DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Research Methods Posc 302

MANAGING AND ANALYZING DATA
(Continued)

I. TODAY'S SESSION:
- A. Refinement of chi square
- B. Measures of association and correlation
- C. Writing tips:
  1. Make sure paragraphs are connected by a transition sentences.
     - i. The last sentence of a paragraph should lead naturally to the first sentence of the next one, even if there is table or graph between them.
     - ii. Example:
       1) ***Thus we see that there are important differences between the age cohorts.***
       [Table]
       Social and economic factors such as family income and mobility, however, may explain this relationship and in particular why older people seem more conservative than the younger generation. For example, if we control for...
  2. Make sure that all of the following are properly cited:
     - i. Direct quotes
     - ii. Sources of data
     - iii. Specific references to facts and claims.
  3. Unless you are 100 percent certain you know the proper style use either ***The Chicago Manual of Style*** or MLA formats.
     - i. The course web page has a list of on line guides that Manual of Style and MLA standards. (See http://www.udel.edu/htr/Researchmethods/pollsurvey.html#guide)

II. CHI SQUARE:
- A. Refinement:
  1. The chi square statistic should always be reported with the degrees of freedom that will always be a part of any computer report.
- B. Example:
  1. The chi square for the association between party identification (summary measure) and vote for president in 1996 as reported by the SDA program is: chi square = 639.14 with **6 df** ($P < .0001$).

2. The symbol **df** stands for "degrees of freedom."

C. Although they are always reported, you can easily calculate the degrees of freedom for any chi square based on a simple contingency table from the formula: $df = (R - 1)(C - 1)$, where R is the number of rows and C the number of columns in the table.

1. Example: in the partisanship by vote table there are 2 rows and 7 columns, so the degrees of freedom is df = (7 - 1)(2 - 1) = (6)(1) = 6.

III. PRESENTING QUANTITATIVE DATA:

A. As noted in the previous class, contingency tables can be used to present all of the data collected for the research projects.

1. However, you might want to draw bar or line graphs in order to make your points visually.

i. These tools sometimes save space but usually at the cost of information.

2. Although the choice is up to you, make sure that you follow the guidelines discussed throughout the semester.

B. (Partial) check list:

1. Tables:

i. Title

ii. Columns fully labeled with no abbreviations

iii. Contents of the table indicated either in the table with symbols or in the title with a subtitle.

1) Example 1:

Table 1
Party Identification By Region
(Percentages)

	North	Midwest	West
Democrat	58	44	38
Republican	42	56	62
Total	100 (431)	100 (337)	100 (401)

2) Example 2:

Table 2 Party Identification By Region			
	North	Midwest	West
Democrat	58%	44%	38%
Republican	42	56	62
Total	100% (431)	100% (337)	100% (401)

iv. Question wording and data source must be cited.

1) Example 3:

Table 3 Party Identification By Region[1] (Percentages)			
	North	Midwest	West
Democrat	58	44	38
Republican	42	56	62
Total	100 (431)	100 (337)	100 (401)
Regions: North: Vermont, New Hampshire...Midwest: Illinois, Indiana, Kansas, Ohio...etc.			

2) Example 4:

[1]Data: 1996 American National Election Study.

Table 4
Party Identification By Region
(Percentages)

	North	Midwest	West
Democrat	58	44	38
Republican	42	56	62
Total	100 (431)	100 (337)	100 (401)

For question wording see Appendix B.
Data: General Social Survey Cumulative File

2. Graphs:
   i. Keep ink and color to a minimum
      1) No three dimensional bars, please.
      2) No pie charts
   ii. Scales and axes must be labeled.
      1) Example 1:

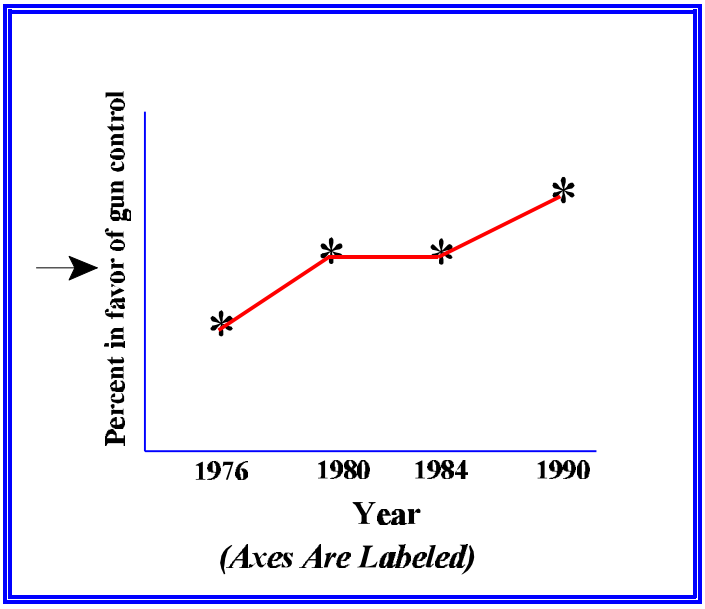

Figure 1: Support For Gun Control 1976 - 1990. Data: General Social Survey Cumulative File

      2) Example 2:

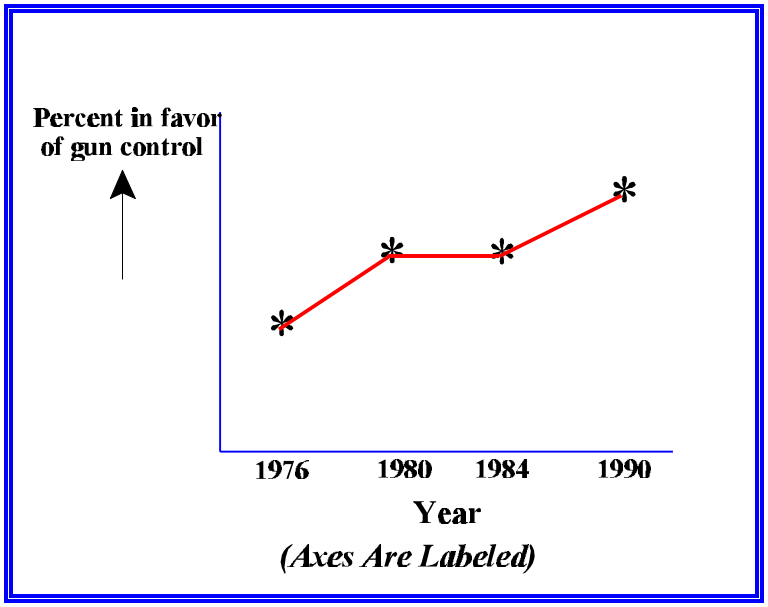

Figure 2: Support For Gun Control 1976 - 1990. Data: General Social Survey Cumulative File

IV. ASSOCIATION AND CORRELATION:

A. **Association**: the values of one variable are related to, tend to be observed in conjunction with the values of another variable.

1. This idea underlies what we have been discussing all semester.
2. Example: values of the variable party identification are often related to the values of political attitudes and beliefs such as the state of the economy or President Clinton's moral character.
3. Statistical independence is the absence of a relationship: the values of one variable are independent of the values of another.

B. Measuring association.

1. We have used contingency tables and to a lesser extent bar and line charts to summarize or display **how** one variable is associated with or related to another variable.
   i. Think about a cross-tabulation: it shows how many people or the percentage of people in one category of X are in particular categories of Y.
   ii. Or, a line graph shows how what percent of the respondents in a given year have some value of a dependent variable.
2. In some instances we can add to these tables or graphs a single numerical indicator that tells one how **strongly** one variable is related to another.
3. This idea is best seen by introducing the notion of correlation.

V. CORRELATION:

A. We are all familiar with patterns among variables.

1. Height and weight are related.
2. So too are number of years of formal education and income.
3. These are examples of relationships called **correlation**.

B. Consider two numerical (quantitative) or ordinal variables, X and Y.

1. Correlation does not apply to qualitative or nominal variables.
2. That is we can speak meaningfully of the correlation between amount of education and income or between age and strength of party identification because the variables are implicitly quantitative or numerical.
   i. But it won't make sense to take about the correlation between "region" and family income because although income is numerical, place of residence is strictly a qualitative variable. Its categories don't stand for numerical quantities.

C. Definition:

1. **Positive correlation: high values of Y are associated with high values of X and, conversely, low values of Y are associated with low values of X.**
2. **Negative correlation: high values of Y are associated with low values of X and, conversely, low values of Y are associated with high values of X.**

D. Pictures may help clarify the notion:

1. Figure 4 shows a "scatterplot"

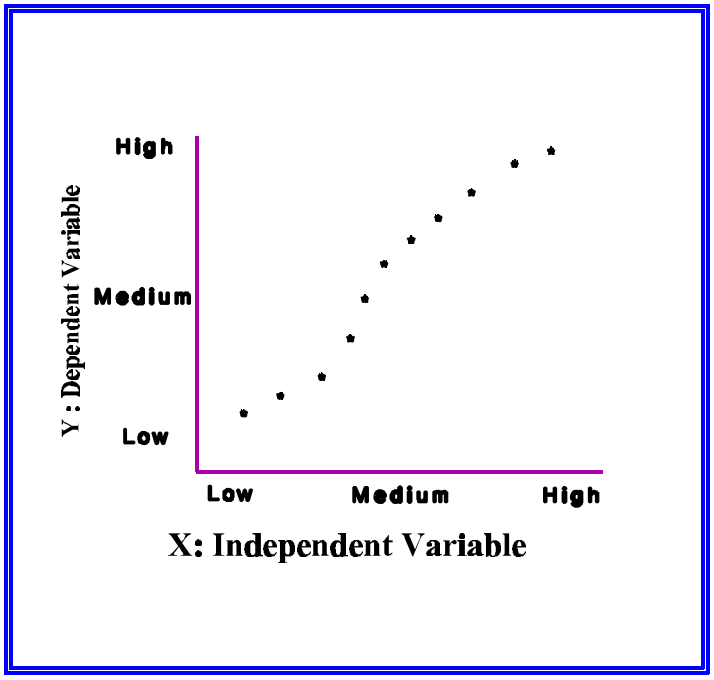

Figure 4: **Example of Correlation**

2. The values of the independent variable are arrayed along the X-axis (horizontal) of a graph, those of the dependent variable along the vertical or Y-axis.
3. The X and Y values of each observation are plotted on this coordinate system.
   i. If, for example, an observation has scores or values of Y = 10 and X = 78, the plot would contain a mark an "o" or "*" (say) at the intersection of 10 and 78.
   ii. A scatterplot contains marks (circles or asteriks or whatever) for each observation according to their values on X **and** Y. The result is a "scatterplot." The plot graphically shows how the values of Y and X are related (if, in fact, they are).
4. Interpretation: notice in the above example that as the value of X increases, the value of Y also increases.
   i. That is, there is a tendency for small values of X to appear with small values of Y, for medium values of X to appear with medium values of Y, and large values of X to appear with large values of Y.
5. This pattern of association (Figure 4) indicates a **positive correlation** between Y and X.
   i. Sometimes one refers to the "direction" of the correlation as being positive.

E. The next figure shows a different pattern of correlated observations

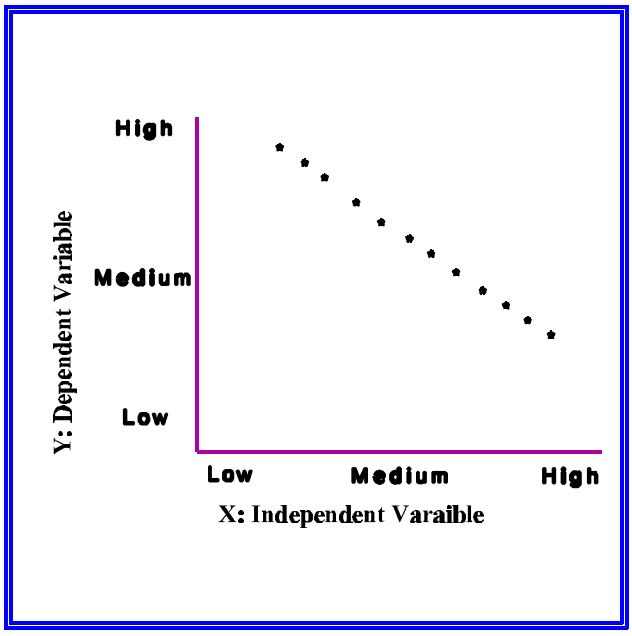

Figure 5: **Perfect Negative Linear Correlation**

i. Here there is a pattern but of a different sort: large values of X are associated with **small** values of Y and small values of X are associated with **large** values of X.
ii. The variables are **negatively** correlated.
1) Or, the direction is negative.

2. The next figure illustrates a pattern of correlation that is less pronounced than the previous two.

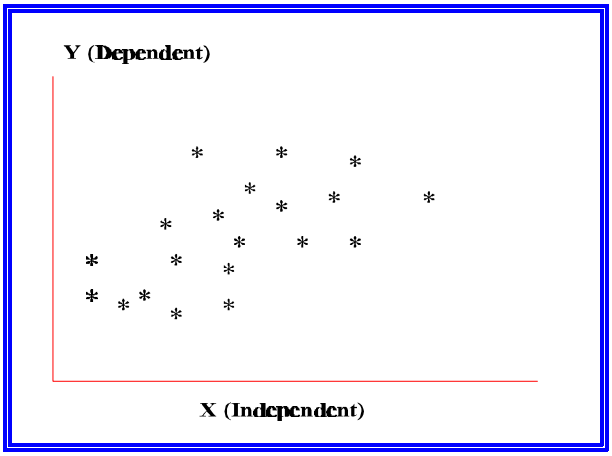

Figure 5: Weak (Positive) Correlation

i. The values of X tend to follow the pattern of positive correlation described above, but the association is "weaker."

F. Summary:
1. Assume people in a survey or other study are measured on two numerical variables such as years of education and annual income.
2. In principle we could plot their income and partisanship values on an X-Y axis system.
3. No doubt the "points" would follow a rough pattern: those people whose education is low would **tend** to have relatively low incomes while those with more education college degrees would have higher incomes.
4. If there were a perfect relationship, the points might all lie on a straight line.
5. In practice we wouldn't expect to find such a perfect pattern so the points would not doubt be scattered.
6. But nevertheless they would "drift" upward.
7. By looking a the pattern we could say that the variables were (positively) correlated but it wouldn't be possible to say precisely how strongly.

G. Although pictures a useful for understanding the concept of correlation, we need to have a numerical indicator, a statistic, that precisely describes the strength and

direction of X-Y relationships.

VI. THE CORRELATION COEFFICIENT:

A. A the correlation coefficient, often called Pearson's r or the product-moment correlation, is perhaps the most commonly cited and used measure of **linear** association or correlation in the social and policy sciences.

1. It "tells numerically" how and to what degree two variables are correlated.
2. A more formal way of saying this is that r, the correlation coefficient, measures how well a batch of X and Y data values "fit" a linear function, which is "described" by a straight line on a graph.

B. Properties:

1. It is a "bounded" measure or index: its value lies between -1.0 and +1.0.
   i. For perfect positive correlation r = 1.0.
   ii. For perfect negative correlation r = - 1.0
2. More formally:

$$-1 \leq r \leq 1.0$$

3. Alternatively one can say that the closer r is to $|1.0|$, where the bars indicate **absolute value**, the closer the distribution is to a straight line.
4. A value of 0 indicates no linear correlation between X and Y.
5. Some pictures illustrate the point.
   i. In the case of perfect positive correlation the value of r is 1.0

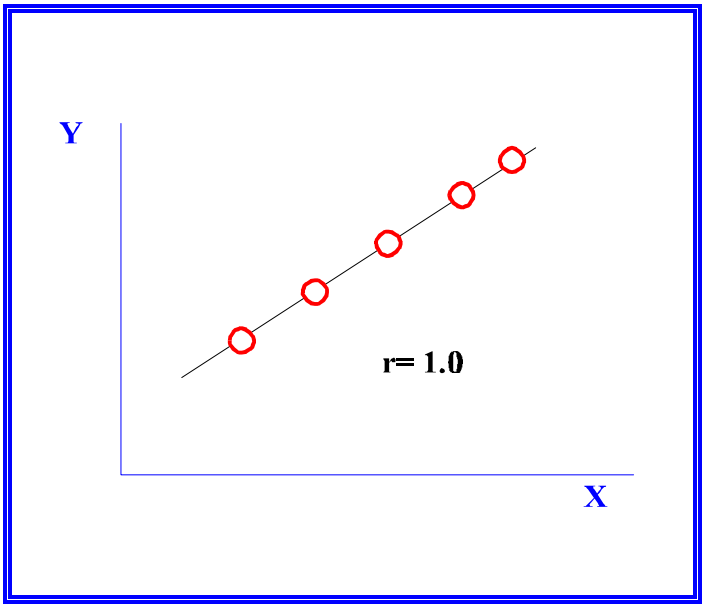

Figure 6: Correlation Coefficient Equals 1.0

ii. In the case of perfect negative correlation it is -1.0.

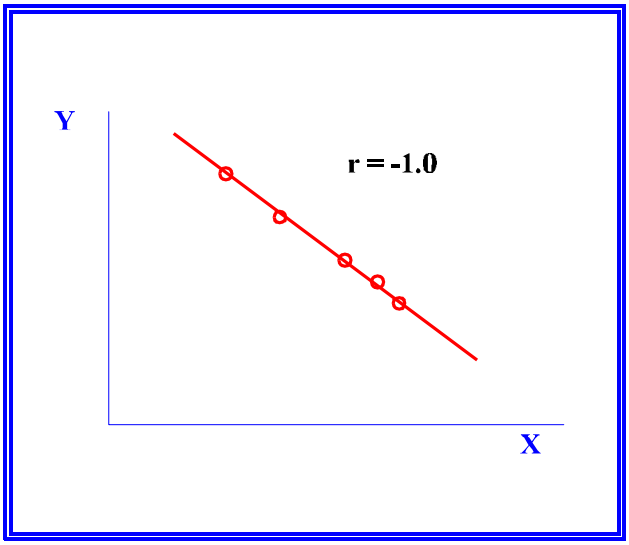

Figure 7: **All points lie on a (negative) straight line**

iii. When there is no **linear** correlation r is 0.0

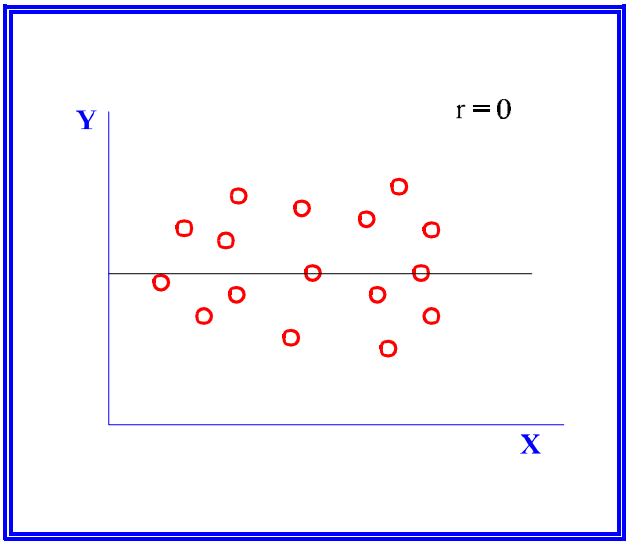

Figure 8: **Coefficient is r = 0**

iv. In this example, the data points are scattered randomly. The best fitting line has a slope of zero (see below).

VII. NEXT TIME:

A. More on measures of association and correlation

B. Reading:

1. Skim Johnson and Joslyn, ***Research Methods***, pages 336 to 353.
   i. But don't worry about the mathematics; try to understand what a measure of association tells you.
2. ***Research Methods***, pages 359 to 367 gives a good discussion of correlation analysis.