\*Steps in Data Analysis

Statistics is process of:

Collecting Data

Summarizing Data

Interpreting Data

Steps in collection of data:

1. Choosing a sample - subgroup of population.

2. Exploratory data analysis - summarization of collected data the meaningful way. Exploratory analysis help scientist to refine questions and sometimes even to create entirely new questions.

3. Inferential data analysis.

\*What is Data?

What do we really mean by data? Put simply, data are pieces of information about individuals organized into variables. By individual, we mean a unit of observation.

An observation or unit of observation refers to a particular person or a particular object, any particular unit of observation within your study sample.

By a variable we need a particular characteristic of the unit of observation.

A data set is literally a set of data that's made up of individual observations and variables. Data sets are typically displayed in tables in which rows represent individuals, or units of observation, and columns represent variables.

Variables can also be classified into one of two types, Quantitative or Categorical.

1. Quantitative variables take numerical values and represent some kind of measurement.

2. Categorical variables, on the other hand, take category or label values and place an observation or individual into one of several groups. In case if coded as numerical codes (0,1) and often referred to as Dummy Codes because they have no arithmetic meaning. That is, it doesn't make sense to add them, subtract them, multiply or divide them. Or even compare the magnitude of these values.

3. Unique identifier - is a variable that is meant to distinctively define each of the units of observation of your data set.

\*Datasets and code books

\*Developing research question

\*Defining exploratory data analysis

Exploratory data analysis is what you use to make sense of the data. You do this by converting data from this raw form to a form that makes sense, that has context, that tells the story you want to tell. Basically, exploratory data analysis consists of:

1. organizing and summarizing raw data

2. looking for important features and patterns in the data

3. looking for any striking deviations from those patterns

4. interpreting your findings in the context of the problem or research question.

1. Univariate/Descriptive analysis - exploratory data analysisis looking at one variable at a time.

Summarize and examine the distribution of any variables.

Distribution of a variable is what valies the variable takes and how often the variables takes those values.

PROC FREQ; TABLES var1 var 2 var3;

RUN;

PROC FREQ produces:

Frequency, Percent, Cumulative Frequency, Cumulative Percent.

DATA steps - to manage and manipulate data

PROC procedures - to analyze and present your data

2. Data management - involves making decisions about data

\*Setting aside missing data

Do you need or not missing data?

Override "unknown" response or data as MISSING, to exclude from analyasis.

\*Coding in valid data and recording values

\*Creating secondary variables

One of the later steps in data management is evaluating whether you might want to create secondary variables.

Secondary variables are variables that include information from two or more primary variables. We can create secondary variables by using a mathematical or logical operation on two or more variables.

\*Grouping variables with individual variables

Create categorical variable that devides the sample into roughly three, or four or ... equal-sized groups.

\*Visualizing Data

Bar charts are most commonly used to examine the distribution of individual variables.

A bar chart helps us display the distribution of a categorical variable, for example a percentage of observations in each category.

Categorical variables can be visualized one at a time with univariate graphs, that is with single variable bar charts.

For Categorical variables:

PROC GCHART; VBAR categorical variable/Discrete type=PCT width=30;

Discrete is an alternate term for categorical.

Type=PCT is telling SAS to generate a graph using percentages, rather than raw numbers. That is, the percentage of observations in each level of the named categorical variable.

For Quantitative variables:

PROC GCHART; VBAR Quat\_Var/ type=PCT;

\*Describing distributions visually

When describing the shape of a distribution, we should consider symmetry or skewness of a distribution and peakness or modality. That is the number of peaks or modes that the distribution has.

Unimodal - when one bell (one mode value around which distribution concetrated)

Bimodal - when roughly 2 bells (2 modes/values around which distributions concetrated)

Uniform - Basically flat, no bells present (no modes)

Skewed-right - if the right tail, the larger values, is much longer than the left tail, or smaller values. An example of a real-life variable that has a skewed-right distribution is salary. Most people earn in the low to medium range of salaries with a few exceptions such as CEO's professional athletes etc. That are distributed along a large range that is the long tail of higher values.

Skewed-left - if the left tail, or smaller values, is much longer than the right tail, or larger values. Note that in the skewed-left distribution, the bulk of the observations are medium to large with a few observations that are much smaller than the rest. Most deaths from natural causes happen at older ages with fewer cases happening in younger ages.

Spread - The spread of the distribution, also called variability, can be described by the approximate range covered by the data.

\*Measures of Center and Spread

mean=sum of values / number of observations

mode=value with highest frequency

median=value of the middle item

Standard Deviation - The standard deviation gives the average or typical distance between a data point and the mean. The idea behind the standard deviation is to quantify the spread of the distribution by measuring how far the observations are from their mean.

To calculate Standard Deviation in SAS we use PROC UNIVARIATE:

PROC UNIVARIATE; VAR quat\_var; RUN;

To describe categorical variables, use:

PROC FREQ and GCHART procedures

To describe Quantitative variables, use:

PROC UNIVARIATE

\*Designing the role each of variables

To answer questions if our two variables have relationship.

Examples:

If we want to explore whether the outcome of the study, the test score, is affected by the test-takers gender, we would designate gender as the explanatory variable and test score as the response variable.

If we want to see whether a person's pass/fail outcome on a driving test can be explained by the length of time that they practice driving, prior to the test, time would be the explanatory variable and driving test the response variable.

When we graph the association between two variables, the independent, or explanatory variable, is plotted on the X axis. The dependent, or response variable, is plotted on the Y axis.

\*Graphing Categorical response variable

To graph the relationship between a categorical explanatory variable and a categorical response variable, we use PROC GCHART:

PROC GCHART; VBAR categoricalexplanatoryvariable/discrete TYPE=mean SUMVAR=categoricalresponsevariable;

RUN;

PROC GCHART; VBAR PACKCATEGORY/discrete TYPE=mean SUMVAR=TAB12MDX;

RUN;

categoricalexplanatoryvariable/discrete - X axis

SUMVAR=categoricalresponsevariable - Y axis

\*Graphing Quantitative response variable

A Scatterplot, by definition, is a graph of plotted points that show the relationship between two quantitative variables.

In a scatterplot data for each observation's explanatory and response variable are plotted.

SAS provides scatterplots in response to the PROC GPLOT command:

PROC GPLOT; PLOT quant\_var1\*Quany\_var2;

RUN;

To characterize the relationship that we see in this scatterplot, it can be helpful to draw a line of best fit through the observations as a way of trying to determine how the dots line up. That is, do they seem to line up in a positive or negative direction? Or with a positive or negative slope?