MAT 2379, Introduction to Biostatistics

Chapter 9. Introduction to statistics
9.1 Random Sampling and Data Description

We will start by examining the question:

What is a statistical problem?

For example, a researcher is examining a group of 50 people of age 18+ and is interested in examining the factors which are associated with the development of heart disease. These factors include: age, weight, smoking status, and the family history.

This problem has all the characteristics of a statistical problem:

1. Associated to this problem, there is a large group of objects about which we have to draw a conclusion. This group of objects is called the *population*. In the example above, the population consists of all people of age 18+.
2. We are interested in certain characteristics of the members of the population. These are called *variables* and are denoted by the capital letters X, Y, Z, etc. In the example above, X =age, Y = weight, Z =smoking status.
3. The population is too large to study. So, we must draw conclusions by studying only a portion of the population called a *sample*. The number of objects in the sample is called the *sample size* and is denoted by n. In the example above, $n = 50$. The observed values of the variable X (say age) are denoted by $x_1, \ldots, x_n$.

In Section 9.1, we will examine questions about the behavior of variables:

- What graphical methods to use to summarize the observed values?
- About what values does the variable fluctuate?
- What is the variation in the observed values?

Variables are of 2 types:

1. **Categorical variables**. These are variables whose values fall in several categories. For instance: blood type, sex, color of a flower.
2. **Quantitative variables**. These are variables which take numeric values. They are of two subtypes:
   a) *discrete variables*, whose values can be listed (e.g. age, number of seeds which germinate)
   b) *continuous variables*, whose values lie in a certain range but cannot be listed (e.g. weight, height, blood pressure).

Example: In each case below, specify what is the population, the sample size, the variables of interest (and their type).
a) In a large hospital, the birth weight and mother's blood type were recorded for 15 newborns.
b) A pediatrician measured the height increase over 1 year for 25 patients.
c) At a center blood donation, the cholesterol level was recorded for 100 donors.

Part 1: Frequency Distributions

Below will introduce various graphical methods for summarizing a data set.

a) The Bar Charts

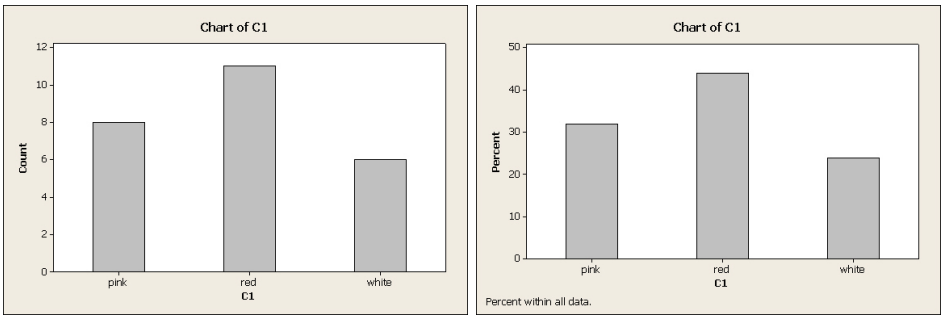
A bar chart is a graphical method used for categorical variables. The vertical bars have the same width and are separated by some arbitrary space. The heights of the bars represent the frequencies for each category of the variable. Alternatively, one can draw a bar chart of the relative frequencies:

$$\text{The relative frequency} = \frac{\text{the frequency}}{n}$$

Example 1. (Color of poinsettias) A sample of 25 poinsettias were classified according to their color as follows:

Color	Frequency	Relative frequency
Pink	8	0.32
Red	11	0.44
White	6	0.24
Total	25	1.00

Here are the bar charts of frequencies and relative frequencies:



b) Histograms

A histogram is a graphical method used for quantitative variables. The scale of the variable determines the placement of the bars. Usually, the vertical bars have the same width. Unlike the bar charts, the bars are not separated by some arbitrary space.

b1) Histograms for discrete variables. In this case, the width of the bars is (usually) 1 unit.

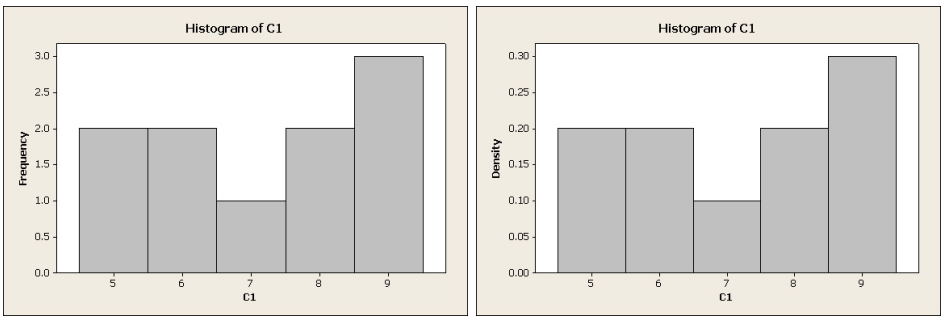
Example 2. (piglets) A company who owns a large number of pig farms is interested in the distribution of the number X of surviving piglets per sow. X is a discrete variable which can take the values: $0, 1, \ldots, 20$. A sample of 10 sows is selected. We record the observed values of X for this sample:

i	1	2	3	4	5	6	7	8	9	10
x_i	5	8	6	5	9	7	6	9	9	8

Below is the table of frequencies and relative frequencies for this data:

Number of Surviving Piglets	Frequency	Relative Frequency
5	2	0.2
6	2	0.2
7	1	0.1
8	2	0.2
9	3	0.3
Total	10	1.0

Here are the histograms of frequencies and relative frequencies:



b2) Histograms for continuous variables. In this case, we have to arrange the data into groups (called *bins*) and count how many values lie in each bin. One has to select the number of bins (this is a delicate issue, which will not be discussed here). Usually the bins have the same width.

Example 3. (height) A sample of 15 college students were asked how tall they were. Here is the data (in inches):

66.5 61.2 63.9 62.7 65.1 68.7 64.3 73.3 69.3 66.5 70.1 71.3 68.1 67.4 66.7

We arrange the data into 7 bins of equal width: bin 1 contains the values between 61 and 63, bin 2 contains the values between 63 and 65, etc. The following table gives the frequencies and the relative frequencies:

Height	Frequency	Relative Frequency
61-63	2	2/15
63-65	2	2/15
65-67	4	4/15
67-69	3	3/15
69-71	2	2/15
71-73	1	2/15
73-75	1	1/15
Total	15	1.0

Here are the histograms of the frequencies and relative frequencies: