
  - Notice that you have to add another column to denote which matrix is which.

| Breast Feeding Status | Mom's Age | Marital Status | Birth Weight |
|---|---|---|---|
| No | 18 | Married | 1.550 |
| No | 33 | Single | 1.990 |
| No | 34 | Married | |
| No | 36 | Married | 1.640 |
| Yes | 28 | Single | 2.381 |
| Yes | | | 1.130 |
| Yes | 26 | Married | 2.060 |
| Lost | 28 | Married | 1.685 |
| Lost | | | 2.435 |

# 20. Assigning codes

- If you have categorical data, assign a code to each category level. Use the code during data entry to save time and minimize errors.
- Here are some examples of codes:
  - Gender 1=Male, 2=Female, 9=Unknown;
  - Race 1=White, 2=Black, 3=Asian, 4=Hispanic, 5=Native American, 8=Multiracial, 9=Unknown;
  - Likert scale 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree, 9=No answer.

# 21. Assigning codes

- **While I prefer to use number codes, there are some advantages to using short letter codes**.
- Here are some examples of letter codes:
  - Gender M, F, and U (Male, Female, and Unknown);
  - Race W, B, A, H, N, M, and U (White, Black, Asian-American, Hispanic, Native American, Mixed, and Unknown);
  - Likert scale SD, D, N, A, SA, NA (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree, No Answer).

# 22. Number versus letter codes

- Letter codes are easier to remember, and sometimes can be used effectively as plotting symbols.
- **I prefer number codes because they offer more flexibility during statistical analysis**. For example, IBM SPSS will not allow you to draw a scatterplot when one of your variables uses letter codes.

# 23. Binary variables

- For binary variables, I prefer to use 0-1 coding rather than 1-2. It is similar to how computers work (0=off, 1=on). Here are some examples:
  - Treatment: 0=placebo, 1=active drug.
  - Exposure: 0=unexposed, 1=exposed.
  - Disease status: 0=healthy, 1=diseased.
  - Gender: 0=female, 1=male.

# 24. Example using number codes

- Let's assign number codes for the categorical variables in the breast feeding data example.

| Breast Feeding Status | Mom's Age | Marital Status | Birth Weight |
|---|---|---|---|
| 0 | 18 | 1 | 1.550 |
| 0 | 33 | 0 | 1.990 |
| 0 | 34 | 1 | |
| 0 | 36 | 1 | 1.640 |
| 1 | 28 | 0 | 2.381 |
| 1 | | | 1.130 |
| 1 | 26 | 1 | 2.060 |
| | 28 | 1 | 1.685 |
| | | | 2.435 |

# 25. Missing value codes

- Note also that there are still some blank spots in this data. These represents missing data. **Never let a empty field represent missing data**. Explicitly create a code for missing, and be sure to explain why the data are missing to anyone involved with analysis of your data.

# 26. Missing value codes

- Let -1 represent a missing value for Mom's Age and Birth Weight. Let 9 represent a missing value for Breast Feeding Status and Marital Status.

| Breast Feeding Status | Mom's Age | Marital Status | Birth Weight |
|---|---|---|---|
| 0 | 18 | 1 | 1.550 |
| 0 | 33 | 0 | 1.990 |
| 0 | 34 | 1 | -1 |
| 0 | 36 | 1 | 1.640 |
| 1 | 28 | 0 | 2.381 |
| 1 | -1 | 9 | 1.130 |
| 1 | 26 | 1 | 2.060 |
| 9 | 28 | 1 | 1.685 |
| 9 | -1 | 9 | 2.435 |

# 27. Create short names for each column of data

- If you are using a spreadsheet, place a descriptive variable name at the top of each column. If you are using a database, provide a descriptive name for each field. You will use this variable or field name in statistical software like SPSS to specify the variables that you want to analyze. **Try to be reasonably descriptive with your variable names; avoid generic names like VAR01, VAR02, etc.**

# 28. Guidelines for variable names

- Here are some general guidelines that will help avoid trouble with variable names.
  - Use a brief name (eight to sixteen characters long).
  - A mixture of numbers and letters is okay, but avoid special symbols such as $, &, or %.
  - Don't rely on upper/lower case to distinguish among variable names
  - Avoid embedded blanks.

# 29. Use a brief name

- Use a brief name (eight to sixteen characters long). A long time ago, (version 11 and earlier of SPSS), you could not use a name longer than eight charcter long. Now you can use up to 255 characters, but you should show some restraint. It is convenient to have a short "handle" that you can refer to for any column of data in your data set.

# 30. Avoid special symbols

- A mixture of numbers and letters is okay, but avoid special symbols such as $, &, or %. Most statistical software will reserve these special symbols for other purposes. The one major exception is the underscore (_), which is found usually paired on the same key with the minus sign.

# 31. Upper/lower case

- Don't rely on upper/lower case to distinguish among variable names. For example, don't name one variable x and the next one X. Some packages are case insensitive and even if they are not, having two variables with names that look almost identical is a formula for trouble.

# 32. Avoid embedded blanks

- In most statistical software, an embedded blank will cause problems. A variable with a name like “mom age” will possibly be interpreted as two variables “mom” and “age” in certain situations. This can lead to lots of problems.

# 33. Avoid embedded blanks

- But you shouldn’t just strip out the blanks. There's a story about a website for a group known as Writer's Exchange. They used www.writersexchage.com for their website, but someone noticed that this could be read as writer sex change.

# 34. Avoid embedded blanks

- If your variable name consists of two or three short words, here are three strategies that work well.
  1. Use an upper case letter at the start of each new word: MomAge, BreastFeedingStatus.
  2. Separate words using a period: mom.age, breast.feeding.status
  3. Separate words using an underscore mom_age, breast_feeding_status.

# 35. Avoid embedded blanks

- Note that separating using a dash (minus sign) is not a good idea. A variable with the name mom-age looks too much to some computers like a subtraction calculation. Some software programs also use a dash to indicate a range of variables (a1-a8).

# 36. Example of variable names

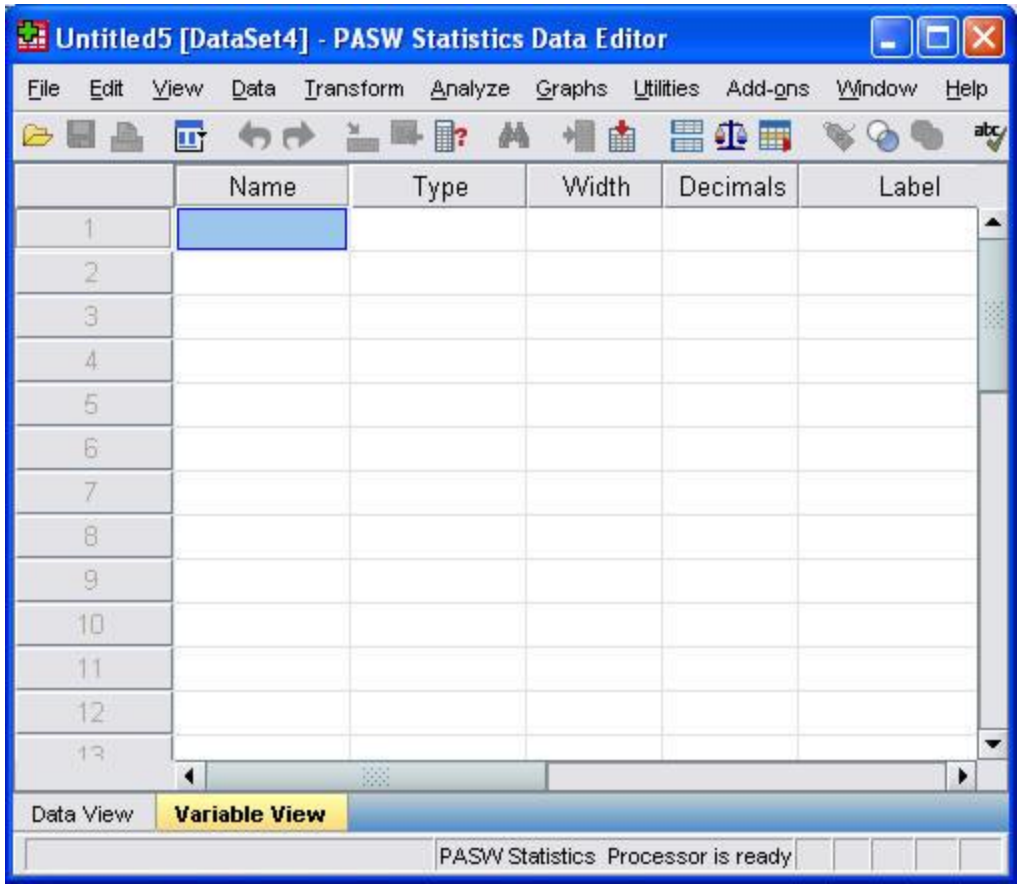
- Here's what the data set looks like with variable names.

| br_feed | mom_age | mar_st | birth_wt |
|---|---|---|---|
| 0 | 18 | 1 | 1.550 |
| 0 | 33 | 0 | 1.990 |
| 0 | 34 | 1 | -1 |
| 0 | 36 | 1 | 1.640 |
| 1 | 28 | 0 | 2.381 |
| 1 | -1 | 9 | 1.130 |
| 1 | 26 | 1 | 2.060 |
| 9 | 28 | 1 | 1.685 |
| 9 | -1 | 9 | 2.435 |

# 37. IBM SPSS Variable View tab

- Once you have names for each column of data, you should document what goes into each column, and control how it is displayed. In IBM SPSS, it starts with selecting the VARIABLE VIEW tab to add documentation to your data.

# 38. IBM SPSS Variable View tab



# 39. IBM SPSS Variable View tab

- From VARIABLE VIEW tab, you can tell SPSS **how to display the data in the SPSS data editor window** (how many decimal places shown, how dates are displayed, and how wide the columns are). You can also provide SPSS with **informational labels that will appear in your output window** (labels for the variable itself and, if needed, labels for category levels). You would also use the dialog box to **specify any codes that represent missing data**.

# 40. NAME

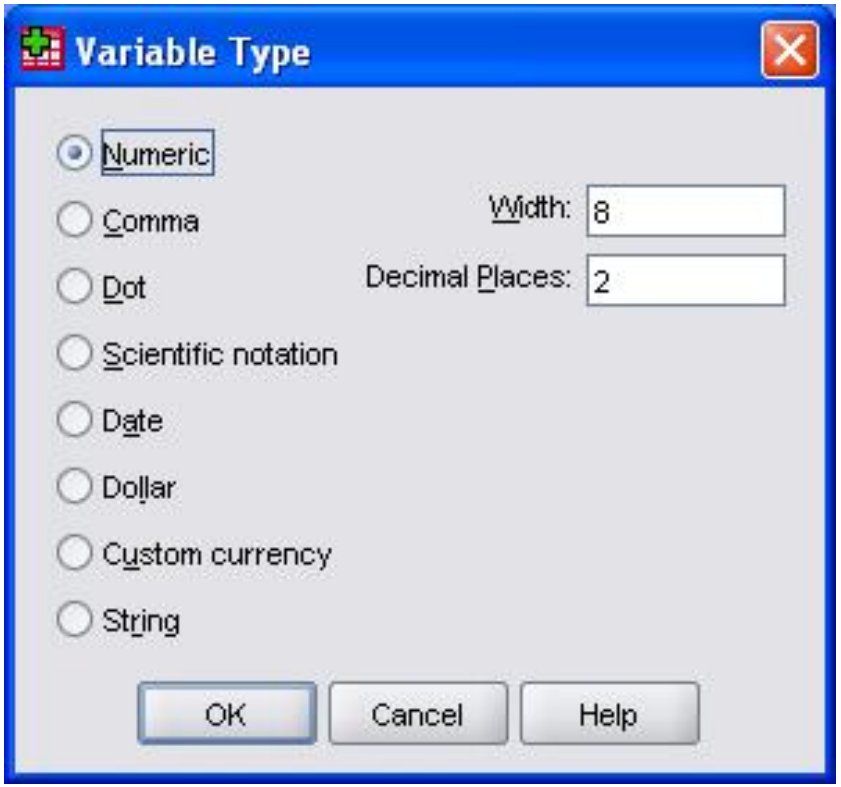
- When documenting your data, your first step should be to **provide a brief but descriptive variable name**. This goes into the NAME column of the VARIABLE VIEW tab. Please spend some time to provide descriptive variable names. As noted above, these names should be short (8 to 16 characters). You will get a chance to provide additional details in the LABEL field.

# 41. TYPE

- WhenClick in the TYPE column to add or change the format type. You will notice a gray button appear on the right hand side. Click on it to get VARIABLE TYPE dialog box.
- This dialog box has information about the type of data that you want to use.

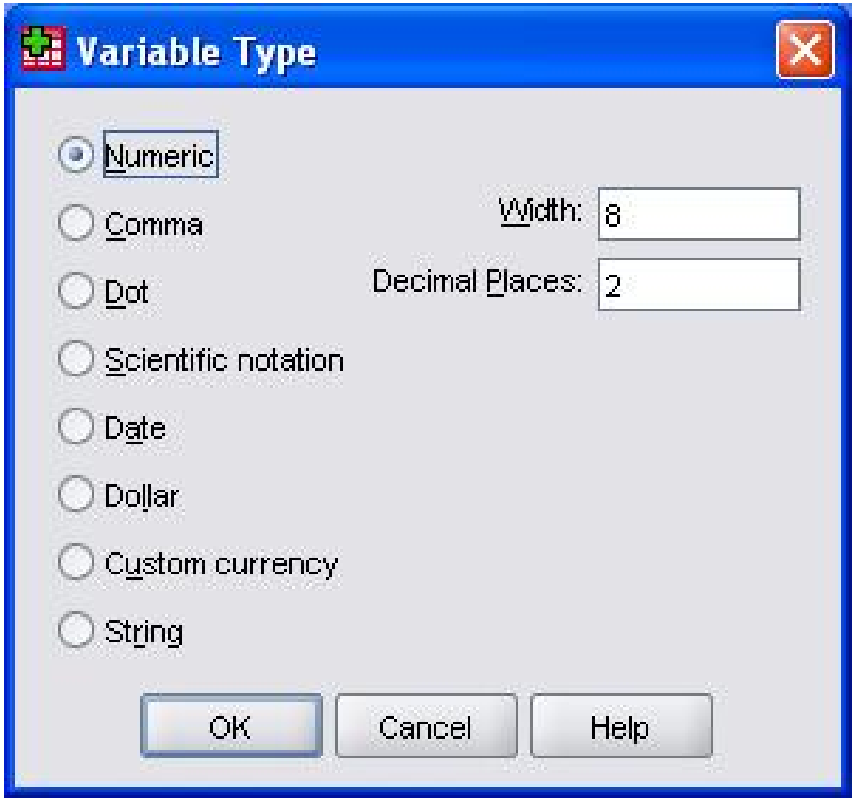
# 42. TYPE (Numeric)

- The most common data type is **NUMERIC, which is used for any data that can be represented solely by numbers**. Unless you are dealing with unusually large numbers, the default width of 8 works well. For some situations, you might be tempted to use a smaller makes, but this can make it more difficult to view the variable name and the value labels.



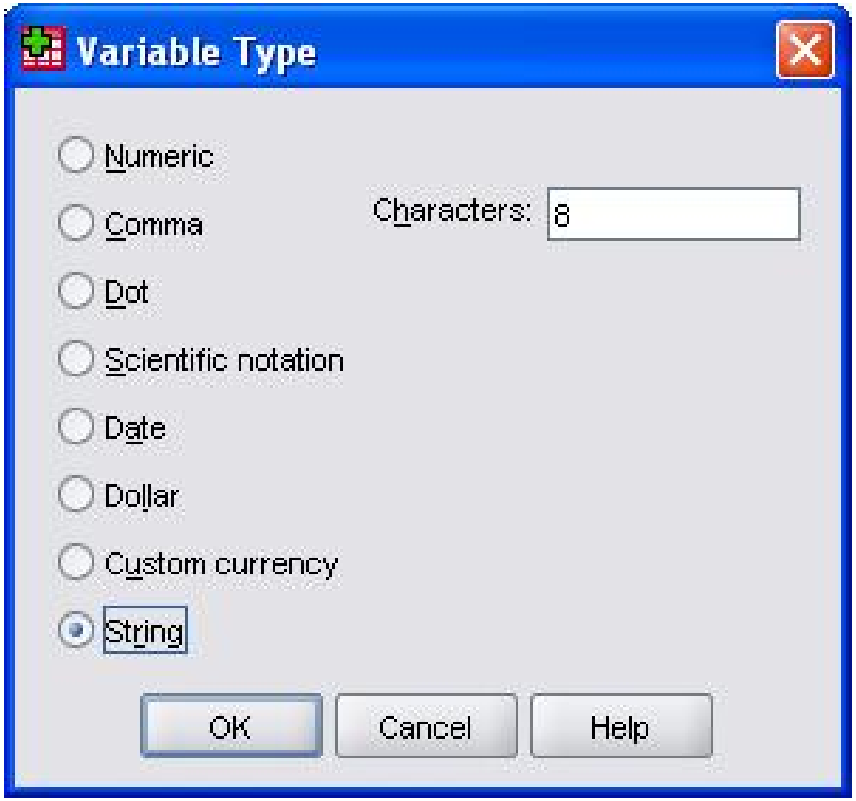
# 43. TYPE (Decimal places)

- Be sure to set the number of decimal places. Please do not display decimal places that you don't really need.



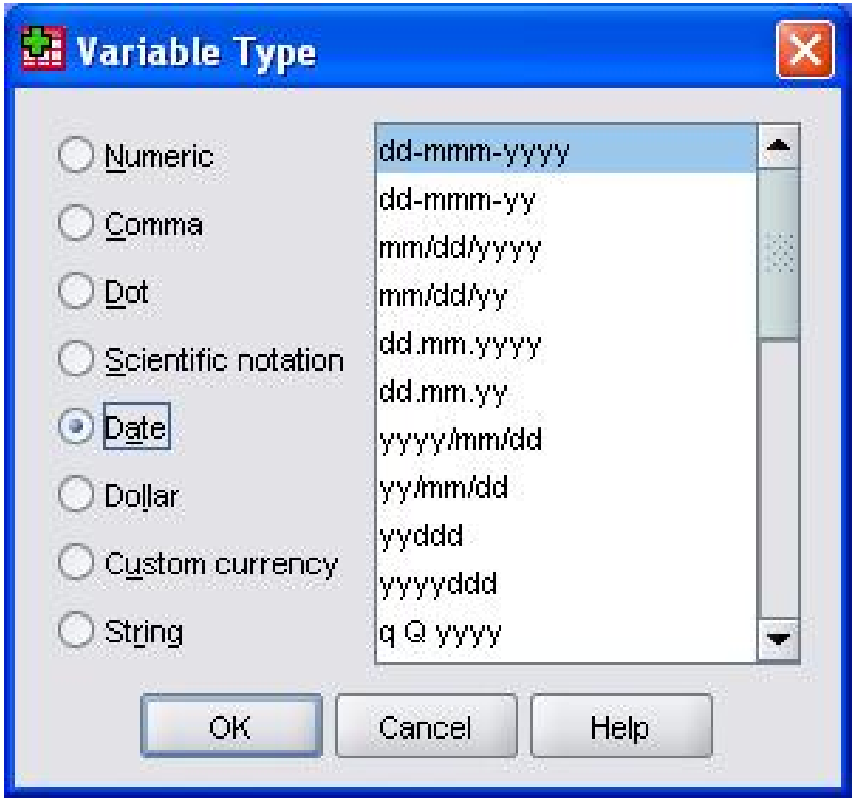
# 44. TYPE (String)

- Select the **STRING options for data that is all letters or a mixture of letters and numbers**. When you select this option, SPSS provides a chance for you to tell how long the strings are.



# 45. TYPE (Date)

- If you click on the DATE option, you will be given choices between various display formats (month names versus month numbers, two digit versus four digit years, etc.).



# 46. LABEL

- Click on the label field to add a variable label. A variable label is a longer description of your data. **Variable labels appear in your output and make it easier to follow what is going on**. You can use a mixture of upper and lower case here, which I recommend for improving readability. AVOID USING ALL UPPERCASE HERE BECAUSE IT IS FAR LESS READABLE THAN A MIXTURE OF CASES.

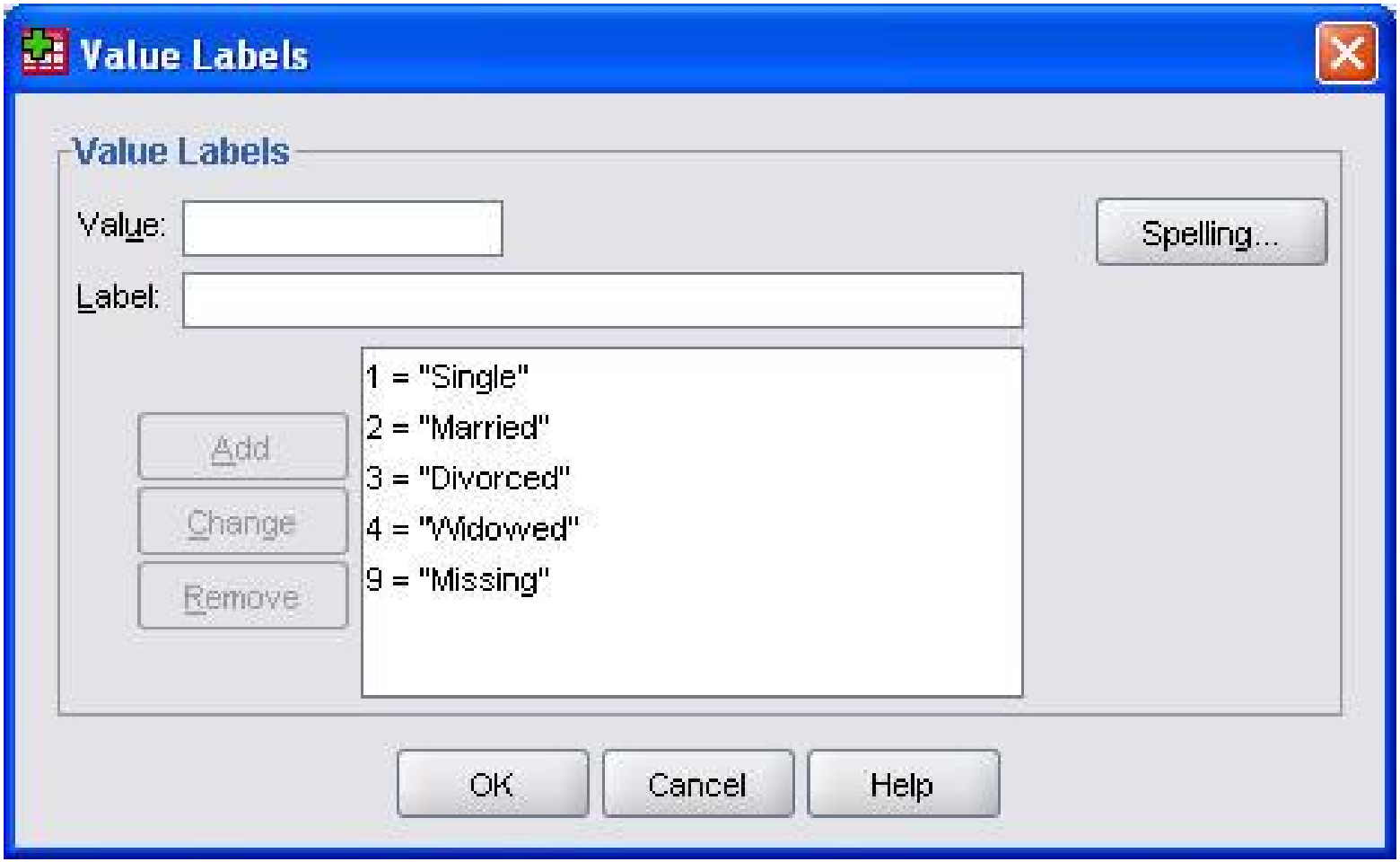
# 47. LABEL

- **You can put blanks and special symbols in your variable label**. If you are very excited about a variable, spice it up with a couple of exclamation points. Go ahead and type to your heart's content. Just a small warning though. A variable label that is too long can make your output look a bit unwieldy. Although you can type up to 255 characters here, it looks strange to have a six inch label underneath a two inch histogram. A variable label of around **20 to 40 characters in length** works well in practice.

# 48. VALUES

- **Value labels provide informative names for levels in any categorical variable**. Leave the value labels blank for continuous data like weight or height. They do make sense, though, for categorical data like gender. This will serve as a reminder that data values of 1 represents males and 2 females. The last thing you want is for people to think that you can't tell the difference between males and females.

# 49. VALUES


Value Labels
Value Labels
Value:
Label:
Spelling...
1 = "Single"
2 = "Married"
3 = "Divorced"
4 = "Widowed"
9 = "Missing"
Add
Change
Remove
OK
Cancel
Help

# 50. MISSING

- If needed, click on the MISSING VALUES button to designate missing value codes. **Missing value codes are useful for designating data in SPSS where the value is unknown, not applicable or otherwise not provided**.

# 51. MISSING

- Be careful about missing values. Make sure you understand why your data is missing and discuss this issue with anyone you are consulting with. **The statistical handling of missing values can vary greatly depending on how the value came to be missing**.

# 52. MISSING

- When you are planning your project, it is a good idea to **select a very clearly impossible code for your missing value**. For example, use -1 for a birth weight because any infant with a negative birth weight would float up to the ceiling after delivery. Use a value of 9 to code missing for gender, since it is obvious to most of us that the number of possible genders is much smaller than 9.

# 53. VALUES


Missing Values
No missing values
Discrete missing values
9
Range plus one optional discrete missing value
Low:
High:
Discrete value:
OK
Cancel
Help

# 54. Summary

- When you are planning your project, it is a good idea to **select a very clearly impossible code for your missing value**. For example, use -1 for a birth weight because any infant with a negative birth weight would float up to the ceiling after delivery. Use a value of 9 to code missing for gender, since it is obvious to most of us that the number of possible genders is much smaller than 9.

# 55. Summary

- To get started with data entry, follow these three steps.
  1. Arrange your data in rectangular format.
  2. Create codes for category levels and missing values.
  3. Create variable names (8 characters or less).

# 56. Two by two table

- Data in two by two tables occur commonly in research, but they are a bit tricky to handle. Here is an example of a two by two table.

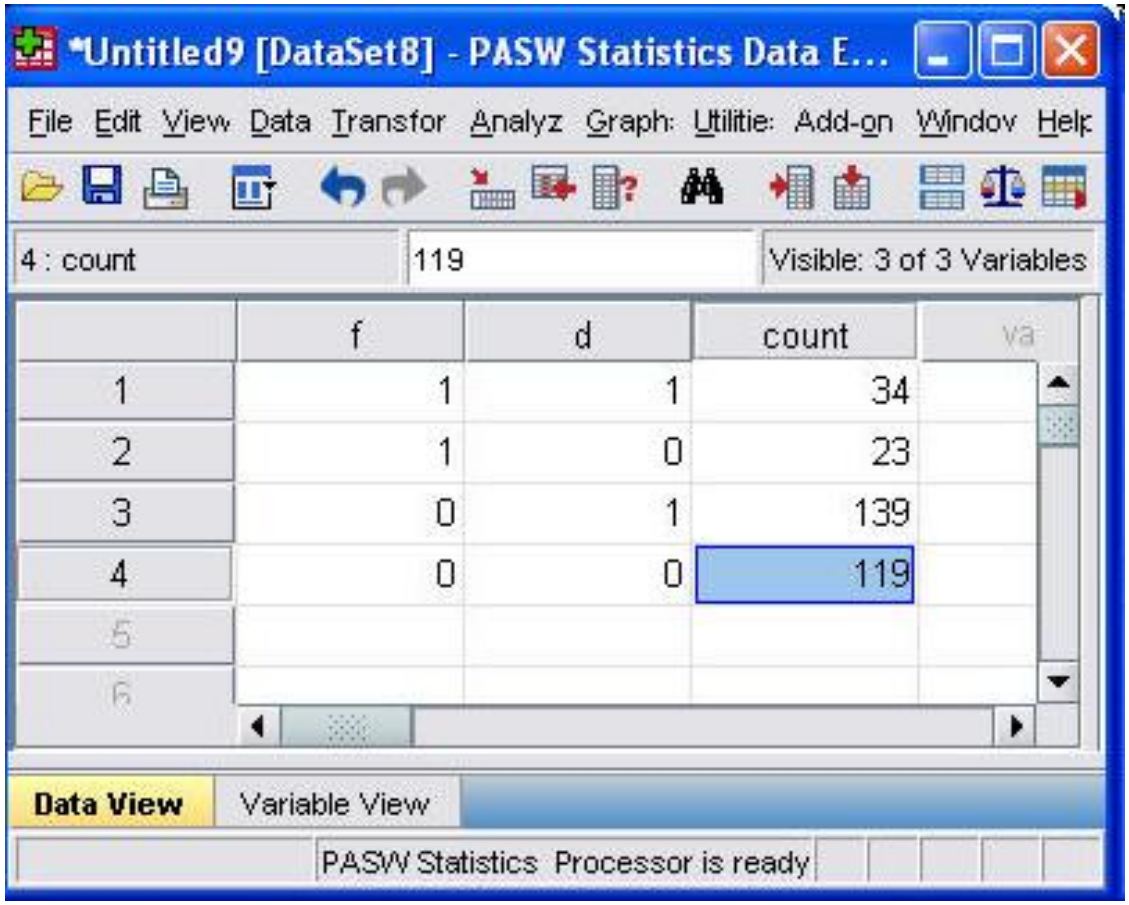
| | D+ | D- | Total |
|---|---|---|---|
| F+ | 34 | 23 | 57 |
| F- | 139 | 119 | 258 |
| Total | 173 | 142 | 315 |

# 57. Two by two table

- For data like this, you have to re-arrange things and then apply weights. To re-arrange the data, you need to specify three variables: F, D, and COUNT.
  - F takes the value of 1 for F+ and 0 for F-.
  - D takes the value of 1 for D+ and 0 for D-.
  - COUNT represents the total number of subjects for each combination of F and D.

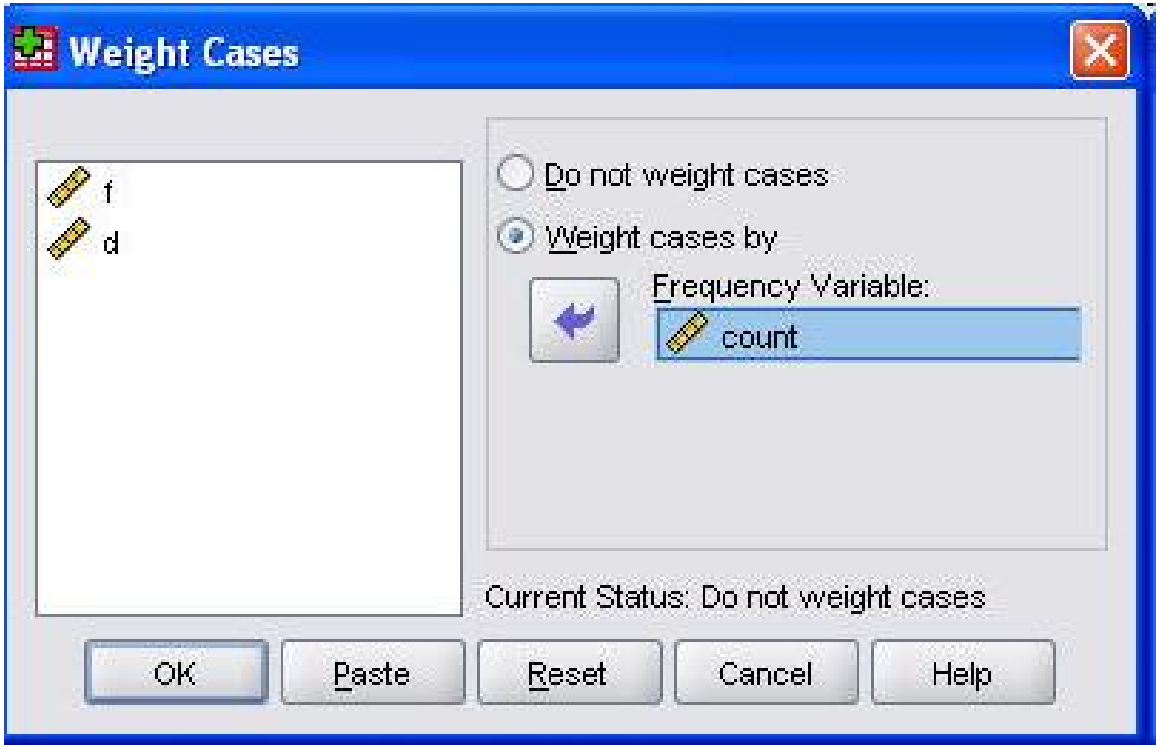
# 58. Two by two table

- Here's what your re-arranged data would look like.



# 59. Two by two table

- **Enter the data, and tell SPSS that W represents a weighting variable**, and you're ready to rock and roll. You do this by selecting **Data | Weight Cases from the SPSS menu.**



# 60. Two by two table

- Here's what a typical output from PSAW would look like.

**d * f Crosstabulation**

| | | | f | | |
|---|---|---|---|---|---|
| | | | 0 | 1 | Total |
| d | 0 | Count | 119 | 23 | 142 |
| | | % within d | 83.8% | 16.2% | 100.0% |
| | 1 | Count | 139 | 34 | 173 |
| | | % within d | 80.3% | 19.7% | 100.0% |
| Total | | Count | 258 | 57 | 315 |
| | | % within d | 81.9% | 18.1% | 100.0% |

# 61. Two by two table

- If you forgot to use WEIGHT CASES, you would get the following.

**f * d Crosstabulation**

| | | | d | | |
|---|---|---|---|---|---|
| | | | 0 | 1 | Total |
| f | 0 | Count | 1 | 1 | 2 |
| | | % within f | 50.0% | 50.0% | 100.0% |
| | 1 | Count | 1 | 1 | 2 |
| | | % within f | 50.0% | 50.0% | 100.0% |
| Total | | Count | 2 | 2 | 4 |
| | | % within f | 50.0% | 50.0% | 100.0% |

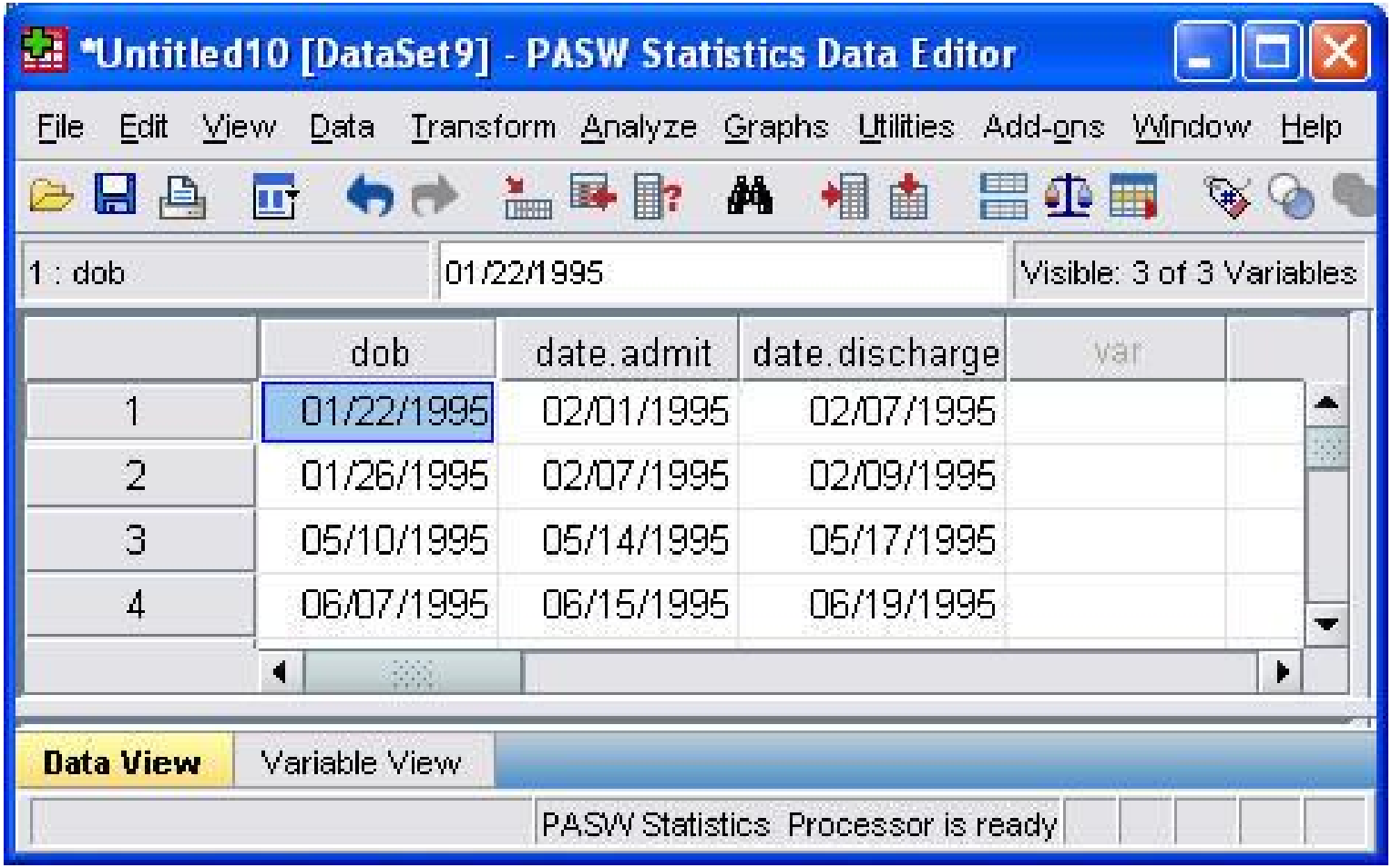
# 62. Dates in IBM SPSS

- A while back I got the following email inquiry:
  - *Dear Professor Mean, I am trying to use dates in SPSS for certain calculations. For example, I want to use a compute statement in SPSS to create a new variable called duration of injury (durinj). I know that I must subtract the date of injury from the date of interview. However, when I do this, I get a number in the millions. What am I doing wrong? -- Stumped Sharon*
  - Dear Stumped, Maybe your patients were waiting for their HMO to approve a visit to a specialist.

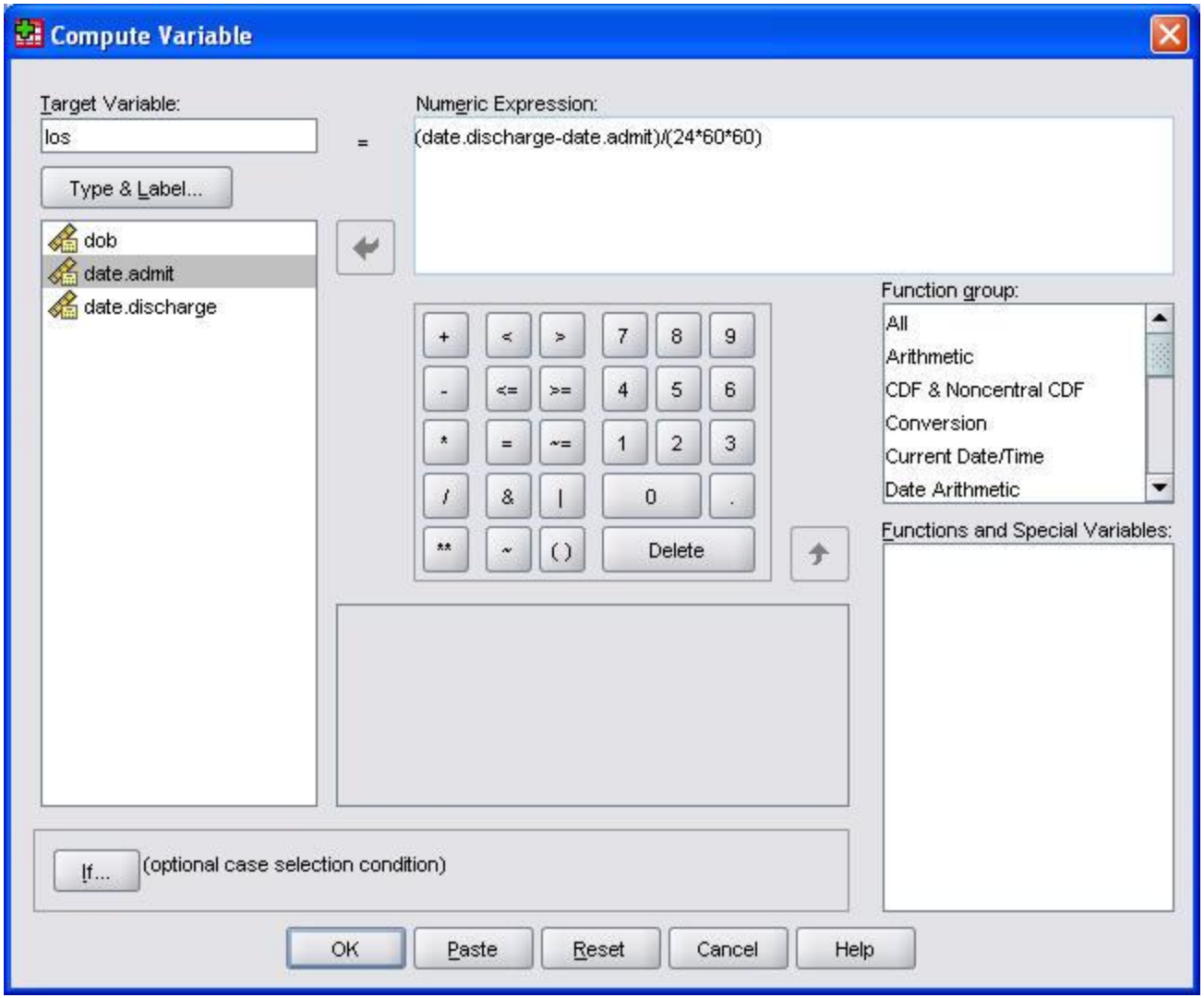
# 63. Dates in IBM SPSS

- Dates in SPSS are actually a bit tricky. **SPSS stores date/time values as the number of seconds since October 14, 1582** (the start of the Gregorian calendar). If you specify only a date and not a time, then SPSS sets the time to midnight. **When you subtract two dates, you get the duration of injury in seconds**. Divide by 86,400 (=24*60*60) to get the duration of injury in days. Divide again by 7, 30, or 365.25 to get duration in weeks, months, or years.

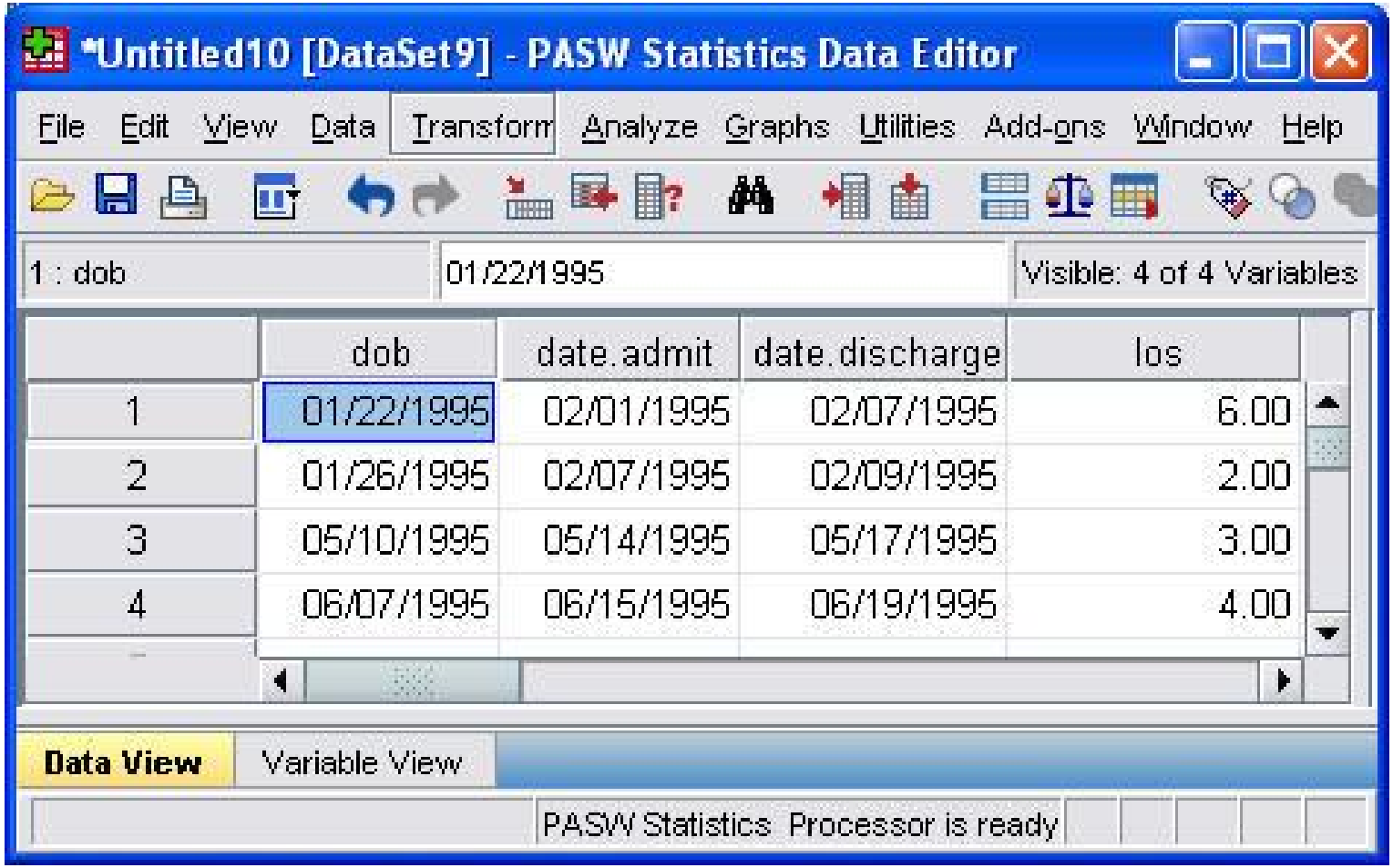
# 64. Date example



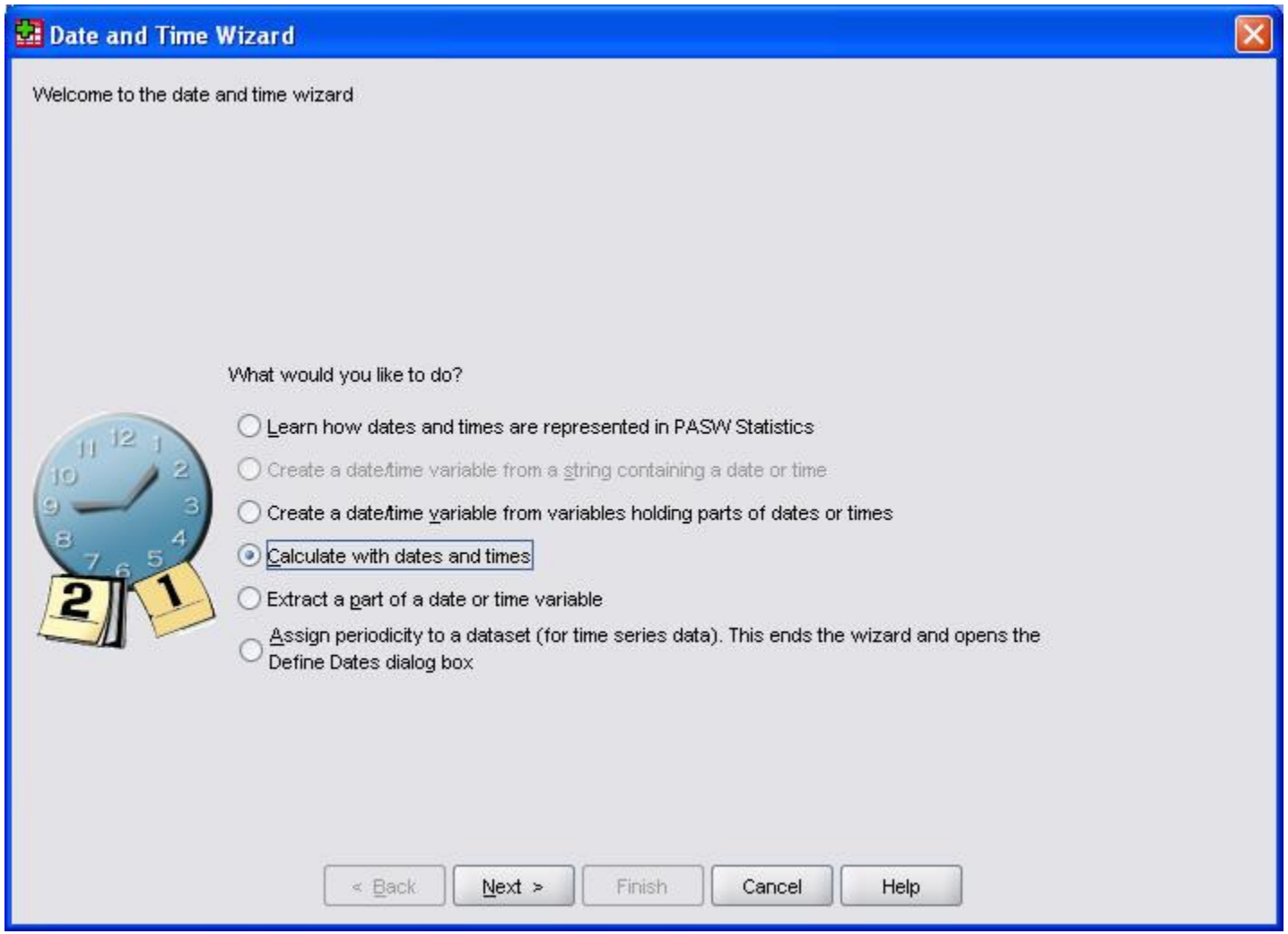
# 65. Date example



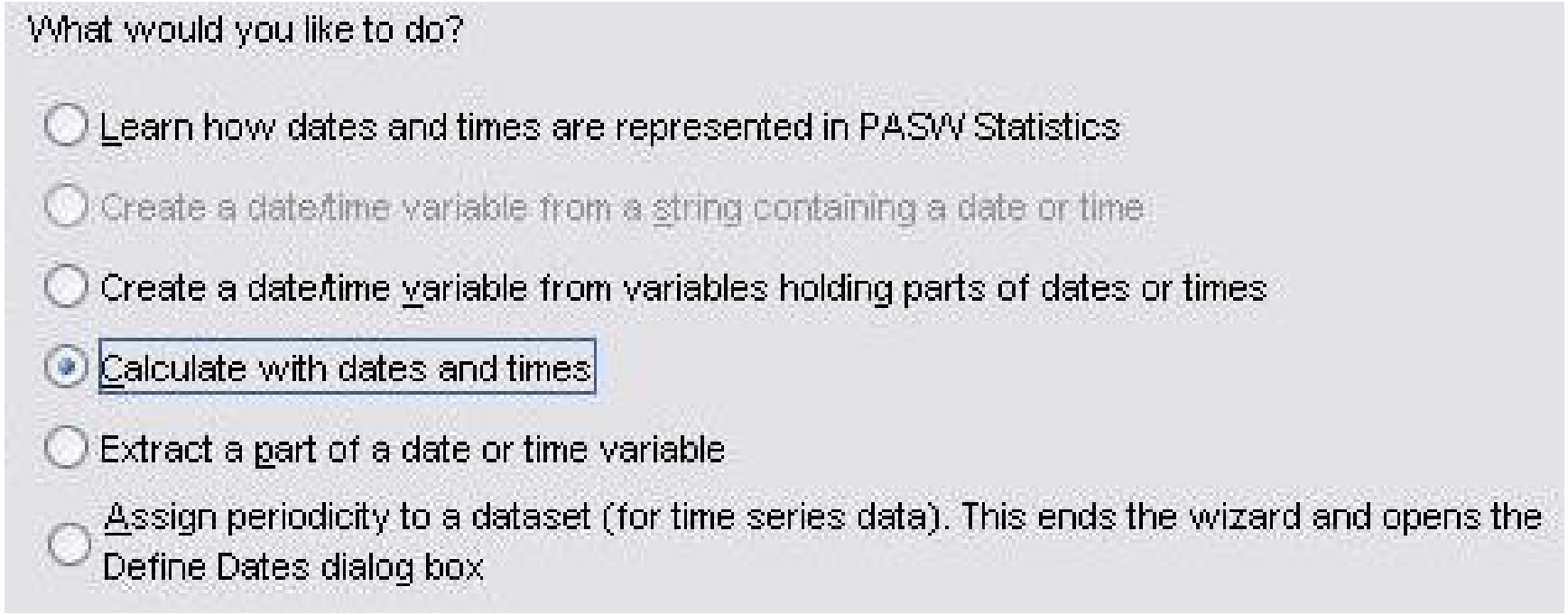
# 66. Date example



# 67. Date and time wizard



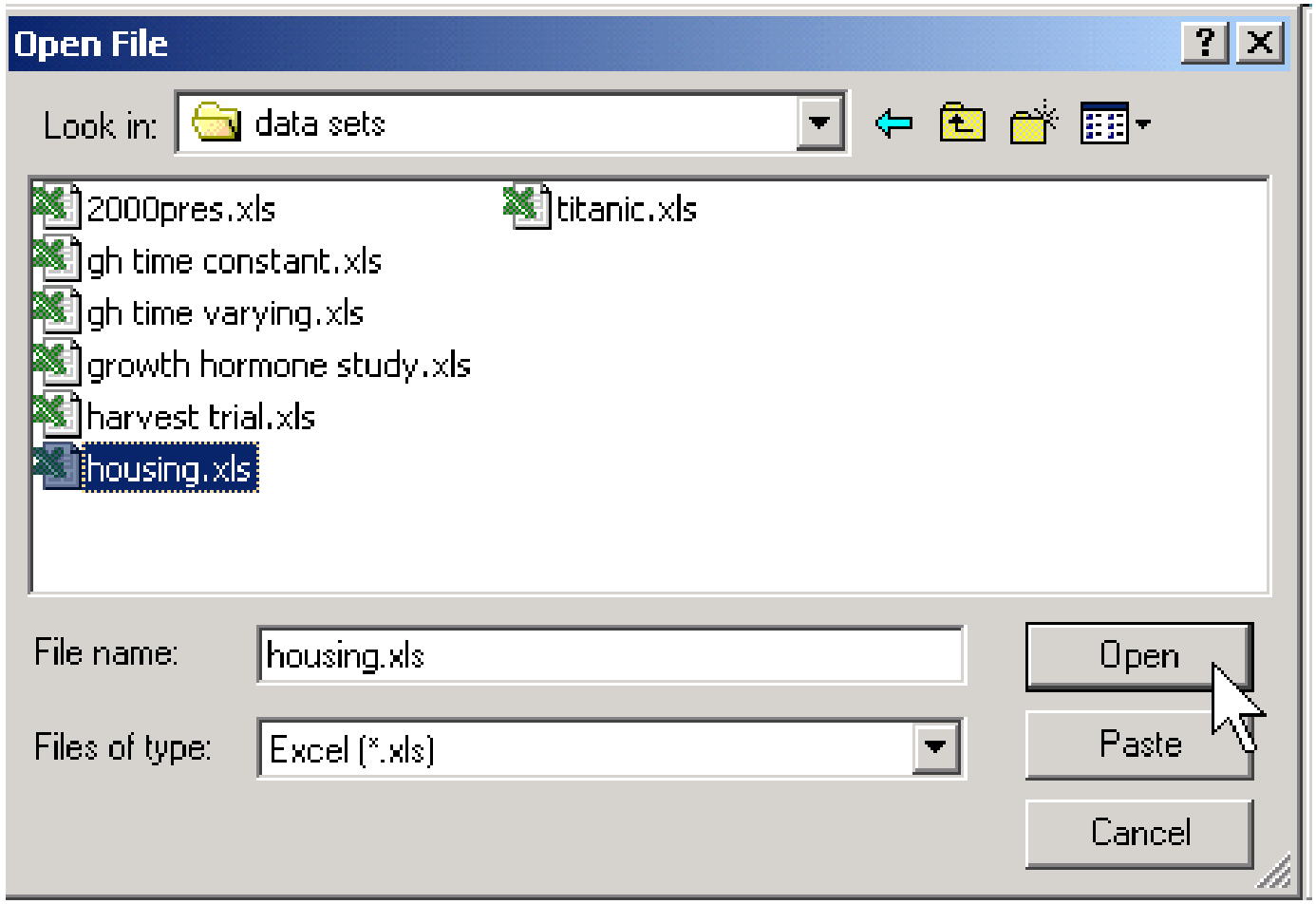
# 68. Date and time wizard



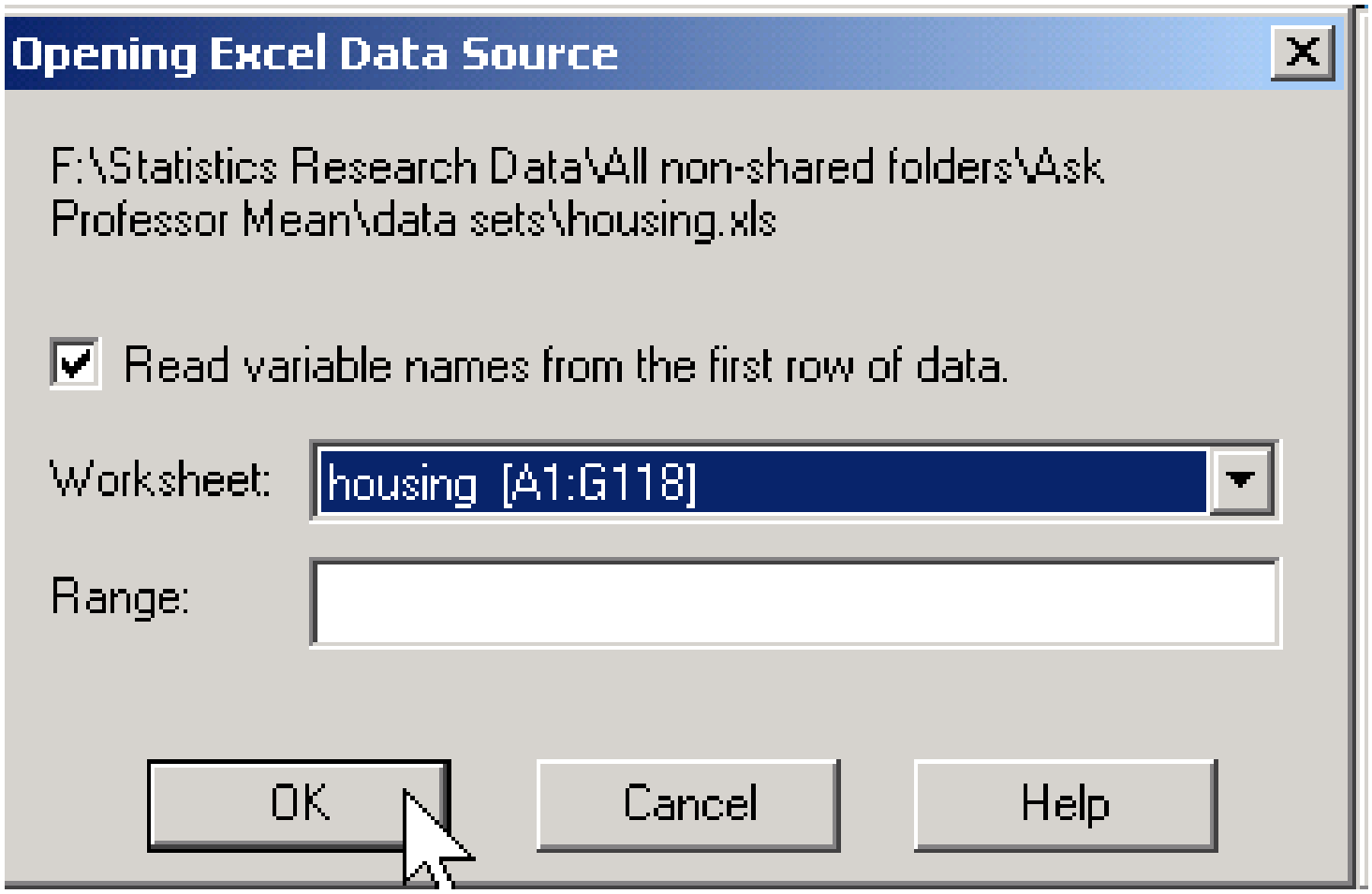
# 69. Importing data from Excel to IBM SPSS

- You should also do some prep work before you import Excel data. Excel is an extremely flexible program that allows you to put your data in just about any way you like.
  1. Arrange the data in a **rectangular grid**
  2. **Don't mix** strings and numbers in a single column.
  3. Put **descriptive names** in your first row.

# 70. Importing data from Excel to IBM SPSS



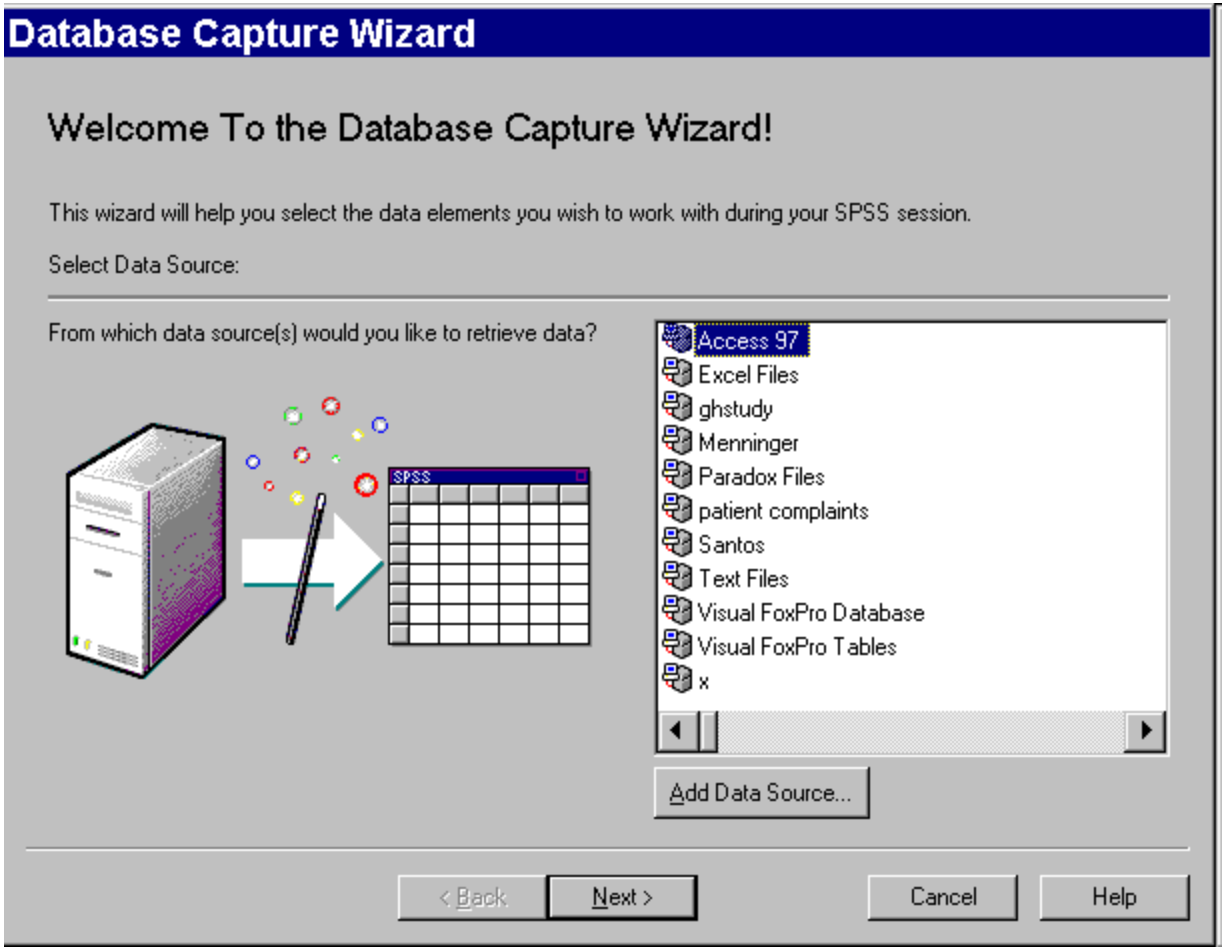
# 71. Importing data from Excel to IBM SPSS



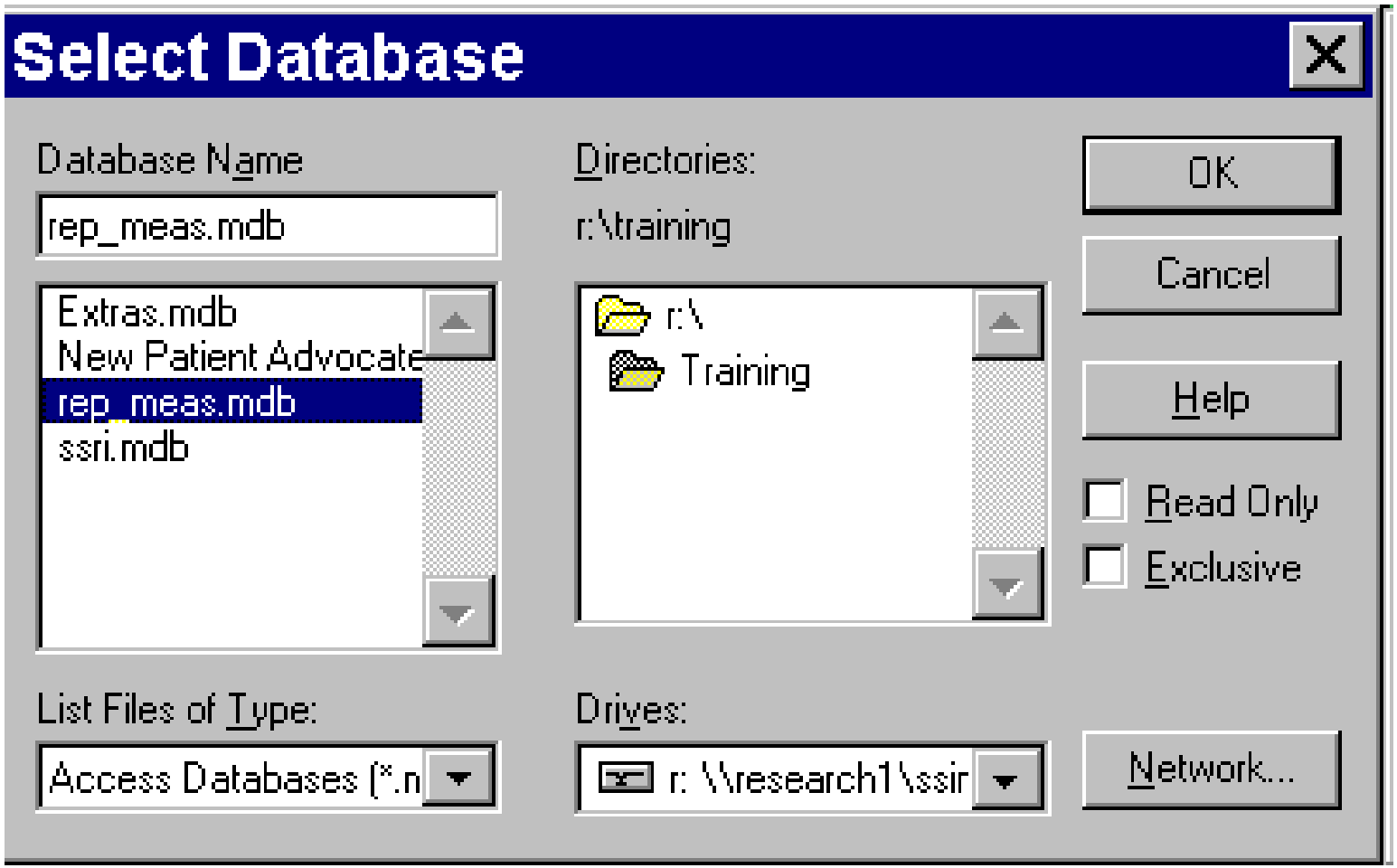
# 72. Importing data from Access to IBM SPSS

- SPSS can import data from a variety of sources using a system known as ODBC (Object Data Base Connectivity). **ODBC has links to just about every database that you would ever need to use**.
- I'll show you an example using Microsoft Access, but this would work just as well on other database systems, such as Oracle and Informix. **To import data from Access, select FILE | DATABASE CAPTURE | NEW QUERY from the SPSS menu.**

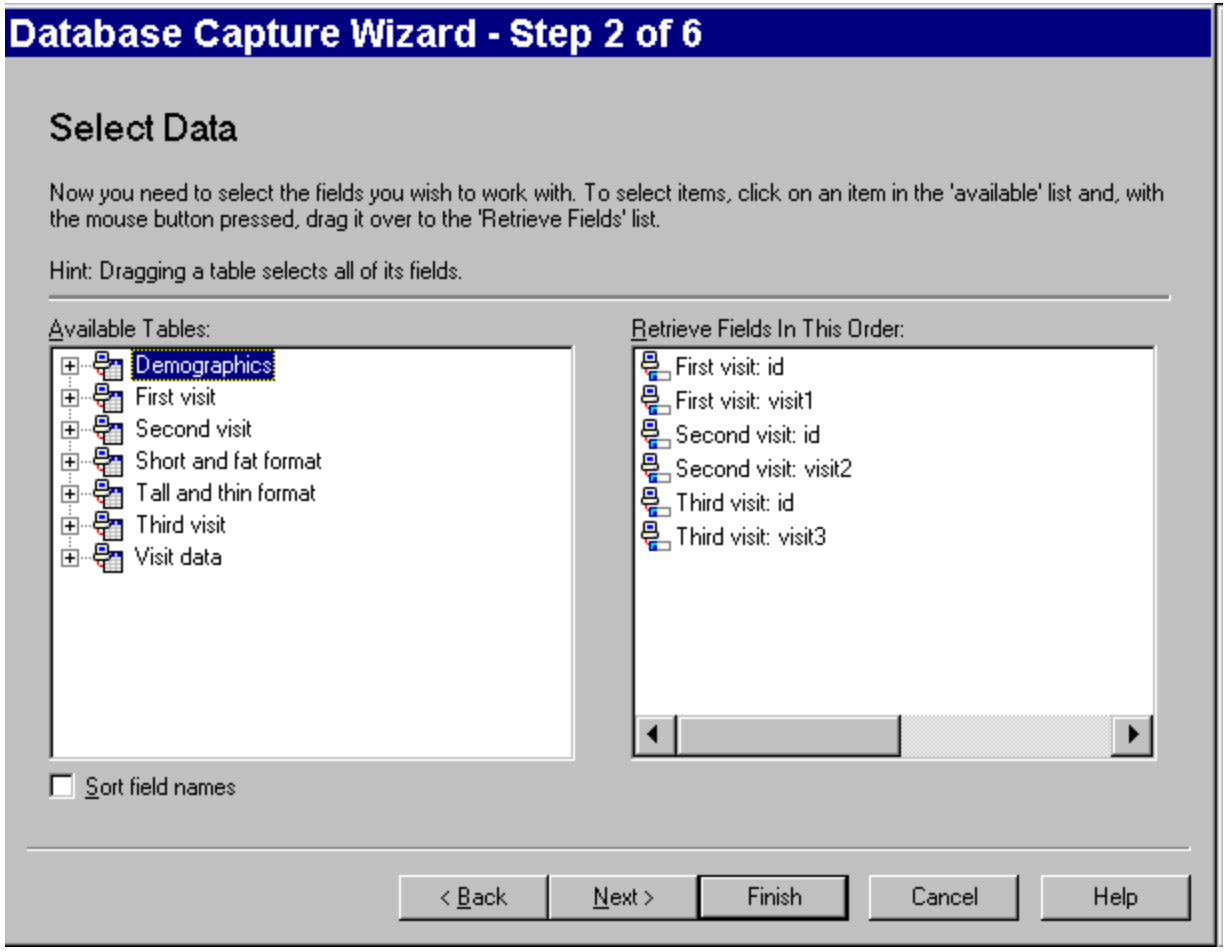
# 73. Importing data from Access to IBM SPSS



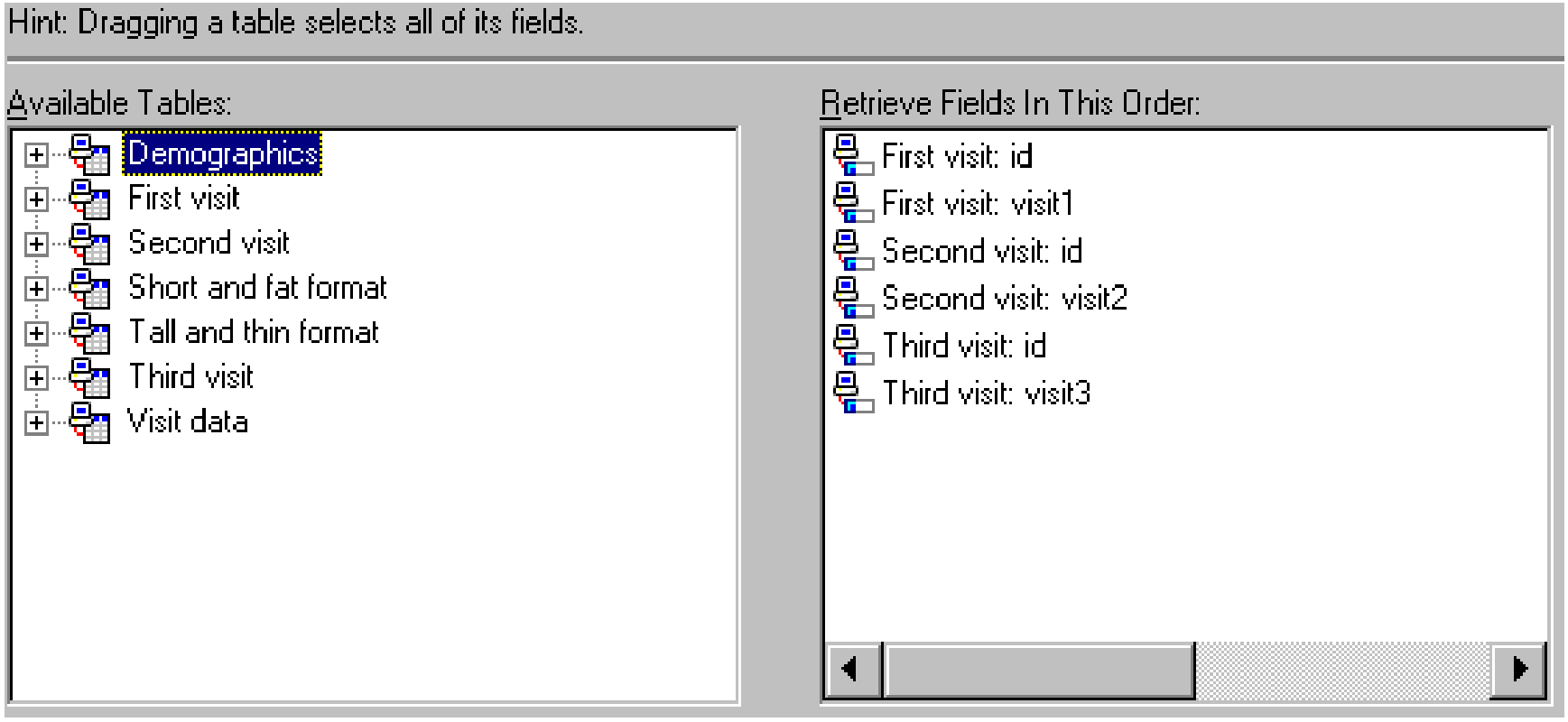
# 74. Importing data from Access to IBM SPSS



# 75. Importing data from Access to IBM SPSS



# 76. Importing data from Access to IBM SPSS



# 77. Pop quiz

1. IBM SPSS provides documentation about the individual levels of a categorical variable using
   - Format type
   - Missing value codes
   - Variable labels
   - Value labels

# 78. Pop quiz

2. In IBM SPSS, if you subtract one date from another to compute the number of days between two events you will get the following result.
   - An error message
   - A missing value
   - A result that is far too large to be correct
   - A warning message

# 79. Pop quiz

3. In IBM SPSS, you can simplify the data entry for a two by two table by using
   - Automatic recode
   - Crosstabs
   - Restructure wizard
   - Weight cases

# 80. The next webinar

Jumpstart Statistics:

Simple descriptive statistics

Friday, March 4, noon-1pm

Details to appear at

www.pmean.com/webinars/jumpstart.html