EXAMINING DATA

The following pages were designed to provide a rough guide as to how to examine data. These instructions are by no means exhaustive or complete and the statistics suggested do not take into consideration special situations and unusual data characteristics. After the database is complete and clean, you need to start the process of analysis. Progressing through these steps will improve the efficiency of your analysis and help prevent you from spending hours in front of the computer on "fishing trips." Use the process outlined below when you begin to think about writing the report for your project.

The questions contained in the next few pages were designed to be generic and useful for most projects. Our purpose was to develop a systematic process so that you can assemble in one place, the relevant information needed to write up your methodology and select the appropriate way to analyze the data.

Four phases are outlined:

1. The first step is to clarify the metadata associated with the information you are about to examine. This means that you need to know some basic things about the way the data were collected. Use this information to help write the "methodology" section of your report (of course you will need to draw upon additional information to provide details about the methodology when you are actually writing this section). However, this form will provide a good place to start. Some of this information may already be recorded in the codebook.

2. The next step is to identify which variables you are interested in examining. This is a critical step because the temptation is to look at the relationship between everything. Statistically this is considered "fishing" for good results.[1] Avoid fishing at all costs. It wastes time and leads to a dramatic increase in the sampling error.

3. The third step is to select the appropriate statistics for the variables you want to examine. This procedure is dependent on a number of characteristics of the data.

4. The final step is to create tables of your findings. These tables along with appropriate graphs or charts will then be inserted into your report along with your interpretations of the results.

[1] Fishing is also referred to as "data dredging."

STEP 1: DEVELOP METADATA FOR THE PROJECT

Prior to sitting down to examine a dataset, it is important to review the data collection methodology (sometimes we refer to this as metadata). Some details about the data collection should be found in the project files and also at the bottom of the codebook associated with your dataset. When writing up the results you will need to write a couple of paragraphs outlining the procedures that were followed to gather the information. The more specific you are on this step the better the final report will be. After sorting out the metadata move on to Step 2. Answer as completely as you can, each of the following:

A. Describe the sampling frame (population + method of selecting the sample).

B. Describe/ list the selection criteria.

C. When was the data collected (i.e. dates and time of day)?

D. Who collected the data (i.e. number and names of staff)?

E. How was the data collected (i.e. method of administration and collection)?

____

____

____

F. Non-response rate (i.e. if a survey was used, how many people decline to participate):

____

____

____

G. What are the main limitations (i.e. related to the measures and the data collection process) and how much does these factors affect the inferential strength of the results?

____

____

____

____

STEP 2: MAKING PREDICTIONS

Prior to sitting down to a database to examine the information, it is important to have a clear sense of what you want to look for. Your ideas should be based in a thorough examination of the relevant literature. This intensive literature review was done early in the project, probably when the grant proposal was written or before the data collection instruments were created. While you may not have been involved in this aspect of the project's development, the project director will be able to tell you what patterns are expected. Also, after spending many hours collecting and cleaning the data, you will also have your own ideas developed about which variables may be related. Isolating specific variables to examine is very important, especially when dealing with large datasets. If you do not narrow down your focus, you will spend hours, if not days, fishing for significant findings. Follow the steps below to select key variables to examine. You may elect to modify or repeat this process if you have more than three key variables or you are developing a complicated multivariate model. Once you have completed this process move on to Step 3.

A. Select one dependent variable: ____

B. Of all the independent variables included in this study which two do you think are most likely to be related to your dependent variable? (name the variables)

1. ___________________________ 2. ___________________________

C. Write a hypothesis for each predicted relationship.

1.
____
____
____
____

2.
____
____
____
____

D. Are there any problems with these measures that would impact on the external validity and reliability of these findings?

____
____
____
____
____
____

STEP 3: SELECTING THE APPROPRIATE STATISTIC

Selecting the appropriate statistic to examine your data (both univariate and bivariate analysis) can be very confusing; there are many statistics to choose from. Basically, the selection of the statistic depends on three factors:

- the level of measurement for the variables you are looking at,
- whether you expect that the data will have a certain underlying distribution (i.e. linear); and,
- the sampling frame (whether it was a random sample).

The first two aspects are the most important to the selection of appropriate statistics. To select the appropriate statistic for your analysis please fill in the table below identifying the level of measurement for each variable involved.

Concept	**Variable Name** (from codebook)	**Level of Measurement**
Independent 1. 2.		
Dependent		

UNIVARIATE ANALYSIS

Always begin your analysis by looking at the descriptive statistics for each variable. Generally, this information is reported when you discuss the sample or the way variable were constructed (before the results section). We refer to this process as doing the univariate analysis because you are examining each variable by itself. The selection of statistics used depends on the level of measurement for each variable. Remember there are four different levels of measurement:

DISCRETE VARIABLES

- nominal: response set has choices that have a name only (i.e. male and female are the two categories or choices for the variable gender)

- ordinal: response sets can be rank ordered (i.e. strongly agree, agree, neutral, disagree, strongly disagree for someone's opinion of a statement).

CONTINUOUS VARIABLES

- interval: numeric value of response set has meaning, and can undergo mathematical operations (i.e. age in years)
- ratio: numeric value of response set has meaning, can undergo mathematical operations, and there is a meaningful zero (i.e. usually a count of some kind like the number of times a person was arrested in the last 6 months).

The table below will help you to determine what statistics to look at for each type variable. Please note that certain graphic representations are appropriate for different kinds of variables.

Table 1. Type of Statistic and Graphic to use based on the Level of Measurement

Level of Measurement	**Central Tendency**	**Dispersion**	**Graphics**
Nominal	Valid percents for categories, with particular attention to the mode (modal category)	Index of Dispersion (also called Index of Qualitative Variation)	Pie chart, bar chart (bars do not touch).
Ordinal	Valid percents for categories, with particular attention to the mode.	Index of Dispersion (also called Index of Qualitative Variation)	Pie chart, bar chart (bars do not touch).
Interval	Mean and median; if you see a dramatic difference then the distribution is skewed – report both (no difference then report just the mean).	Standard Deviation Distribution (i.e. normal, j-curve, bi-modal)	Histogram (like a bar chart but the bars touch), or a line chart
Ratio	Mean and median- same above.	Standard Deviation Distribution (i.e. normal, j-curve, bi-modal)	Histogram (like a bar chart but the bars touch), or a line chart

Univariate analyses is generally done for three sets of variables: sample demographics or control variables, key independent variables, and dependent variables. The results are reported in a table and there is a paragraph or two for each category of variables. This text draws the readers' attention to the most predominant patterns in the table.

BIVARIATE ANALYSIS

Bivariate analysis requires that you examine the relationship between two variables. To describe a relationship you must:

- identify the form of the relationship (linear or curvilinear),
- assess the strength of the relationship (also referred to as the degree of association),
- determine the direction of the relationship (positive or negative), and
- establish the significance of the relationship (testing the hypothesis).

A. What kind of distribution do you expect to see? If the relationship between the variables is linear then you should select parametric statistics for this variable(s). If the relationship between the variables is non-linear then you will need to choose non-parametric statistics. To make this determination you may view a scatterplot to see how the variables are interacting.

B. Assessing the nature and strength of a relationship also depends on the level of measurement for each variable, the difference here is that you must take into account both variables. The chart below offers some guidance as to which statistics to select. The following pages provide much more detailed information about the different options available to you.

Table 2. Type of Statistic to Use with Two Levels of Measurement

Level of Measurement	Analytic Process	Measure of Association (PRE)	Significance Test	Direction/Shape
2 discrete variables nominal/nominal nominal/ordinal	Contingency Tables	Phi	Chi-square	Sign/ diagonal concentration
Discrete/Continuous	ANOVA	Eta (for linear) (square for PRE)	F Test	Sign
Continuous/ Continuous	Correlations	Pearson's Correlation Coefficient (square for PRE) LINEAR	T test	Sign/ scatter plot
Ordinal/ Ordinal	Correlations	Spearman's Rho (square for PRE) LINEAR	T test	Sign

Tables 3 to 6 that follow provide greater detail about many of the most popular statistics used in research reports.

Table 3. Measures of Association for Nonparametric Statistics -- Nominal

Level of Measurement	Measure of Association	Formula	Use/ Assumptions	Range	Direction
2 nominal variables or nominal w/ ordinal	**Phi (f)**	$f = \sqrt{\frac{c^2}{N}}$	2x2 table; independent SRS to gather data; observations independent (no times series data); no cell has less than 5 freq; f^2 demonstrates the amount of influence of IV on DV, take % for PRE.	-1 to +1	Always + when used w/ c^2
Nominal w/ ordinal	Contingency Coefficient (C)	$C = \sqrt{\frac{c^2}{c^2 + N}}$	> 2 categories; strength depends on upper limit; cannot be squared to get an estimate of variance	0 to $\sqrt{\frac{(r-1)}{r}}$ r = # of rows	Always + due to square root
Nominal w/ ordinal	Cramer's V (V)	$V = \sqrt{\frac{c^2}{N(L-1)}}$ L=lesser of (r-1) or (c-1)	> 2 categories; cannot be squared to get an estimate of variance	-1 to +1	+ variables move together, - inverse
2 nominal variables	Lambda (?)	$l = \frac{E_1 - E_2}{E_1}$	Specify IV and DV, is an asymmetric measure; If IV and DV have same modal category it can't be used; PRE measure when take %.	0 to 1	Always +

Table 4. Measures of Association for Nonparametric Statistics – 2 Ordinal Variables

Measure of Association	Formula	Use/ Assumptions	Range	Direction
Gamma (?)	$g = \frac{N_s - N_d}{N_s + N_d}$	General use	-1 to +1	+ more pairs are similar - more dissimilar
Somer's d (*d*)	$d = \frac{N_s - N_d}{N_s + N_d + T_y}$	Tied on DV Takes tied pairs into account, weakens assoc.	-1 to +1	+ more similar pairs - more dissimilar
Tau-b	$tau-b = \frac{N_s - N_d}{\sqrt{(Ns + Nd + Ty)(Ns + Nd + Tx)}}$	Tied on x but not on y Includes pairs tied on y but not x. Must have equal # r and c	-1 to +1 (1 if all *f* fall on the diagonal)	+ more similar pairs - more dissimilar
Spearman's rank order cor. (rho) (r_s)	$r_s = 1 - \frac{6(\sum D^2)}{N(N^2 - 1)}$	Variables ranked on a case-by-case basis r_s^2 (rho)^2 is a PRE measure. Error is reduced by that amount when we know IV. **Most popular.**	-1 to +1	+, both variables move together, - inverse

Strength level for all but Contingency Coefficient (C) is: .00 - .10 no relationship; 20-.40 = weak to moderate; .40-.65 = strong; .65-.09 = very strong; 1.0 perfect

Table 5. Hypothesis Testing for Discrete Data (Nonparametric Statistics). All Alpha levels follow the .05, .01, or .001 convention.

Test	Formula	Use	Measure of Association	Notes
Chi-square (x^2) contingency table	$x^2 = \sum \frac{(O-E)^2}{E}$ $E = \frac{(ColumnN)(RowN)}{TotalN}$ (O = observed frequency; E = expected frequency in each cell.) DF = (r-1) (c-1)	-Nominal or Ordinal -Independent samples -Expected cell *f* ≥ 5	**Phi (f) (use w/ 2x2 table)** Contingency Coefficient (C) (use w/ >2 categories) Cramer's V (V) (use w/ > 2 categories) Gamma (?) (use w/ ordinal) Lambda (?) (use w/ asymmetric data) **Somer's d (*d*) (use with ordinal that are tied on DV)** **Tau-b (use with ordinal that are tied on x but not on y)** Yule's Q (*Q*) (use in place of Gamma for 2x2 tables) **Spearman's rank order correlation (rho) (r_s) (used for variables ranked by case)**	f^2 is used for variance once you convert to percentage, represents the amount of variation in the DV accounted for by the IV. PRE measures are used to determine how much more accurate the estimate of the DV is with IV information than without that info.
Yates' Correction ($x^2_{corrected}$) for continuity	$x^2_{corrected} = \sum \frac{(\lvert O-E \rvert - 0.5)^2}{E}$ $E = \frac{(ColumnN)(RowN)}{TotalN}$ DF = (r-1)(c-1)	-Nominal or Ordinal -Independent sample -Cell *f* < 5	Same	Same
Kruskal-Wallis test (H)	$H = \frac{12}{N(N+1)}\left[\frac{s_1^{\ 2}}{N_1} + \frac{s_2^{\ 2}}{N_2} + \frac{s_3^{\ 2}}{N_3}\right] - 3(N+1)$ DF = *k*-1	Ordinal data when ranked	Eta $h^2 = \frac{H}{N-1}$	Can use eta here since ranked data are treated as continuous data

Table 6. Testing Hypotheses for Continuous Data (Parametric Statistics)

Test	Formula	Use	Alpha level	Rule	df	Measure of Association	Notes
Z score	$Z = \frac{x - \overline{X}}{s}$	Determine difference between a score and mean	N/A	N/A	N/A	N/A	Look up either Z score or probability between score and mean in "Areas of Normal Curve" chart
Decision Rule (z)	$z = \frac{\overline{X} - m}{s / \sqrt{N}}$	Determine difference for 2 groups of scores	.05, .01, or .001	Retain if $< t_{crit}$ Reject if $\geq t_{crit}$	N/A	N/A	
Student's t (t-test) – independent means	$t = \frac{\overline{X}_1 - \overline{X}_2}{\left(\sqrt{\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2}}\right)\sqrt{\frac{N_1 + N_2}{(N_1)(N_2)}}}$	Total N is less than 120 and means are independent	.05, .01, or .001	Retain if $< t_{crit}$ Reject if $\geq t_{crit}$	N - 2; when 2 is the # of groups	Eta $h^2 = \frac{t^2}{t^2 + df}$	Must decide if one or two tailed Use % of h^2 to get the variance explained.
Correlated or dependent **means (t-test)**	$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{N}}{N(N-1)}}}$	Total N is less than 120 and means are dependent	.05, .01, or .001	Retain if $< t_{crit}$ Reject if $\geq t_{crit}$	N-1	Eta $h^2 = \frac{t^2}{t^2 + df}$	Must decide if one or two tailed Use % of h^2 to get the variance explained.
ANOVA- Analysis of Variance	$F = \frac{MS_{BETWEEN}}{MS_{WITHIN}}$ * * Must use SS formulas to get $MS_{BETWEEN}$ and MS_{WITHIN} when doing calculations by hand.	Comparison of grouped data using variance	.05, .01, or .001	Retain if $< F_{crit}$ Reject if $\geq F_{crit}$	$df_b = k-1$ $df_w = N-k$	Eta $h^2 = \frac{SS_{BETWEEN}}{SS_{TOTAL}}$	Use Scheffé Test to determine which means differ

MULTIVARIATE ANALYSIS

Multivariate analysis involves examining a model with many different independent and control variables with one dependent variable. The three most popular forms of multivariate analysis are:

- multiple OLS regression: used when you have a continuous dependent variable and continuous independent variables (or dichotomous independent variables and at least one continuous independent variable)
- logistic regression: used when you have a discrete dependent variable and continuous independent variables.
- path analysis: used when you have mediating variables or expect that the independent variables work through another factor and may have indirect effects on the dependent variable.

See your project director for directions about how to do this for your project.

Very helpful resources:

Morgan, S.E., T. Reichert, and T.R. Harrison, 2002. *From Numbers to Words: Reporting Statistical Results for the Social Sciences*. Allyn and Bacon.

Walsh, A. & J.C. Ollenburger, 2001. *Essential Statistics for the Social and Behavioral Sciences*. Prentice-Hall.

Vogt, W. P., 1999. *Dictionary of Statistics and Methodology: a Nontechnical Guide for the Social Sciences*. 2nd edition. Sage Publications.

STEP 4: CREATING TABLES AND/OR GRAPHICS

All tables and graphics must have the following elements:

1. **Informative Title:** describes what appears in the image or what variables/analysis are presented in the table.
2. **Notes:** located below the table or image to provide additional information needed to interpret the data or to identify its source:
   a. data source(s)
   b. time period (when data was collected and the time period of the data)
   c. sample size or data coverage

d. significance levels or notes about the statistics presented (tables only)

3. **Legend:** if needed, to explain symbolism or truncated variable names

4. **Axis Labels and value labels** (when appropriate if creating a chart)

A number of different tables are presented in the following pages. These are by no means the most perfect tables but they will provide some guidance as to some ideas of formatting for different kinds of analyses. Note that the decimal points must always line up. Also, if you are in doubt, pick-up a few Criminology journals and model your tables after their formats.

Table 7. Sample Characteristics for theVictimization Study, 2000.

Variable	Frequency	Percent*
Ethnicity		
Hispanic	400	40
European	600	60
Gender		
Male	863	86.3
Age Groups		
18-29	620	62
30-49	350	35
50+	30	3
Mode of Conviction		
Guilty Pleas	915	91.5
Bench Trials	56	5.6
Jury Trials	29	2.9
In/Out		
Not Incarcerated	363	36.3
Incarcerated	637	63.7

* Valid percents reported.

Table 8. Comparing the DUI Survey respondents with the Undergraduate Campus Population for the 2000 – 2001 Academic Year.

Demographic Variables	Survey Respondents[a]	Undergraduate Population (Fall 2000)[b]
Status	(440)	
Freshman	16.4 %	22.8 %
Sophomore	21.4	13.6
Junior	32.5	28.3
Senior	29.8	35.3
Average Age	22.1 (443)	25.4 years

Female	54.1% (451)	62.9 %
Ethnicity	(444)	
Hispanic	30.0	28.9 %
African American	8.6	10.7
White European	41.4	39.2
Asian	6.3	8.1
Mixed	13.7	13.1
Majors	(441)	
Arts and Letters	23.4	28.6 %
Business	19.0	20.9
Education	2.7	0.3
Natural Sciences	17.0	18.6
Social Sciences & Behavioral	32.9	22.8
Undecided	5.0	8.9
Average Credit Load	14.39 units (441)	12.99 units
Working	(447)	
Not working	20.0	Not available
Up to 20 hrs/week	27.3	
21- 40 hrs/week	40.7	
40+ hrs/week	13.0	
Living Arrangements	(447)	
Live with Parents	42.5	Not available
Roommates	32.7	
Significant Other	9.6	
Alone	10.1	
Combination	5.1	

Notes:

[a] While 470 students responded to the survey, 18 were removed because they were graduate students. This brings the total sample size to 452. There was no attempt to capture a representative proportion of graduate students. Valid percents used for survey respondents.

[b] Campus figures are for undergraduate students only and do not include extended education or certificate program students. These figures are generated by the Office of Institutional Research and reported in the Annual Statistical Factbook (2001).

Table 9. Mean, Standard Deviation and Range for Self Control, 1996 (N=100).

Variable	Mean	S.D.	Minimum	Maximum
Estimated for Both Scenarios				
CB Ratio (Sub to/Exercised)	.50	.27	.12	1.32
Sex (Female = 1)	.49	.50	0.00	1.00
Religiosity	.34	.39	0.00	1.00
Low Self-Control	37.12	10.45	16.00	67.00
Age	22.83	4.24	18.00	46.00
Predation Specific				
Intentions to Commit Predation	.42	.24	0.00	1.00
Prior Predation	.12	.36	0.00	1.00
Moral Beliefs –Predation	6.19	2.79	0.00	10.00
Peers Commit Predation	6.01	3.65	0.00	10.00
Perceived Risk – Predation	6.07	2.95	0.00	10.00
Exciting to Commit Predation	3.10	3.69	0.00	10.00
Defiance Specific				
Intentions to Commit Defiance	.51	.50	0.00	1.00
Prior Unwanted Sex	.18	.52	0.00	3.00
Moral Beliefs – Defiance	5.79	3.45	0.00	10.00
Peers Commit Defiance	3.33	3.19	0.00	10.00
Perceived Risk – Defiance	2.59	2.71	0.00	10.00
Exciting to Commit Defiance	4.85	3.92	0.00	10.00

Table 10. OLS Regression Analysis of Homicide Rates (1999-2001).

	Family Homicide			Acquaintance Homicide			Stranger Homicide		
Variable	*b*	$	*t* Ratio	*b*	$	*t* Ratio	*b*	$	*t* Ratio
Gini Index	3.56*	.59	2.44	2.38	.18	1.67	4.21*	.21	2.22
Percent Black	.03	.14	1.90	.03*	.15	2.2	.00	-.02	-.34
Unemployment	.04	.14	2.09	.07*	.22	2.78	.05	.14	1.61
Divorce Rate	.03	.32	3.33	.03*	.56	7.24	.02*	.98	4.31
South	.22	.10	1.39	.22	.09	1.98	-.04	-.02	-.36
Inverse Population	.00	.19	1.90	.00	.33	1.35	.00	.00	-.07
Percent Young	-.50*	-.16	-3.00	.00	.00	-.01	-.09	-.66	-.56
Poverty	-.60*	-.16	-3.09	-.70	-.67	-3.48	-.10*	-.30	-4.12
Log Density	-.22	-.19	-2.44	.00	.00	-.01	.19	-.03	.96
Pop. Change 80-92	.00	.09	.80	.00	.02	.09	.06	.15	1.57
City Share	-.52*	-1.62	-2.28	-.30	-.09	-1.25	-.46	-.80	-1.73
Constant	1.05			-.49			-2.09		
Adjust R^2	.49			.67			.40		

* Significant at $p < .05$.

Table 11. t-Test Results for Difference of Means in Factors of Social Control (N=604).

	Intent to shoplift			Intent to drive drunk		
	M			M		
Variable	Men (n=280)	Women (n=324)	t ratio	Men (n=280)	Women (n=324)	t ratio
Low Control	78.90	68.35	-4.56*	64.58	68.33	-4.48*
Shame	7.95	7.25	3.40**	8.12	7.52	2.66**
Priors	57.04	34.23	5.64*	48.99	34.39	5.04*
Pleasure	1.60	2.30	-3.26**	1.62	2.43	-3.89**
Sanctions	4,705.30	5,036.03	4.93*	3,805.82	4,078.50	4.90*
Morals	0.78	0.84	-1.05	2.02	2.02	-3.51**
Intent to shoplift	2.43	2.34	-3.89**			
Intent to drive drunk				2.00	2.56	-4.80**

*p<.001; *p<.01

Table 12. Frequencies, Coding, and Factor and Reliability Analysis of Dependent Variables (N = 100).

Dependent Variables	Percent	
	Shame at Wave 1	Shame at Wave 3
I certainly feel useless at times.		
4 = Strongly Agree	7.8 %	3.7%
3 = Agree	38.3%	30.5%
2 = Disagree	32.5%	31.9%
1 = Strongly Disagree	12.3%	11.6%
Missing	9.1%	22.3%
Factor Loading	**.790**	**.821**
My plans hardly ever work out, so planning really makes me unhappy.		
4 = Strongly Agree	4.8%	2.9%
3 = Agree	13.0%	11.8%
2 = Disagree	48.3%	45.9%
1 = Strongly Disagree	25.3%	16.9%
Missing	8.5%	22.6%
Factor Loading	**.661**	**.721**
Scale Alpha	**.772**	**.785**

Table 13. Logistic Regression Coefficients Predicting Early Onset by Level of SES.

	Low SES				High SES			
Variable	**B**	SE(*B*)	Wald	Exp(*B*)	**B**	SE(*B*)	Wald	Exp(*B*)
Low Birth Weight	2.56	.65	5.53*	5.13	.58	.38	2.96	1.20
SES	-.11	.07	2.30	.90	-.03	.01	5.55*	.97
Weak Family Structure	.23	.13	2.10	1.01	.02	.07	.04	1.02
Gender	.64	.33	.97	1.90	-.56	.41	1.76	.11
Constant	-.39	1.09	5.88*	1.14	.11	.55	.08	
X^2		11.26				10.19		
df		5				5		
p		.02				.08		
Model Prediction Rate		68%				63%		

Table 14. Bivariate Correlations Matrix for Independent Variables and Composite Dependent Variables (N = 428).

Variable	Age	Female	White	Education	Protestant	No Religion	Conservative
Age	---						
Female	-.04	---					
White	.10*	-.03	---				
Education	-.08	-.06	.22**	---			
Protestant	.16**	.00	-.31**	-.17	---		
No Religion	-.20**	.01	.05	-.03	-.42**	---	
Conservative	.30**	.13**	-.03	-.13**	.17**	-.11*	---

**p* <.05 (two-tailed); ***p* <.01 (two-tailed)