Data Analysis Plan

The appropriate methods of data analysis are determined by your data types and variables of interest, the actual distribution of the variables, and the number of cases. Different analyses of the same dataset may reflect or represent different aspects of the underlying data structure.

Once a plan has been established, it could contain any combination of the following types of data **analysis strategies**:

- **Exploratory**: This type of data analysis often occurs when a program is new, and it is unclear what to expect from the data.
- **Descriptive**: The most common type of data analysis, this approach will summarize your findings and describe the sample.
- **Inferential**: Inferential statistics allow us to draw conclusions about the larger population from which the sample is drawn. These powerful techniques may be able demonstrate if a change has occurred as a result of your program. PDA's Statistics Division specializes in many sophisticated data analysis techniques.

4.1 Exploratory Analysis:
Once the data is collected and entered, the first question is: "What do the data look like?"

Exploratory data analysis uses numerical and graphical methods to display important features of your data set. Exploratory data analysis helps us to highlight general features of your data to direct future analyses. It also pinpoints problem areas in the data. For example:

- Should outliers be included or excluded in the analyses?
- Do the data need cleaning for consistency?
- How much missing data is there and how should it be handled?
- What do the distributions look like for key variables?

Distribution of the data: What's the "shape" of the data? Where do most of the values lie? Are they clumped around a central value, and if so, are there roughly as many above this value as below it? We look at the distribution for each variable to determine which analyses would be most appropriate. Sometimes it is necessary to examine distributions of data partitioned by other key variables.

Missing Values: In a survey, missing values correspond to skipped questions or unendorsed options. A discussion between analyst and client should take place in determining how missing values should be handled. In some cases, missing values might be perfectly normal (e.g. the variable "result of pregnancy test" for a male is blank). However, in some cases missing values for important variables might exclude a record from certain analyses. Sometimes it is appropriate to place normalized values in place of missing values.

Outliers: "Unusually" large or small values that are dramatically separated from the rest of the data might be: 'out-of-range' or physically impossible values that resulted from entry or processing error. Merely "weird" values might represent entry error,

random fluctuation, or members of a population other than the one you want to study.

Data Cleaning: Exploratory data analyses might indicate a particular need to clean data. Data cleaning is extremely important when the data collection method allows inconsistencies. All data cleaning work must be carefully documented and available in a report. Date cleaning includes the following activities as needed:

- **Removal of outliers**: Invalid, impossible, or extreme values may be removed from the dataset. Outliers might also be marked for exclusion for the purpose of certain analyses.
- **Labeling missing values**: It may be necessary to label each missing value with the reason it is considered missing in order to guarantee accurate bases for analysis.

4.2 Descriptive Data Analysis:
Descriptive statistics tell you how your data look, and what the relationships are between the different variables in your data set. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Descriptive Statistics are used to present quantitative descriptions in a manageable form. Each descriptive statistic reduces lots of data into a simpler summary. For instance, consider a simple number used to summarize how well a batter is performing in baseball, the batting average. This single number is simply the number of hits divided by the number of times at bat (reported to three significant digits). A batter who is hitting .333 is getting a hit one time in every three at bats. One batting .250 is hitting one time in four. The single number describes a large number of discrete events.

Every time you try to describe a large set of observations with a single indicator you run the risk of distorting the original data or losing important detail. The batting average doesn't tell you whether the batter is hitting home runs or singles. It doesn't tell whether she's been in a slump or on a streak. Even given these limitations, descriptive statistics provide a powerful summary that may enable comparisons across people or other units.

Univariate Analysis. Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the central tendency
- the dispersion

In most situations, we would describe all three of these characteristics for each of the variables in our study.

The Distribution. The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value. For instance, a typical way to describe the distribution of college students is by year in college, listing the number or percent of students at each of the four years. Or, we describe gender by

listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. With variables that can have a large number of possible values, with relatively few people having each one, we group the raw scores into categories according to ranges of values.

Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

Table 1. Frequency distribution table.

One of the most common ways to describe a single variable is with a ***frequency distribution***. Depending on the particular variable, all of the data values may be represented, or you may group the values into categories first (e.g., with age, price, or temperature variables, it would usually not be sensible to determine the frequencies for each value. Rather, the values are grouped into ranges and the frequencies determined.). Frequency distributions can be depicted in two ways, as a table or as a graph. Table 1 shows an age frequency distribution with five categories of age ranges defined. The same frequency distribution can be depicted in a graph as shown in Figure 2. This type of graph is often referred to as a *histogram* or *bar chart*.

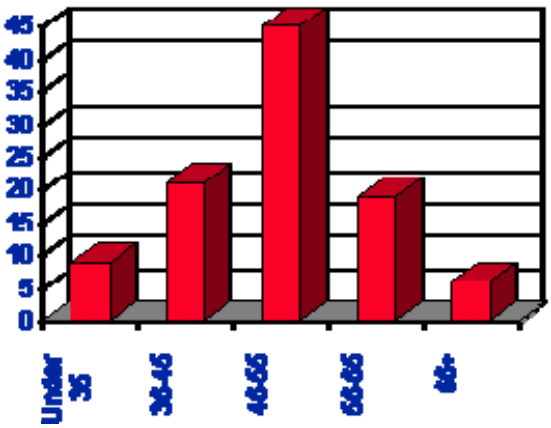

Table 2. Frequency distribution bar chart.

Distributions may also be displayed using percentages. For example, you could use percentages to describe the:

- percentage of people in different income levels
- percentage of people in different age ranges
- percentage of people in different ranges of standardized test scores

Central Tendency. The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The **Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values. For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167, so the mean is 167/8 = 20.875.
The **Median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score #250 would be the median. If we order the 8 scores shown above, we would get:

15,15,15,20,20,21,25,36

There are 8 scores and score #4 and #5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

The **mode** is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

Notice that for the same set of 8 scores we got three different values -- 20.875, 20, and 15 -- for the mean, median and mode respectively. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other.

Dispersion. Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 - 15 = 21.

The **Standard Deviation** is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The Standard Deviation shows the relation that set of scores has to the mean of the sample. Again lets take the set of scores:

15,20,21,20,36,15,25,15

To compute the standard deviation, we first find the distance between each value and the mean. We know from above that the mean is 20.875. So, the differences from the mean are:
15 - 20.875 = -5.875
20 - 20.875 = -0.875
21 - 20.875 = +0.125
20 - 20.875 = -0.875
36 - 20.875 = 15.125
15 - 20.875 = -5.875
25 - 20.875 = +4.125
15 - 20.875 = -5.875

Notice that values that are below the mean have negative discrepancies and values above it have positive ones. Next, we square each discrepancy:
-5.875 * -5.875 = 34.515625
-0.875 * -0.875 = 0.765625
+0.125 * +0.125 = 0.015625
-0.875 * -0.875 = 0.765625
15.125 * 15.125 = 228.765625
-5.875 * -5.875 = 34.515625
+4.125 * +4.125 = 17.015625
-5.875 * -5.875 = 34.515625

Now, we take these "squares" and sum them to get the Sum of Squares (SS) value. Here, the sum is 350.875. Next, we divide this sum by the number of scores minus 1. Here, the result is 350.875 / 7 = 50.125. This value is known as the **variance**. To get the standard deviation, we take the square root of the variance (remember that we squared the deviations earlier). This would be SQRT(50.125) = 7.079901129253. Although this computation may seem convoluted, it's actually quite simple. To see this, consider the formula for the standard deviation:

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

In the top part of the ratio, the numerator, we see that each score has the the mean subtracted from it, the difference is squared, and the squares are summed. In the bottom part, we take the number of scores minus 1. The ratio is the variance and the square root is the standard deviation. In English, we can describe the standard deviation as: **the square root of the sum of the squared deviations from the mean divided by the number of scores minus one.**

Although we can calculate these univariate statistics by hand, it gets quite tedious when you have more than a few values and variables. Every statistics program is capable of calculating them easily for you. For instance, I put the eight scores into a

statistical software program (eg, Epi Info, Excel, SPPS, Stata, SAS) and got the following table as a result, which confirms the calculations done by hand:

N	8
Mean	20.8750
Median	20.0000
Mode	15.00
Std. Deviation	7.0799
Variance	50.1250
Range	21.00

The standard deviation allows us to reach some conclusions about specific scores in our distribution. Assuming that the distribution of scores is normal or bell-shaped (or close to it!), the following conclusions can be reached:

- approximately 68% of the scores in the sample fall within one standard deviation of the mean
- approximately 95% of the scores in the sample fall within two standard deviations of the mean
- approximately 99% of the scores in the sample fall within three standard deviations of the mean

For instance, since the mean in our example is 20.875 and the standard deviation is 7.0799, we can from the above statement estimate that approximately 95% of the scores will fall in the range of 20.875-(2*7.0799) to 20.875+(2*7.0799) or between 6.7152 and 35.0348. This kind of information is a critical stepping stone to enabling us to compare the performance of an individual on one variable with their performance on another, even when the variables are measured on entirely different scales.

4.3 Inferential Statistics:

Inferential statistics test hypotheses about the data and may permit you to generalize beyond your dataset. Examples include comparing means (averages) for a given measurement between several different groups or for the same individuals across time:

- **Trends Analysis**: If your study involves repeating the same measurements at different points in time, repeated measures or time series analyses can help find changes or patterns over time.
- **Analysis Of Variance Models**: Analysis of variance is used to compare average scores for different groups. Many different ANOVA models are available to tease out the effect of covariates (factors expected to relate to the outcome, e.g., stressful life events and depression) and handle multiple dependent variables (outcomes).
- **Multiple Regression & Correlation**: Correlation measures the strength of the relationship between different variables in your data (e.g., there is a strong relationship between height and shoe size, a weak relationship between height and IQ). Multiple regression examines how well one set of variables (e.g. hours studied, IQ, interest in topic) predicts an outcome variable (grade on an exam) and specifies the unique contributions of each predictor (e.g.,

hours studied is more important than IQ in predicting grade). Canonical correlation analysis relates one set of variables to another set.

References

AO Foundation (n.d.). Step-by-step guide to DOING clinical research. Retrieved 09 October 2006 from http://www.aofoundation.org/portal/wps/portal/!ut/p/.cmd/cs/.ce/7_0_A/.s/7_0_7T5/_s.7_0_A/7_0_7T5

Professional Data Analysts (n.d.). Stage 3: Data Analysis.Retrieved 09 October 2006 from http://www.pdastats.com/default.asp

Trochim, W.M.K. (2006).*Descriptive Statistics.* Retrieved 09 October 2006 from *http://www.socialresearchmethods.net/kb/statdesc.htm*

Wikipedia (n.d.). Statistical Bias. Retrieved on 09 October 2006 from http://en.wikipedia.org/wiki/Bias_(statistics)#External_link