* Questions that are too general will require us to consider too many variables.

• Questions that don’t address the purpose of the study will cause us to choose the wrong variables.

• We always determine each variable’s data type. (categorical or quantitative)

• Units should be specified when working with quantitative variables

1. WHY was the data collected? (What was the intended purpose of the study?)

2. WHAT data was collected? (What do the data values represent? If the values should have units, what are they?)

3. WHO was the data collected from? (This refers to the individuals from which the data was actually obtained, also called the sample)

4. HOW was the data collected?

5. WHEN was the data collected?

6. WHERE was the data collected?

The first type is called quantitative, measures the amount of some quantity.

Three common quantitative variables we encounter are:

• Physical measurements: like time, weight, height

• Count data: number of students in a class, number of years in school

• Percentages: percentage of points on an exam, percentile of an SAT score The values of the second data type a variable can have, called categorical, represent categories or groups.

For the data type of a variable having numerical values to be quantitative, its values must have units. Another test sometime used is that a variable has a quantitative data type if we can do meaningful math with the values (ex, averages and differences have meaning).

Three common examples of categorical variables we will encounter are variables that

• Sort individuals into categories or classes: like gender, age group, race

• Identifiers: Area codes, social security numbers

• Rakings: 1 to 5 scale (so called Likert scale), pain scale, educational level

Variables, parameter,

* The parameter for a variable is the one number we would compute if we had the data for a variable from everyone in the population.

• For a categorical variable, the parameter is the population percentage or proportion in one of the categories.

• For a quantitative variable, the parameter is the average of all the variable values.

Sample Statistics estimate to the population *parameters*.

"The greatest value of a picture is when it forces us to notice what we never expected to see.”

John Tukey

Four Things We Always Do When We Summarize a Data Set

1) Look at the data values

2) Visually summarize (Make a plot and/or table of the data)

3) Numerically summarize (Compute numbers that summarize main features of the plot)

4) Describe the plot (Using the visual and numerical summaries)

What is the relationship between invalid data and outliers?