LECTURE 3: THE POPULATION AND THE SAMPLE 2.1 – 2.4

Population Parameters

Suppose there are N elements in a population. Let $X_1, \ldots, X_N$ denote the true values of a certain characteristic of interest for all elements in this population.

Notation. The population total is denoted by

$$X = \sum_{i=1}^{N} X_i \,.$$

The population mean is

$$\bar{X} \;=\; \frac{X}{N} = \frac{\sum_{i=1}^{N} X_i}{N} \,.$$

The population variance is

$$\sigma_X^2 \;=\; \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N} \,.$$

The population standard deviation is

$$\sigma_X \;=\; \sqrt{\sigma_X^2} \,.$$

The population coefficient of variation is

$$V_X \;=\; \frac{\sigma_X}{\bar{X}} \,.$$

It represents the ratio of the population standard deviation to the population mean.

The population relative variance is V_X^2.

Special case. Let $X_i = 1$ if a certain characteristic is present in element i, and 0, otherwise. Then the population proportion is

$$P_X \;=\; \frac{X}{N} \;=\; \frac{\sum_{i=1}^{N} X_i}{N} \;=\; \frac{\#\text{ with characteristic}}{N} \,,$$

and the population variance in this case is

$$\sigma_X^2 \;=\; P_X \,(1 \,-\, P_X) \,.$$

The population coefficient of variation is

$$V_X \;=\; \sqrt{\frac{1 \,-\, P_X}{P_X}} \,.$$

Sample Parameters

Suppose there are n elements in a population. If $N \gg n$, the sample is said to be taken from a large (or infinite) population. Let $x_1, \ldots, x_n$ denote the values of a certain characteristic of interest for all elements in this sample.

Notation. The sample total is denoted by

$$x = \sum_{i=1}^{n} x_i \,.$$

The sample mean is

$$\bar{x} = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \,.$$

The sample variance is

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \,.$$

The sample standard deviation is

$$s_x = \sqrt{s_x^2} \,.$$

Special case. Let $x_i = 1$ if a certain characteristic is present in element i, and 0, otherwise. Then the sample proportion is

$$p_x = \frac{x}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\#\text{ with characteristic}}{n} \,,$$

and the sample variance in this case is

$$s_x^2 = \frac{n p_x (1 - p_x)}{n-1} \,.$$

If n is large, $s_x^2 \simeq p_x(1 - p_x)$.

Estimation of Population Characteristics

The sample mean is an unbiased estimator of the population mean, $\hat{\bar{X}} = \bar{x}$. An unbiased estimator of the population variance is

$$\hat{\sigma}_X^2 = s_x^2 \left(\frac{N-1}{N} \right) .$$

2.3 Sampling Distribution

Denote a population parameter of interest by d. Suppose a sampling plan can result in T possible samples from the population. Let $\hat{d}$ be an estimator of d. Suppose that $\hat{d}$ can assume C values $\hat{d}_1, \ldots, \hat{d}_C$ with respective probabilities $\pi_1, \ldots, \pi_C$. The sampling distribution of $\hat{d}$ with respect to the specified

sampling plan is the values $\hat{d}_1, \ldots, \hat{d}_C$ and the probabilities $\pi_1, \ldots, \pi_C$.

The mean of the sampling distribution is the expected value of $\hat{d}$,

$$\mathbb{E}(\hat{d}) \;=\; \sum_{i=1}^{C} \hat{d}_i \, \pi_i \,.$$

The variance of the sampling distribution is the variance of $\hat{d}$,

$$\mathbb{V}ar(\hat{d}) \;=\; \sum_{i=1}^{C} \big[\, \hat{d}_i - \mathbb{E}(\hat{d}) \,\big]^2 \pi_i \;=\; \sum_{i=1}^{C} \hat{d}_i^2 \pi_i \;-\; \big[\, \mathbb{E}(\hat{d}) \,\big]^2 \,.$$

The standard deviation of the sampling distribution (or standard error) is the square root of the variance,

$$SE(\hat{d}) \;=\; \sqrt{\mathbb{V}ar(\hat{d})} \,.$$

Special case. If all T samples are equally likely, then

$$\pi_i \;=\; \frac{\#\text{ of times } \hat{d}_i \text{ occurs}}{T} \,.$$

The mean of the sampling distribution has the form

$$\mathbb{E}(\hat{d}) \;=\; \frac{\sum_{i=1}^{T} \hat{d}_i}{T} \,,$$

and the variance of the sampling distribution is

$$\mathbb{V}ar(\hat{d}) \;=\; \frac{\sum_{i=1}^{T} \big[\, \hat{d}_i \;-\; \mathbb{E}(\hat{d}) \big]^2}{T} \,.$$

2.4 Characteristics of Estimates of Population Parameters

The bias of an estimate $\hat{d}$ of a population parameter d is the deviation of the mean of the sampling distribution of $\hat{d}$ from the true parameter d, that is,

$$B(\hat{d}) \;=\; \mathbb{E}(\hat{d}) \;-\; d \,.$$

If the bias is equal to zero, the estimate $\hat{d}$ is called unbiased.

The mean square error of an estimate $\hat{d}$ is defined by

$$MSE(\hat{d}) \;=\; \sum_{i=1}^{C} \big(\hat{d}_i \;-\; d\big)^2 \pi_i \,.$$

Note that

$$MSE(\hat{d}) \;=\; \mathbb{V}ar(\hat{d}) \;+\; B^2(\hat{d}) \,.$$

Proof.

The reliability of an estimator $\hat{d}$ refers to how reproducible the estimator is over repetitions of the process yielding the estimator. Reliability can be stated in terms of variance of the sampling distribution of $\hat{d}$. The smaller the variance, the greater is its reliability.
The validity of $\hat{d}$ refers to how the mean of the estimator $\mathbb{E}(\hat{d})$ differs from the true population value d, that is, the smaller the bias, the greater is the validity.
The accuracy of the estimator $\hat{d}$ refers to how far on average a particular value of $\hat{d}$ is from the true value d, that is, the smaller $MSE(\hat{d})$, the greater is its accuracy.