# CHAPTER TWO

# DESCRIBING AND USING DATA

Data is the foundation from which all scientific inferences are made. The observations we make on our sample patients allow us to answer hypotheses about the population-at-large. The collection of data is one of the more tedious aspects of medical research and biostatistics. It is also one of the most crucial aspects if meaningful and accurate conclusions are to be drawn from the data. No amount of statistical manipulation can overcome data which is either not collected to begin with or is collected improperly. It is therefore essential to consider the types of data that will be necessary to answer the proposed research hypotheses before data collection begins to ensure that the correct types of data will be available for analysis later.

## TYPES OF VARIABLES

There are many different types of data each of which requires specific statistical tests for correct analysis. The characteristics of interest in any study are called **variables**. Study variables may include age, sex, height, weight, blood pressure, cardiac output, survival, or any of a multitude of other characteristics. There are two major types of variables: **discrete** (or **categorical**) and **continuous** (Figure 2-1).

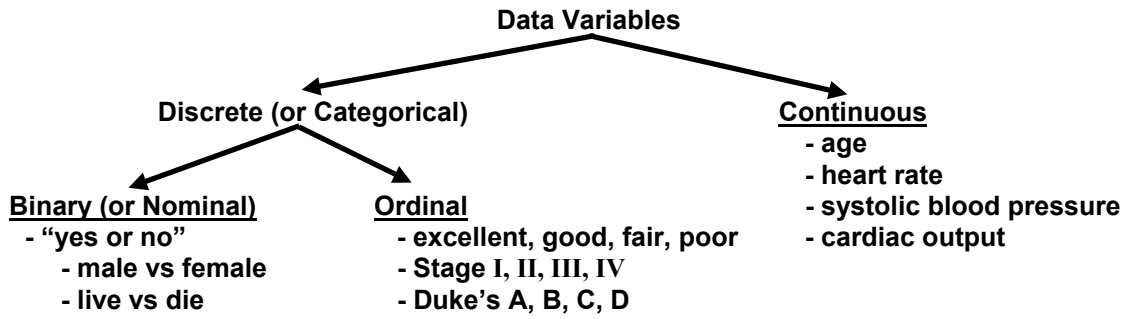



**Figure 2-1: Types of Data Variables**

**Discrete** or **categorical** variables are those which can only assume certain fixed or integer values. Discrete variables can be further divided into two categories: **binary** (or **nominal**) and **ordinal**. Binary or nominal variables are those whose values are either "yes" or "no." Examples include sex (male vs female), survival (live vs die), disease (cancer vs no cancer), or outcome (extubated vs reintubated). Ordinal variables are those whose values fall into categories, but are not limited to only two results. Examples of ordinal variables include outcome (excellent, good, fair, poor, no change) or disease stages (such as tumor staging classifications). **Continuous** variables are those which can assume an infinite range of values. Age, heart rate, systolic blood pressure, and cardiac output are examples of continuous variables. It is important to make a distinction between discrete and continuous variables as each requires a different set of statistical tests for proper analysis.

Variables can be further categorized as being either **dependent** or **independent**. **Dependent** variables are those whose value depends on that of another factor which is termed an **independent** variable. An example of a dependent variable is mean arterial blood pressure which is calculated from the systolic and diastolic blood pressures (both of which are examples of independent variables). A **confounding** variable is one which is closely associated with the outcome of interest (the dependent variable) such that it is unclear whether an observation is a reflection of the relationship between the confounding and dependent variables or between the independent and dependent variables. Consider a study that addresses whether therapy with $H_2$ blockers (the independent variable) influences the occurrence of peptic ulcer disease (the dependent variable). If fewer ulcers are seen in patients receiving $H_2$ blockers, we might conclude that $H_2$ blockers decrease the incidence of peptic ulcer disease. If the study patients were also receiving antacid therapy (a potential confounding variable), however, our conclusions might be in error as the decrease in ulcer incidence could be due to the $H_2$ blockers or the antacids or a combination of the two. There is an infinite list of potential confounding variables in any study and the goal of study design is to minimize the potential effect of such

variables on data analysis and study conclusions. It is important to account for confounding variables in the design and statistical analysis of studies in order to avoid incorrect conclusions. The most effective way to accomplish this is to perform a randomized study in which the effect of confounding variables is theoretically be distributed equally between the study and control groups (see Chapter Nine).

Observations collected on a study sample will tend to be grouped in some form of pattern. Only rarely will all of our observations be identical. In most situations, they will all be slightly different due to natural biologic variability. They will, however, tend to be grouped towards a middle or central point. If our sample closely approximates the population of interest, observations from the sample and the population should be centered around the same point. One way to confirm that our sample data is representative of the population is to plot or graph the observations. This results in a **frequency distribution** or **histogram** such as that below (Figure 2-2) where each "x" indicates a patient with a post-operative myocardial infarction.

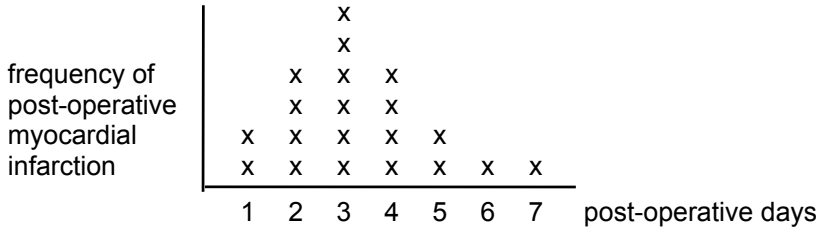
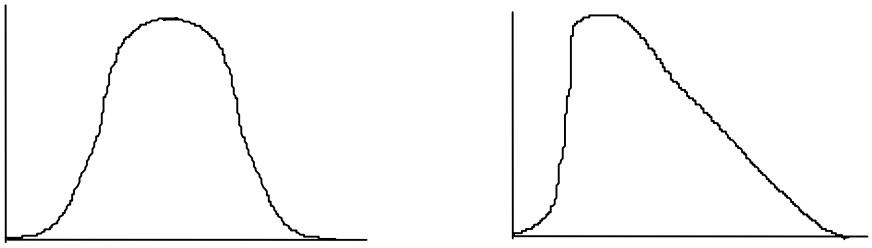



**Figure 2-2: Frequency histogram**

From this histogram it is clear that the observations are more or less centered around the 3rd day and appear to be equally distributed on either side. We can conclude from this graph that most post-operative myocardial infarctions occur on the 3rd post-operative day, but can occur with less frequency on any of the first 7 days following surgery.

Such a symmetric pattern (which takes the appearance of a "bell-shaped" curve) is called a **normal**, **gaussian**, or **z distribution** (after the German mathematician, Johann Gauss, who first described it) and is characterized by the distribution having the same slope on both sides of the center of the data (Figure 2-3a). Not all observations are symmetrically distributed, however. If the observations fall predominantly to one side or the other and are not evenly distributed around the center of the data (as in a normal distribution), the data are said to be **skewed** or **non-normally distributed** (Figure 2-3b). Just as the type of variable determines the correct statistical test to be used, the type of distribution determines whether a statistical test for normally or non-normally distributed data must be used.



a) normal (z) distribution b) non-normal distribution

**Figure 2-3: Types of distributions**

Data which are skewed or non-normally distributed can present problems with statistical analysis. These problems can be overcome by using tests designed specifically for non-normally distributed data or by "transforming" the data such that it fits a normal distribution (see Chapter Six).

## SOURCES AND TYPES OF BIAS

There are three potential sources of error in statistical analysis. **Chance** error occurs as a result of the normal random variation that is inherent in nature and requires us to use statistical analysis. There is little that can be done to control error due to natural variation other than to account for it in any statistical analysis.

The second potential source of error is **errors in data entry** or **errors in statistical analysis**. Errors in data entry, if undetected, can significantly alter the conclusions of a study. It is therefore a good practice to always review the raw data to ensure that the data is accurate and valid. Errors in statistical analysis result when an inappropriate test is used or the appropriate test is used incorrectly. Errors such as these may result in false conclusions (a Type I error) or in ignoring an important conclusion that is valid (a Type II error). An understanding of basic biostatistics and the assistance of a biostatistician should minimize these sources of error.

The third potential source of error, **statistical bias**, is perhaps the most difficult problem encountered in statistics. Sources of bias may be very obvious or quite subtle. Bias can enter a study at any stage from study design to data collection to data analysis and interpretation. The avoidance of statistical bias is a central concern in study design and its detection, when present, a goal of study interpretation.

There are numerous sources of statistical bias. Perhaps the easiest to avoid is **measurement bias** which occurs when the study groups are not evaluated in a consistent and uniform manner. This may include the way in which data is collected, measurements are made, or data is analyzed. It may also occur when measuring equipment is not used correctly or is improperly calibrated resulting in a systematic error being propagated throughout the data collection and analysis. It is important that all study groups are treated and evaluated similarly in order to avoid this source of bias.

**Confounding bias**, as discussed previously, occurs when two variables are interrelated and lead to an erroneous conclusion. The independent variable is concluded to have caused a change in the dependent variable when in reality it is the relationship between the confounding and dependent variables that is responsible for the change. There are multiple confounding variables in any study and careful study design and data interpretation are essential in order to avoid making such erroneous conclusions.

**Assembly or selection bias** is said to be present when observations are made on patients which have been assembled incorrectly such that they differ from the other study groups or population-at-large with regards to the characteristic of interest. An example would be a study of disease survival in which sicker patients are consistently placed in one of the groups such that the study groups are not equal. Physicians can introduce selection bias into a study by including or excluding patients from studies who they feel (either consciously or unconsciously) would benefit or be harmed by the study protocol. Exclusion of patients who meet the entry criteria is a common source of selection bias and can significantly alter the study's conclusions. A patient’s refusal to enter a study may likewise lead to bias if it results in the study groups being dissimilar or not representative of the patient population of interest. Selection bias is a major reason for performing **double blinded controlled trials** in which neither the physician nor patient is aware of which treatment or therapy will be received.

**Procedure bias** occurs when the study groups are not treated in a similar manner. Study patients may receive more attention than the control patients simply by virtue of being in a study or having a disease. This increased attention may result in differences between the study and control groups that are not related to the therapeutic intervention under study. Procedure bias is also a concern in retrospective reviews where a newer therapy is compared to an older therapy in a previous patient group. If the current and retrospective groups received differing treatment aside from the therapy of interest, they may not be comparable and invalid conclusions may result.

**Migration bias** occurs when patients move from one study group to another, are lost to follow-up, or drop out of the study. All of these situations may change the data sufficiently to affect the validity of the ultimate study conclusions.

**Compliance bias** is a concern in trials that require patients to take a medication or perform an event (such as exercise). Patients may find it preferable to not take the medication (due to taste, route of administration, or frequency of dosing) or avoid the required therapy. This can significantly confuse the study results and conclusions. **Intent-to-treat bias** is a form of compliance bias which results when investigators

attempt to take patients' noncompliance into account in their data analysis. Intent-to-treat bias is also commonly seen in the medical literature when a patient failed to receive the therapy or intervention to which they were originally randomized and are analyzed with another group of patients. Randomized studies are designed to minimize confounding variables and the problems associated with selection bias. By analyzing patients based on the therapy they actually received rather than by that to which they were originally randomized, the benefits of randomization are lost and significant bias may be introduced into the study. In such situations, **intent-to-treat analysis** should be performed where patients are analyzed by their original randomization and not by their compliance to treatment. This may result in a smaller clinical effect being seen, but avoids the potential for introducing further bias.

Statistical bias is common in medical research and can never be completely eliminated from a study or clinical trial. Every attempt should be made, however, to minimize the influence of bias on the data analysis and study conclusions. Further methods of addressing statistical bias are discussed in Chapter Nine.

## SUGGESTED READING

1. Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals. New York: Springer-Verlag, 1990:7-8, 74-75.
2. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics (2nd Ed). Norwalk, CT: Appleton and Lange, 1994:21-22, 271-274.
3. Moses LE. Statistical concepts fundamental to investigations. In: Bailar JC, Mosteller F (Eds). Medical uses of statistics (2nd Ed). Boston: NEJM Books, 1992:5-26.
4. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology - the essentials. Baltimore: Williams and Wilkins, 1982:6-11, 118-122.