These notes are supplementary material for Mathematics 243 and are not intended to stand alone. They should be used in conjunction with the textbook and the classroom discussion. Please inform the author of any errors, typographical or otherwise that you find in them.

1 Types of Data

The OED defines data as “facts and statistics used for reference or analysis.” (And the OED notes that while the word data is technically the plural of datum, it is often used with a singular verb and that usage is now generally deemed to be acceptable.) For our purposes, the sort of data that we will use comes to us in collections or datasets. A dataset consists of a set of objects, variously called observations, items, instances, units, or subjects, together with a record of the value of a certain variable or variables defined on the items.

Definition 1.1. A variable is a function defined on the set of objects.

For example, we record the SAT scores and GPAs of 30 seniors. The 30 seniors are the items. The SAT score of a given senior can be thought of as the value of a function defined on the set of 30 seniors. In this instance then, a variable is simply a numerical characteristic of the items. We will usually use uppercase letters to denote variables (e.g., $X$) and lowercase letters (e.g., $x$) to denote specific values of the variables. A number of distinctions among types of variables are important.

1.1 Categorical and Numerical Variables

Some variables, like the SAT and GPA variables mentioned above, are numerical variables. In other words, the range of these variables is a set of real numbers. Note that these numbers almost always come with units. Other variables are nonnumerical and for the purpose of statistical analyses simply serve to divide the subjects up into categories. For example, the variable CLASS when applied to Calvin students, has values FR, SO, JR, SR and serves to divide Calvin undergraduates into four categories. Often categorical variables are transformed into numerical variables by numbering the categories (as in 1,2,3,4 in the example of the variable CLASS). As you might expect, we usually ask different kinds of questions about numerical and categorical variables and do different kinds of analyses.
1.2 Continuous and Discrete Variables

This distinction refers to the possible values that a variable that have, i.e., the range of the variable.

**Definition 1.2.** A variable is **continuous** if for every \( \epsilon > 0 \) and possible value \( x \) of the variable, there is a possible value \( y \) of the variable such \( |x - y| < \epsilon \).

Almost always, our continuous variables have as their range an interval of real numbers. For example, AGE is a variable defined on all living persons and is a real number in the interval \((0, \infty)\) (once we specify the units such as minutes).

**Definition 1.3.** A variable is **discrete** if there is \( \epsilon > 0 \) such that if \( x, y \) are two possible values of the variable then \( |x - y| > \epsilon \).

The variable SAT described above is discrete since the possible values of SAT are the numbers in the set \( \{200, 210, 220, \ldots, 780, 790, 800\} \). In this instance, \( \epsilon = 5 \) suffices. Discrete variables usually have values that are natural numbers (or sometimes rational numbers with fixed denominator) such as counts. Not all variables are either discrete or continuous. For example, a variable with possible values in \( \{0, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1\} \) is neither.

We will have different methods of analyses depending on whether a variable is discrete or continuous. However, given the precision of our measuring instruments, most variables that are theoretically continuous are practically discrete. HEIGHT (in inches) is a continuous variable but in practice we report heights to the nearest half or quarter of inch. HEIGHT (to the nearest quarter inch) is a discrete variable. Nevertheless we we will usually assume that the variable is continuous and often assume that our measurement is exact. Similarly, sometimes a discrete variable will be treated as continuous if it has too many different values. We will even profit from treating SAT as a continuous variable even though the definition of continuous in this case is not satisfied for \( \epsilon = 10! \).

1.3 Measurement Scales

Numerical variables are generally used to measure something, some underlying property of the object on which the variable is defined. If \( x_1 \) and \( x_2 \) are values of a variable \( X \) on objects \( O_1 \) and \( O_2 \) respectively, we make inferences about the differences between \( O_1 \) and \( O_2 \) on the underlying property that variable \( X \) measures. For example, if \( X \) is AGE, we say someone aged 10 is “twice” as old as someone aged 5. However we do not say that someone with SAT score 400 did twice as well on the SAT test as someone who scored 200. This distinction has to do with the scale of values used to measure these variables.

1. **Nominal scale.** A measurement scale for a variable \( X \) is a **nominal scale** if \( x_1 \neq x_2 \) means only that \( O_1 \) and \( O_2 \) are different (with respect to \( X \)). A variable with a nominal scale is essentially a categorical variable for which numbers are the names of the categories. For example, for the variable GENDER we could use 1 to mean female and 0 to mean male (or vice versa) or we could use the numbers 37 and 25 respectively. The actual values of the number carry no other significance about gender.

2. **Ordinal scales.** A measurement scale for the variable \( X \) is an **ordinal scale** if \( x_1 < x_2 \) implies that \( O_1 \) comes before \( O_2 \) in a natural ordering of the variables by what \( X \) measures. For example, FINISH-ORDER of runners in a race is an ordinal variable (as is any ranking variable) since the order of the values of the variable is significant.
3. **Interval scales.** A measurement scale for a variable $X$ is an **interval scale** if $x_1 - x_2 = r$ means that object $X_1$ has $r$ more of some quantity than object $O_2$. For example, TEMPERATURE as a measure of heat is an interval variable since a difference of 10 degrees means a certain heat difference regardless of where on the scale it falls. The information in an interval scale is preserved by any linear function but not by a nonlinear function. FINISH-ORDER is not an interval variable (as a measure of speed, at least) as the difference in speed or time to finish between runners 1 and 2 is not necessarily the same as that between runners 2 and 3.

4. **Ratio variables.** A measurement scale for a variable $X$ is a **ratio scale** if $x_1 / x_2 = r$ implies that $O_1$ has $r$ times as much of some quantity as does object $O_2$. HEIGHT is a ratio variable since “twice as tall” means twice of something (linear space consumed). TEMPERATURE is not a ratio variable (as a measure of heat) since 20 degrees is not twice as hot as 10 degrees. Ratio scales have a natural 0 whereas interval scales do not. (We could add 10 to all temperatures and preserve the information in the relative temperature readings.)

Many of the analyses we perform will assume that we are dealing with a variable that is on an interval scale. Sometimes ratio variables will be assumed. Sometimes we will pretend the variable is on an interval scale (SAT scores is a prime example) when it is really only on an ordinal scale. (SAT scores only order students on the basis of how many questions they got right.)

**Exercises**

1.1 Classify the following variables as discrete or continuous and as nominal, ordinal, interval or ratio
   a. CLASS
   b. CLASS-RANK
   c. GPA
   d. WEIGHT
   e. GENDER
   f. FT-PCT (the free-throw percentage of a basketball player)
   g. RED-BLOOD-CELL-COUNT (the number of red blood cells per cc of blood)
   h. SCORE (the score of a golfer in a particular round of tournament golf)

2  **Sampling**

Many of the datasets that statisticians analyze arise from sampling from some population.

**Definition 2.1.** A **population** is a (well-defined) set of objects.

As with any mathematical sets, sometimes we define a population by a **census** or enumeration of the elements of the population. The registrar can easily produce an enumeration of the population of all currently registered Calvin students. Other times, we define a population by properties that determine membership in the population. (In mathematics, we define sets like this all the time since many sets
in mathematics are infinite and so do not admit enumeration.) For example, the set of all Michigan registered voters is a population even though a census of the population would be very difficult to produce. It is perfectly clear for an individual whether that individual is a Michigan registered voter or not.

**Definition 2.2.** A subset \( S \) of population \( P \) is called a **sample** from \( P \).

Quite typically, we are studying a population \( P \) but have only a sample \( S \) and have the values of one or several variables for each element of \( S \). The canonical problem of (inferential) statistics is:

**Problem 2.3.** Given a sample \( S \) from population \( P \) and values of variables \( X_1, \ldots, X_k \) on elements of \( S \), make inferences about the values of \( X_1, \ldots, X_k \) on the elements of \( P \).

Obviously, our success at solving this problem will depend to a large extent on how representative \( S \) is of the whole population \( P \) with respect to the properties measured by \( X_1, \ldots, X_k \). In turn, the representativeness of the sample will depend how the sample is chosen. A **convenience sample** is a sample chosen simply by locating units that conveniently present themselves. A convenience sample of Calvin students could be produced by grabbing the first 40 students that come through the doors of Johnny’s. It’s pretty obvious that in this case, and for convenience samples in general, there is no guarantee that the sample is likely to be representative of the whole population. In fact we can predict some ways in which a “Johnny’s sample” would not be representative of the whole student population.

One might suppose that we could construct a representative sample by carefully choosing the sample according to the important characteristics of the units. For example, to choose a sample of 40 Calvin students, we might ensure that the sample contains 22 females and 18 males. Continuing, we would then ensure a representative proportion of first-year students, dorm-livers, etc. There are two problems with this strategy. First, there are usually so many characteristics that we might consider that we would have to take too large a sample so as to get enough subjects to represent all the possible combinations of characteristics in the proportions that we desire. Second, even if we list many characteristics, it might be the case that the sample will be unrepresentative according to some other characteristic that we didn’t think of and that characteristic might turn out to be important for the problem at hand.

Statisticians have settled on using sampling procedures that employ chance mechanisms. The simplest such procedure (and also by far the most important) is known as **simple random sampling** (SRS).

**Definition 2.4.** A **simple random sample** of size \( k \) from a population of size \( n \) is a sample that results from a procedure for which every subset of size \( k \) has the same chance to be the sample chosen.

For example, to pick a random sample of size 30 of Calvin students, we might write the names of all Calvin students on index cards and choose 30 of these cards from a well-mixed bag of all the cards. In practice, random samples are often picked by computers that produce “random numbers.” (A computer can’t really produce random numbers since a computer can only execute a deterministic algorithm. However computers can produce numbers that “look” random. More on this issue later in the course.) In this case, we would number all students from 1 to 4135 and then choose 30 numbers from 1 to 4135 in such a way that any set of 30 numbers has the same chance of occurring. The R command `sample(1:4135,30,replace=F)` will choose such a set of 30 numbers.

Now it is certainly possible that a random sample is unrepresentative in some significant way. Since all possible samples are equally likely to be chosen, by definition it is possible that we choose a
bad sample. For example, a random sample of Calvin students might fail to have any seniors in it. However the fact that a sample is chosen by simple random sampling enables us to make quantitative statements about the likelihood of certain kinds of nonrepresentativeness. This in turn will enable us to make inferences about the population and to make statements about how likely it is that our inferences are accurate.

The concept of random sampling can be extended to produce samples other than simple random samples.

**Definition 2.5. A random sample** of size $k$ from a population of size $n$ is a sample that results from a procedure for which every object has a fixed likelihood of being in the sample chosen.

This definition differs from that of a simple random sample in two ways. First, it does not require that each object has the same likelihood of being the sample chosen. In opinion polls, sometimes specific groups are “oversampled.” For example, we might want to ensure in a sample of Calvin students that there are enough international students so that we can make some inferences about subgroups. To do this we might construct the sampling method to make it more likely that an international student is chosen in the sample. Second, it does not require that equal likelihood extends to groups. A sampling method that we might employ given a list of Calvin students is to choose one of the first 30 students in the list and then choose every 30th student thereafter. Obviously some subsets can never occur as the sample since two students whose names are next to each other in the list can never be in the same sample. Such a sample might indeed be representative however. Stratified random sampling as described in the textbook is an example of a random sampling method that does not result in simple random samples.

**Exercises**

2.1 Often, we take a sample by some convenient method (a convenience sample) and hope that the sample “behaves like” a random sample. For each of the following convenient methods for sampling Calvin students, indicate in what ways that the sample is likely not to be representative of the population of all Calvin students.

a. The students in Mathematics 243A.
b. The students in Nursing 329.
c. The first 30 students who walk into the FAC west door after 12:30 PM today.
d. The first 30 students you meet on the sidewalk outside Hiemenga after 12:30 PM today.
e. The first 30 students named in the bod book.
f. The men’s basketball team.

2.2 For the convenience samples in the previous problem, which do you suppose is likely to behave most like a random sample?

2.3 Donald Knuth, the famous computer scientist wrote a book entitled “3:16”. This book was a Bible study book that studied the 16th verse of the 3rd chapter of each book of the Bible (that had a 3:16). Knuth’s thesis was that a Bible study of random verses of the Bible might be
edifying. The sample was of course not a random sample of Bible verses and Knuth had ulterior motives in choosing 3:16. Describe a method for choosing a random sample of 60 verses from the Bible.

2.4 Consider the set of natural numbers $P = \{1, 2, \ldots, 30\}$ to be a population.

a. How many prime numbers are there in the population?

b. If a sample of size 10 is representative of the population, how many prime numbers would we expect to be in the sample? How many even numbers would we expect to be in the sample?

c. Using R choose 5 different samples of size 10 from the population $P$. Record how many prime numbers and how many even numbers are in each sample. Make any comments about the results that strike you as relevant.