What is the purpose of Statistics?

The purpose of Statistics is to use the information contained in a sample to make inferences about the population from which the sample is taken. Populations are characterized by numerical descriptive measures called **parameters**. The objective of many statistical investigations is to estimate the value of one or more relevant parameters.

What are parameters?

(1) Proportions - For example, a manufacturer of washing machines might be interested in estimating the proportions $p$ of washers that can be expected to fail prior to the expiration of a 1-year guarantee time.

(2) Population mean

(3) Population variance

(4) Population standard deviation

For example, we might want to estimate the mean waiting time $\mu$ at a supermarket check-out station, or the standard deviation of the error of measurement $\sigma$ of an electronic instrument.

These parameters of interest - parameters of the population we intend to estimate - are called *target parameters*.

Suppose we want to estimate the mean waiting time $\mu$ at a supermarket check-out station. We can give our estimate in two different forms:

(1) We could use a single number, for instance 3 min, that we think is close to the unknown population mean. This type of estimate is called a *point estimate* because a single value, or point, is given as the estimate $\mu$.

(2) We might say that $\mu$ falls between two numbers - for example, 2 min and 4 min. The two values may be used to construct an interval $(2, 4)$ that is intended to enclose the parameter of interest. this typeof an estimate is called an *interval estimate*. When we look for an interval estimate, we can find the two values using the sample data so that the target parameter will fall within the interval with certain probability.

The information in the sample can be used to calculate the value of a point estimate, an interval estimate, or both.

To find an actual estimation for the target parameter, we use a function called an *estimator*.

**Defintion** An estimator is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample.

For example, the sample mean

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

is one possible point estimator of the population mean $\mu$.

Many different estimators (rules for estimating) may be obtained for the same population parameter. Therefore, there are *Good* estimates and *Bad* estimates.

We cannot estimate the goodness of a point estimation procedure on the basis of the value of a single estimate; rather, we must observe the results when the estimation procedure is used many times.

An estimate is number which varies from sample to sample taken from a certain population, thus we can consider the distribution of estimates. Suppose we wish to specify a point estimate for a population parameter $\theta$.

The estimator of $\theta$ is indicated by $\hat{\theta}$. We would like the mean or expected value of the distribution of the estimates to equal the parameter estimated. That is, $E(\hat{\theta}) = \theta$.

Point estimators that satisfy the property $E(\hat{\theta}) = \theta$ are called *unbiased* estimators. When $E(\hat{\theta}) \neq \theta$, we say that $\hat{\theta}$ is a *biased* estimator. The *bias* of a point estimator $\hat{\theta}$ is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

In addition to unbiasedness, we want the variance of the distribution of the estimator $\mathrm{Var}(\hat{\theta})$ to be as small as possible. Given two unbiased estimators of a parameter $\theta$, and all other things being equal, we would select the estimator with the smaller variance.

Rather than using the bias and variance of a point estimator to characterize its goodness, we use $E[(\hat{\theta}-\theta)^2]$, the average of the square of the distance between the estimator and it target parameter.

**Definition** The Mean Square Error of a point estimator $\hat{\theta}$ is

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta}-\theta)^2].$$

It can be shown that $\text{MSE}(\hat{\theta})$ is a function of both its variance and bias.

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [B(\hat{\theta})]^2.$$

**Unbiased Point Estimators**

We can use

(1) sample mean $\overline{Y}$ to estimate the population mean $\mu$.

(2) sample proportion $\hat{p} = \frac{Y}{n}$ to estimate the binomial proportion $p$.

Consider two independent random samples of $n_1$ and $n_2$ observations selected from two different populations. How do we estimate the difference between means $(\mu_1 - \mu_2)$ and the difference in the two binomial proportions $(p_1 - p_2)$?

We can use the difference in sample means $\overline{Y}_1 - \overline{Y}_2$ to estimate $\mu_1 - \mu_2$.

We can use the difference in sample means $\hat{p}_1 - \hat{p}_2$ to estimate $p_1 - p_2$.

$\overline{Y}$, $\hat{p}$, $\overline{Y}_1 - \overline{Y}_2$ and $\hat{p}_1 - \hat{p}_2$ are functions of random variables observed in samples.

It is easy to check that

$$E(\overline{Y}_1 - \overline{Y}_2) = E(\overline{Y}_1) - E(\overline{Y}_2) = \mu_1 - \mu_2 \qquad \text{Var}(\overline{Y}_1 - \overline{Y}_2) = \text{Var}(\overline{Y}_1) + \text{Var}(\overline{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

and

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2 \qquad \text{Var}(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2},$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. The variance of the sampling distribution of the estimator $\hat{\theta}$ is denoted by $\sigma_{\hat{\theta}}^2$, and the standard deviation $\sigma_{\hat{\theta}} = \sqrt{\sigma_{\hat{\theta}}^2}$ is called the *standard error* of the estimator $\hat{\theta}$.

**Expected values and standard errors of some common point estimators**

| Target Parameter $\theta$ | Sample Size(s) | Point Estimator $\hat{\theta}$ | $E(\hat{\theta})$ | Standard Error $\sigma_{\hat{\theta}}$ |
|---|---|---|---|---|
| $\mu$ | $n$ | $\overline{Y}$ | $\mu$ | $\frac{\sigma}{\sqrt{n}}$ |
| $p$ | $n$ | $\hat{p} = \frac{Y}{n}$ | $p$ | $\sqrt{\frac{pq}{n}}$ |
| $\mu_1 - \mu_2$ | $n_1$ and $n_2$ | $\overline{Y}_1 - \overline{Y}_2$ | $\mu_1 - \mu_2$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^{*\dagger}$ |
| $p_1 - p_2$ | $n_1$ and $n_2$ | $\hat{p}_1 - \hat{p}_2$ | $p_1 - p_2$ | $\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}^{\dagger}$ |

*$\sigma_1^2$ and $\sigma_2^2$ are the variances of populations 1 and 2, respectively.

†The two samples are assumed to be independent.