Lesson 2.1

Sampling Distributions Based on the Normal Distribution

Introduction

Recall that a random sample of observations is also referred to as an "iid" (*independent and identically distributed*) sample of observations $Y_1, Y_2, \cdots, Y_n$. That is, these observations are independent and come from the same probability distribution.

Definition

A **statistic**, say T, is a function of the random variables $Y_1, Y_2, \cdots, Y_n$. A statistic can depend om known constants, but it cannot depend on unknown parameters.

To denote the dependence of T on $Y_1, Y_2, \cdots, Y_n$, we may write

$$T = T(Y_1, Y_2, \cdots, Y_n)$$

In addition, while it often be the case that $Y_1, Y_2, \cdots, Y_n$ constitute a random sample, the above definition of T holds in more general setting. In practice, it is common to view $Y_1, Y_2, \cdots, Y_n$ as **data** from an experiment or observational study and T as some summary measure (such as sample mean, sample variance, etc.).

Example 2.1.1

Suppose that $Y_1, Y_2, \cdots, Y_n$ is an iid sample from $f_Y(y)$. The following are statistics:

- $T = T(Y_1, Y_2, \cdots, Y_n) = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- $T = T(Y_1, Y_2, \cdots, Y_n) = \frac{1}{2}[Y_{(n/2)} + Y_{(n/2+1)}]$
- $T = T(Y_1, Y_2, \cdots, Y_n) = Y_{(1)}$

- $T = T(Y_1, Y_2, \cdots, Y_n) = Y_{(n)} - Y_{(1)}$
- $T = T(Y_1, Y_2, \cdots, Y_n) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

It is very important to note that since $Y_1, Y_2, \cdots, Y_n$ are random variables, any statistic $T = T(Y_1, Y_2, \cdots, Y_n)$, being a function of random variables, is also a random variable. Thus, T has its own distribution.

Definition

The probability distribution of a statistic T is called its **sampling distribution**. The sampling distribution of T describes mathematically how the values of T vary in repeated sampling from the population distribution $f_Y(y)$. Sampling distributions play a crucial role in statistics

Sampling Distributions based on the normal distribution

Example 2.1.2

Suppose $Y_1, Y_2, \cdots, Y_n$ is an iid sample from $N(\mu.\sigma^2)$ and consider the statistic

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

the sample mean. It can be shown (via MGF technique) that

$$\overline{Y} \sim N\Big(\mu, \frac{\sigma^2}{n}\Big)$$

Furthermore, the quantity

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Example 2.1.2

In the interest of pollution control, an experimenter records Y, the amount of bacteria per unit volum of water (measured in mg/cm^3). The population distribution for Y is assumed to be normal with mean $\mu = 48$ and variance $\sigma^2 = 100$, that is, $Y \sim N(48, 100)$.

a. What is the probability that the amount of bacteria in single water sample exceeds $50\, mg/cm^3$?

b. Suppose the experimenter takes a random sample of $n = 100$ water samples and denote the observations by $Y_1, Y_2, \cdots, Y_{100}$. What is the probability that the sample mean $\overline{Y}$ will exceed $50\, mg/cm^3$?

c. How large should the sample size n be so that $P(\overline{Y} > 50) < 0.01$?

SOLUTION: [Left as a classroom exercise!]