

# Quantitative Data Input and Analysis using SPSS

Jacob Groshek

[cgroshek@indiana.edu](mailto:cgroshek@indiana.edu)

# Before Opening SPSS

- Data collection is the key
- Levels of measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- These categories determine which data analysis techniques are appropriate
- Higher levels of measurement can be converted to lower levels after collection, but not vice versa

# Levels of Measurement Review

- Nominal: categorical distinctions only
  - religion, gender, musical genres and so on
- Ordinal: rank order without equivalent distinctions
  - horse (and political) races, hierarchical rankings, etc.
- Interval: rank order with equivalent distinctions
  - thermometer, Likert scales of interest, attitudes, more
- Ratio: rank order with equivalent distinctions and true zero
  - age, income, height, formal education, and the like

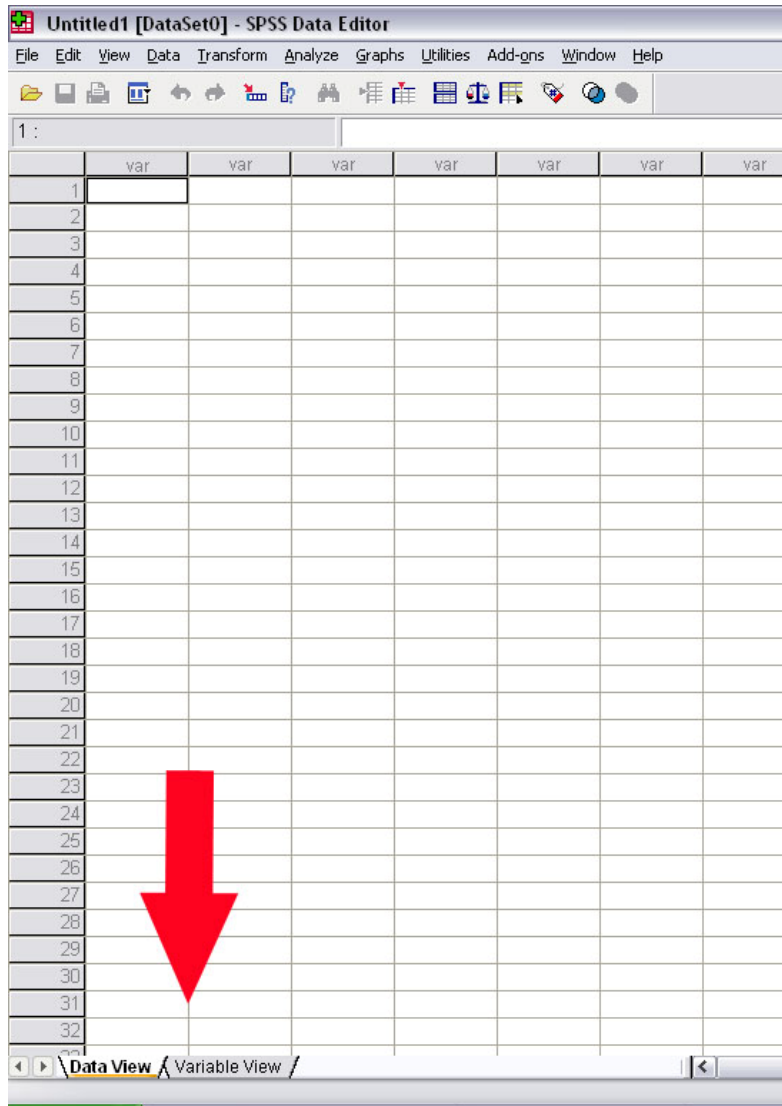
# Using Levels of Measurement

<i>Independent Variable</i>	<i>Dependent Variable</i>	<i>Appropriate Statistical Test</i>
Nominal	Nominal	Chi square
Nominal (two categories)	Interval or Ratio	T-test
Nominal (two or more categories)	Interval or Ratio	Analysis of Variance (ANOVA)
Interval or Ratio	Ordinal	Rank-order Correlation (Rho)
Interval or Ratio	Interval or Ratio	Correlations and/or Regressions

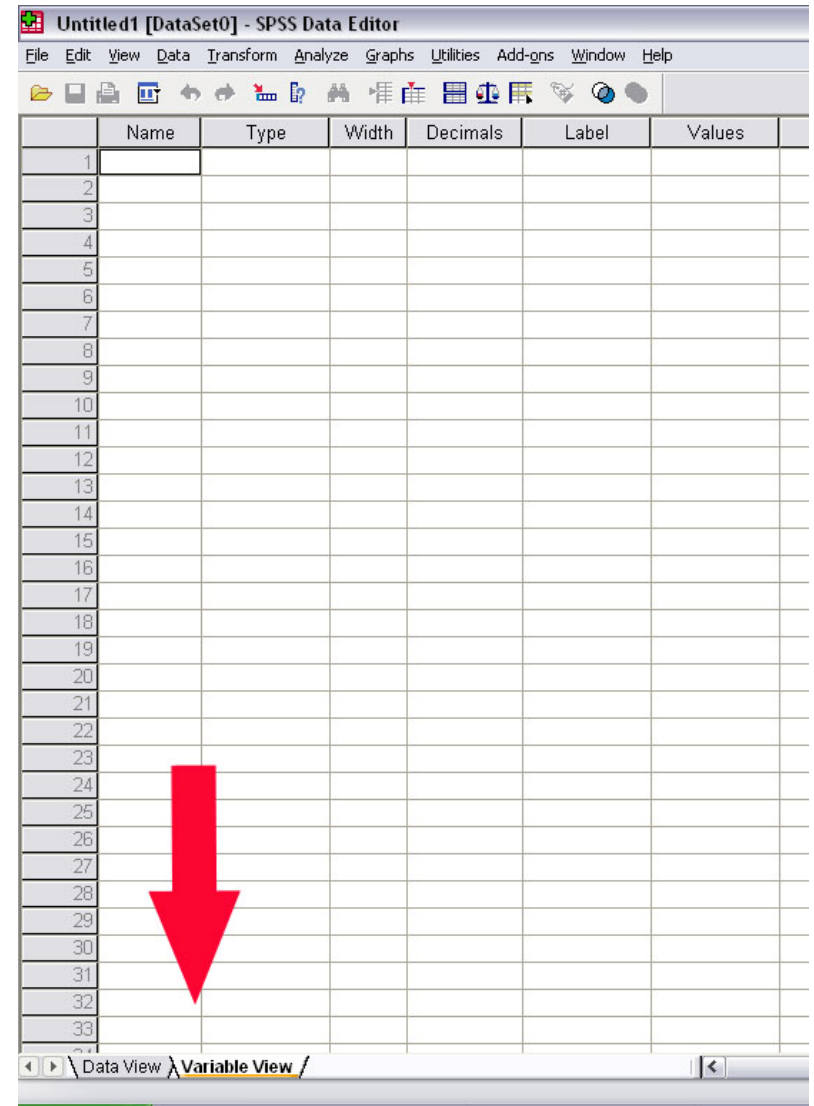
# Data Definition and Entry

- When starting a new dataset, launch SPSS
- Select the *Type in data* option or use the *File* → *New Data* drop down menu
- Opening an existing dataset is the same as opening any typical software program
- One of the most pivotal functions is toggling between the data view and the variable view
  - The data view is where you enter values (numeric or otherwise) of your variables
  - The variable view is where you define and label your variables and their values

# Data View



# Variable View



# Examples of Basic Steps

- Variable View:
  - Naming variables
  - Identifying variable types
  - Labeling variables: A crucial step
  - Notating what values stand for: Equally critical
  - Declaring missing values

# More Examples of Basic Steps

- Data View:
  - Inputting values (*View* → *Value Labels*)
  - Find using the binoculars icon
- Either View:
  - Recoding variables (*Transform* → *Recode into Different Variables*)
  - Computing variables (*Transform* → *Compute Variable*)

# Basic Analytic Commands

- *Analyze* → *Descriptive Statistics* → *Frequencies* is useful for identifying the number of times a value appears and can be used for any level of measurement
  - *Statistics* button allows us to choose specific measures of central tendency (see *Frequencies* handout)

# “Cleaning” and Missing Data

- Frequencies can be used to locate out of range codes
- Similarly, we can find system missing variables
  - Option: go back and find input or coding error if possible
  - Option: leave missing and do not include in analysis
  - Option: replace with mean substitution if continuous (not nominal level) variable

# More Data Cleaning

- Using *Frequencies* can also identify which values represent only a small proportion of cases
  - These outliers may be artificially skewing continuous variables
  - Use *Transform* → *Recode into Different Variables* to group outliers (see Recode handout)

# Basic Analysis: Cross-Tabulation

- Most commonly used to evaluate nominal level variables
  - Can, however, be used with all other levels
  - Remember higher levels of measurement can always be analyzed with lower level tests
  - But not vice versa
- Depending on the number of categories and levels of measurement identified in the contingency table, certain tests are more appropriate than others (pp 162-167)

# Analysis: Chi square

- Chi square tests examine if differences between groups are not due to chance (based on observed and expected values)
- In SPSS, use *Analyze* → *Descriptive Statistics* → *Crosstabs*
  - Under the *Statistics* button, check Chi square
- Report column percentages when the column is the Independent Variable
  - Under the *Cells* button, check column percentages

# Analysis: Chi square

- For any Chi square (and most statistical tests), we need to know the degrees of freedom
  - $df$  is defined as the number of scores that are “free to vary”
  - For chi square  $df = (\# \text{ rows}-1) \times (\# \text{ columns}-1)$
- Evaluating and producing output
  - See crosstabs handout

# Output: Chi square

Americn Focus in Headline \* Which--CNN or CNN International Crosstabulation

			Which--CNN or CNN International		Total
			CNN	CNN International	
Americn Focus in Headline	Yes	Count	285	203	488
		% within Which--CNN or CNN International	71.4%	35.7%	50.5%
	No	Count	114	365	479
		% within Which--CNN or CNN International	28.6%	64.3%	49.5%
Total		Count	399	568	967
		% within Which--CNN or CNN International	100.0%	100.0%	100.0%

Pop Quiz:

Is the relationship observed here due to chance?

How do we know?

Asymp. Sig. is *LESS* than .05 for Pearson Chi-Square value.

Thus, we can be nearly certain ( $p=.000$  in this case) when reporting that framing differences between CNN and CNN International are not due to chance.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	119.417 <sup>b</sup>	1	.000		
Continuity Correction <sup>a</sup>	117.993	1	.000		
Likelihood Ratio	122.481	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	119.293	1	.000		
N of Valid Cases	967				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 197.64.

# Analysis: Comparing Means

- Another (more powerful) way to measure differences between two groups
- Examining changes in mean scores or rankings over time (before and after)
- T-tests compare differences in means against the t (or student) distribution
  - “The probable error of the mean” (Student, 1908)
  - Gossett (the Student author) was a Guinness brewery statistician
  - He was also a student of Pearson’s, who is most noted for developing Pearson’s r correlation measure

# Analysis: T-test

- In SPSS, use *Analyze* → *Compare Means* → *Independent-Samples T Test*
  - $df = \text{sample size} - 2$
- Evaluating and producing output
  - See t-test handout

# Output: T-test

**Group Statistics**

	Which--CNN or CNN International	N	Mean	Std. Deviation	Std. Error Mean
Picture Conflict	CNN	31	2.4839	1.43460	.25766
Frame Recoded	CNN International	74	1.6757	1.13606	.13206

**Independent Samples Test**

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Picture Conflict Frame Recoded	Equal variances assumed	5.033	.027	3.070	103	.003	.80820	.26326	.28608	1.33031
	Equal variances not assumed			2.791	46.514	.008	.80820	.28953	.22557	1.39082

## Pop Quiz:

Is the relationship observed here due to chance?

Levene's Test suggest we should use equal variances not assumed

What should we do? Consider sample size and differences in means; both test specifications are significant so we can report there is difference in the level of conflict framing by CNN network

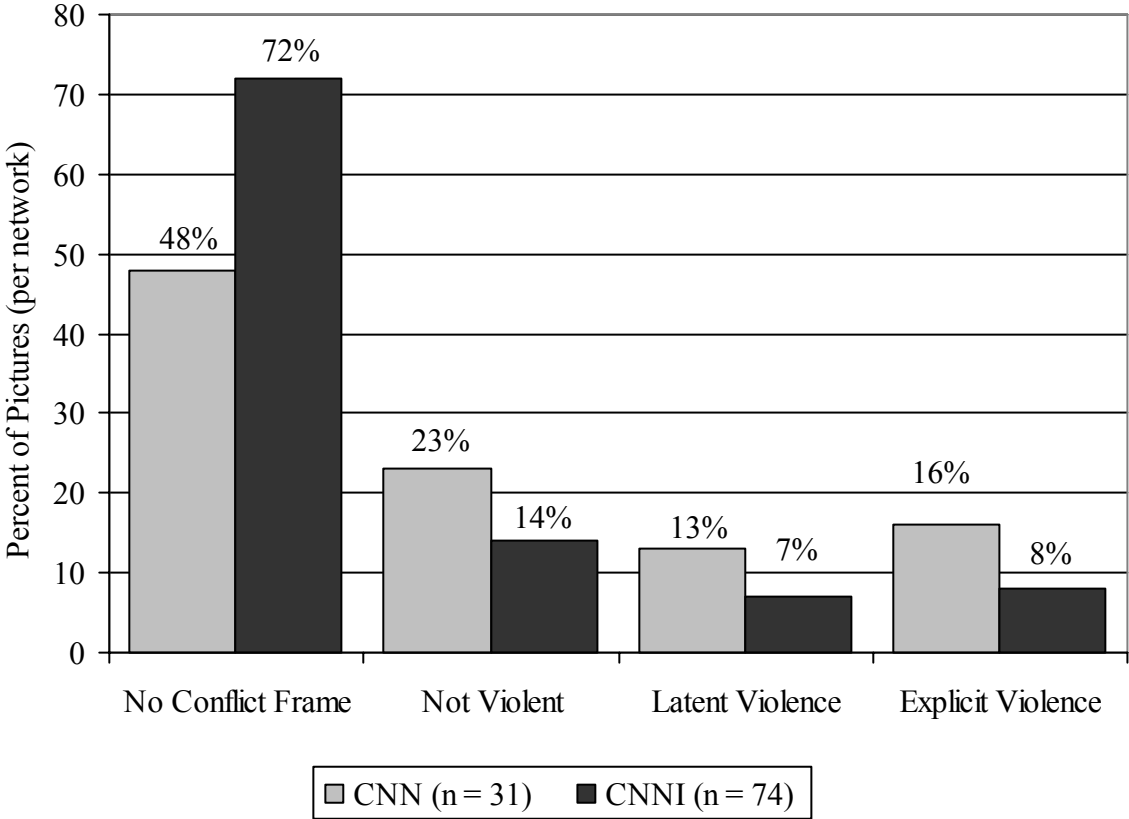
# Guidelines to Reporting Findings

- Most important question: Does the table or graph make interpretation easier?
- If not, you probably don't need it
- SPSS is not a very friendly graphics program; use Word or your preferred editor
- The most important resource you have is the APA Manual (or that of whichever reference style you are using)

# Guidelines to Reporting Findings

- Visualizations should offer an honest, straightforward representation of findings
  - Use double spacing; avoid bolding
  - Only one chart/table/figure per page
  - Report significance levels
  - Write concise titles and labels
- Show your graphs to someone who is completely unfamiliar with your study and gauge their response
- A few examples follow

Figure 1 Frequency of Conflict and Violent Imagery in Coverage on CNN and CNNI



Note:  $\chi^2_3 = 5.2, p = .155$

Table 1 Average Levels of Conflict and Violent Imagery in Coverage on CNN and CNNI

	<i>Research Question 6: Network Differences</i>	
<i>Conflict Framing:</i>	CNN (n = 31)	CNN International (n = 74)
Mean Levels of Violent Imagery	2.48	1.68**

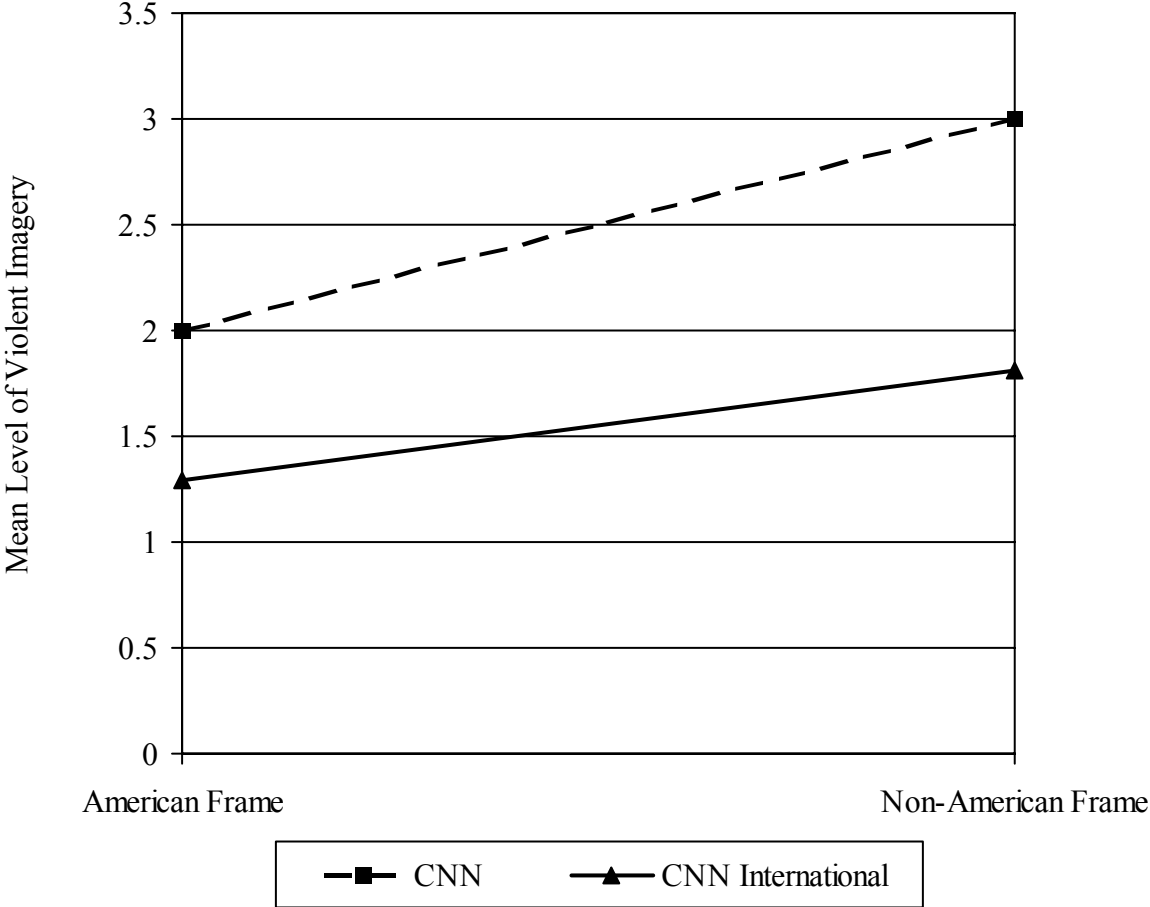
*Note:* \*\*  $p = < .01$ , equal variances assumed

Table 2 Coverage of News Topics on CNN and CNNI

Category	CNN Rank	Percentage	CNNI Rank	Percentage
Crime/law and order	1	25.3%	2	17.3%
Politics	2	13.8%	1	19.4%
War/terrorism	3	13.0%	3	15.5%
Business/economics	4	10.3%	5	6.9%
Health care	4	10.3%	5	6.9%
Oddities	6	9.0%	8	6.0%
Religion/culture	7	6.0%	4	10.6%
Accidents/natural disasters	8	4.8%	7	6.5%
Sports	9	2.8%	9	3.0%
Others	10	4.8%	10	8.2%
Total		100%		100%

Note: Spearman's rho  $r = .89, p = .001$

Figure 2 Relationships Between Levels of Violent Imagery, Frames, and Networks



Note: Main effect between level of violently depicted conflict and American framing ( $F(1, 105) = 5.49, p = .021, \eta_p^2 = .052, \text{observed power} = .640$ )

# Quantitative Data Input and Analysis using SPSS

--PLEASE--

## Feel Free to Contact Me

Jacob Groshek

[cgroshek@indiana.edu](mailto:cgroshek@indiana.edu)