Chapter 6 Statistics and Sampling Distributions

This chapter helps make the transition between probability and inferential statistics.

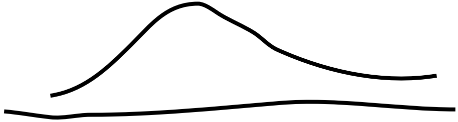
- Each sample quantity such as sample mean is an **estimate** of its population counterpart, which is referred to as a **parameter**.
- The behavior of such estimates in **repeated sampling** is described by what are called **sampling distributions**.
- A sampling distribution will give an indication of how close the estimate is likely to be to the value of the parameter being estimated.

Outline

Consider a **population distribution**, say that of the heights of *all* UNB students.

Let X be the height of a randomly selected student $\Rightarrow$

<u>The distribution of the random variable X</u> is the population distribution (distribution of variable height in the population, which might be represented by a smooth curve).



Height

Suppose now we take a sample $(x_1, x_2, \ldots, x_n)$ and construct a histogram.

The histogram displays the distribution of the variable height in the sample (we may call it the **sample distribution**, compared with the population distribution.)

Consider now selecting samples (of the same size n) repeatedly from the population distribution:

$$\text{Sample 1: } (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}) \Rightarrow \bar{x}^{(1)}, s^{(1)}$$

$$\text{Sample 2: } (x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}), \Rightarrow \bar{x}^{(2)}, s^{(2)}$$

$$\vdots$$

$$\text{Sample } N\text{: } (x_1^{(N)}, x_2^{(N)}, \ldots, x_n^{(N)}), \Rightarrow \bar{x}^{(N)}, s^{(N)}$$

$$\vdots$$

Before we take a sample, there is uncertainty about the value of each x_i. Consequently, prior to obtaining $x_1, \ldots, x_n$, there is uncertainty as to the value of $\bar{x}$, the value of s, and so on.

Because of this uncertainty, before the data become available we view **each observation as a random variable** and denote the sample by $X_1, X_2, \ldots, X_n$ (uppercase letters for random variables).

$x_1, \ldots, x_n$

$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

A **statistic** is any quantity whose value can be calculated from sample data.

Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result.

Therefore, a **statistic is a random variable** and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

$\bar{X}$ is a statistic.

$\bar{x}$ is not.

Thus the sample mean, regarded as a statistic (before a sample has been selected or an experiment has been carried out), is denoted by $\bar{X}$; the calculated value of this statistic is $\bar{x}$.

Similarly, S represents the sample standard deviation regarded as a statistic, and its computed value is s.

$\bar{X} - \mu$ is **not** a statistic.

$X_1, \dots, X_n$

sample summary as a r.v.

Any statistic, being a random variable, has a probability distribution.

The probability distribution of a statistic is referred to as its **sampling distribution** to emphasize that it describes how the statistic varies in values across all samples that could be selected.

The probability distribution of any particular statistic depends on

- the population distribution (normal, uniform, etc.),
- the sample size n, and also
- the method of sampling.

Pop = {1, 5, 10}

Consider selecting a sample of size $n = 2$ from a population consisting of just the three values $1, 5$, and 10, and suppose that the statistic of interest is the sample variance.

If sampling is done "with replacement," then $s^2 = 0$ will result if $x_1 = x_2$.

{1, 1}
{1, 5}
{1, 10}
{5, 5}
{5, 10}
{10, 10}

However, S^2 cannot equal 0 if sampling is "without replacement."

{1, 5}
{1, 10}
{5, 10}

So P$(S^2 = 0) = 0$ for one sampling method, and this probability is positive for the other method.

Random sample

The rv's $X_1, X_2, \ldots, X_n$ are said to form a (simple) **random sample** of size n if

1. The X_i's are independent rv's.
2. Every X_i has the same probability distribution.

Conditions 1 and 2 can be paraphrased by saying that the X_i's are **independent and identically distributed (iid)**.

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

If sampling is with replacement, Conditions 1 and 2 are satisfied exactly.

These conditions will be approximately satisfied if sampling is without replacement, and the sample size n is much smaller than the population size N.

In practice, if $n/N \leq .05$ (at most 5% of the population is sampled), we can proceed as if the X_i's form a random sample.

The virtue of this simple sampling method is that the probability distribution of any statistic can be more easily obtained than for any other sampling method.