# <u>**WEEK 9**</u>

## 9. Estimation

In the previous section, we have studied the basic ideas of sampling and sampling distributions. remember that a random sample is a sequence of independent and identically distributed random variables with a probability density function $f(x;\theta)$. Here, $\theta$ is the parameter which characterize the population. In this section we will try to discuss to estimate (to get some information) the parameter $\theta$. If $X_1, X_2 \ldots, X_n$ is a random sample, then we have seen that as $n \to \infty$

$$\bar{X}_n \xrightarrow{P} \mu, \; S_n^2 \xrightarrow{P} \sigma^2, \; \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1), \; \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{D} N(0,1)$$

where $\bar{X}_n$ and $S_n^2$ are the sample mean and sample variance respectively.

As we mentioned earlier, if $X$ is a random variable then any function of a random variable is also a random variable. That is, since a random variable is a function from the sample space to real line, $g(X)$ is also a function from the sample space to real line,

$$\begin{array}{lll} X : \Omega \to \mathbb{R} & g : \mathbb{R} \to \mathbb{R} & g(X) : \Omega \to \mathbb{R} \\ \quad w \to X(w) & \quad x \to g(x) & \quad w \to g(X(w)). \end{array}$$

<u>**Definition**</u>: Let $X_1, X_2 \ldots, X_n$ be any random sample from a population having a probability density function $f(x;\theta)$. Then any function of the sample is called an "*estimator*" or "*statistic*".

Note that a random sample is a sequence of random variables $X_1, X_2 \ldots, X_n$ denoted by $\underset{\sim}{X} = (X_1, X_2 \ldots, X_n)'$ and $T(\underset{\sim}{X})$ is any function of the sample and therefore it is an estimator. We will use $T$ instead of $T(\underset{\sim}{X})$. The estimator $T(\underset{\sim}{X})$ or just $T$ is a random variable. Since the estimator $T$ is a random variable, it is also a function from the sample space to the real line and therefore the value of $T$ is a real number ($T : \Omega \to \mathbb{R}$, $T(w) \in \mathbb{R}$ ). This number is called an "*estimate"*.

Estimator → A random variable

Estimate → A real number, The value of the estimator.

Let $X_1, X_2 \ldots, X_n$ be any random sample from a population having a probability density function $f(x;\theta)$ such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Note that

$$T(\underset{\sim}{X}) : \Omega \to \mathbb{R}$$
$$w \to T(\underset{\sim}{X})(w)$$

and some well-known estimators are given below:

| | |
|---|---|
| $T_1(\underset{\sim}{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$ | →sample mean |
| $T_2(\underset{\sim}{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ | →sample variance |
| $T_3(\underset{\sim}{X}) = X_{(n)} = \max\{X_1, X_2 \ldots, X_n\}$ | →nth order statistic |
| $T_4(\underset{\sim}{X}) = X_{(1)} = \min\{X_1, X_2 \ldots, X_n\}$ | →first order statistic |
| $T_5(\underset{\sim}{X}) = X_{(n)} - X_{(1)} = \mathcal{R}$ | →sample range |

You can write as many estimators as you can. We can also consider the median as an estimator. Based on the experimental data points, we calculate the values of any of these estimators. These values will be an estimate for the population parameter. The question is "which estimator should we use to estimate the parameter? Therefore, we need to evaluate the estimators based on some statistical properties.

**Example**: Consider an experiment of estimating the new-born baby weights. In order to estimate the new-born baby's weight, for a specific day we randomly select a hospital and randomly select 20 new-born babies and weights them. The results are given below:

Let $X_i$ be a random variable which measures the ith baby's weight e.g. $X_1(w) = x_1 = 3.75kg$. The rest of the values are given below.

| *i* | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| weight | 3.75 | 3.22 | 3.38 | 2.94 | 2.71 | 4.05 | 3.62 | 3.45 | 3.62 | 3.54 |
| *i* | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| weight | 3.65 | 3.70 | 3.85 | 3.27 | 3.45 | 2.90 | 3.64 | 3.28 | 4.12 | 3.86 |

Weights for 20 newborn babies

In order to calculate some statistical tendencies, we ordered the values from the smallest to the largest. The ordered values are given below.

| *i* | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| weight | 2.71 | 2.90 | 2.94 | 3.22 | 3.27 | 3.28 | 3.38 | 3.45 | 3.45 | 3.54 |
| *i* | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| weight | 3.62 | 3.62 | 3.64 | 3.65 | 3.70 | 3.75 | 3.85 | 3.86 | 4.05 | 4.12 |

The Ordered Values

Based on these observed values, we calculate some population parameters. First of all, we need to mention that no distributional properties yet. If the estimator is the sample mean $T(\underset{\sim}{X}) = \bar{X}_n$ which is a random variable and the value of this random variable is

$$\bar{X}_n(w) = \bar{x}_n = \frac{x_1 + x_2 + ... + x_{20}}{20} = \frac{70}{20} = 3.5 .$$

That is, the estimated population mean (say $\mu$) is $3.5kg$ . This means that based on this sample, it is expected that the newborn babies have weights $3.5$ kg. The estimated variance and the standard deviation (which is the positive square root of the variance) are calculated as

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = 0.13741053 \text{ and } s_n = +\sqrt{s_n^2} = 0.37068926 .$$

```
data a; input x@@; cards;

3.75 3.22 3.38 2.94 2.71 4.05 3.62 3.45 3.62 3.54

3.65 3.70 3.85 3.27 3.45 2.90 3.64 3.28 4.12 3.86

;

proc univariate normal plot; var x; run;
```

```
*****************************************************************************
**********

            N                    20    Sum Weights              20

            Mean                  3.5    Sum Observations          70

            Std Deviation      0.37068926    Variance          0.13741053

            Skewness          -0.4637137    Kurtosis          -0.0697348
```

```
          Uncorrected SS       247.6108    Corrected SS           2.6108

          Coeff Variation    10.5911217    Std Error Mean      0.08288864


**************************************************************************
*********

                    Basic Statistical Measures

              Location                  Variability

            Mean     3.500000     Std Deviation            0.37069

            Median   3.580000     Variance                 0.13741

            Mode     3.450000     Range                    1.41000

                          Interquartile Range      0.45000


**************************************************************************
*********

                      Quantiles (Definition 5)

                       Quantile      Estimate

                       100% Max         4.120

                       99%              4.120

                       95%              4.085

                       90%              3.955

                       75% Q3           3.725

                       50% Median       3.580

                       25% Q1           3.275

                       10%              2.920

                       5%               2.805

                       1%               2.710

                       0% Min           2.710
```

********************************************************************************
**********

```
                    Extreme Observations

               Value     Obs      Value     Obs

                2.71      5        3.75      1

                2.90     16        3.85     13

                2.94      4        3.86     20

                3.22      2        4.05      6

                3.27     14        4.12     19
```

********************************************************************************
***********

```
     Stem Leaf              #         Boxplot

       40 52                2            |

       38 56                2            |

       36 224505            6         +-----+

       34 554               3         *--+--*

       32 2788              4         +-----+

       30                                |

       28 04                2            |

       26 1                 1            |
```

**Normal Probability Plot**

```
     4.1+                             *++++*

       |                          *+*++

       |                      ****+**+

       |                   ***+++

       |                *+**+*+

       |             +++++
                 |     ++++*  *
                 |
               2.7+ +++++*
                    +----+----+----+----+----+----+----+----+----+----+
                        -2       -1        0       +1       +2
```

Some other values of the estimators have been calculated (in SAS) and given in the above table. Now let us try to explain these values.

In the table, the first part is the SAS codes to analyze the data. After we enter the data and run the SAS codes, the output is given right below the codes. There are some other outputs but we deleted the rest.

The second part of the table are the basic calculations about the data (mean, variance, standard deviation, skewness, kurtosis, coefficient of variation) which we explained the first section in the notes. The third part of the table contains information about the mean, median, mode, range and interquartile range. As we know, the median of the sample is a number $m$ such that 50% of all observations are less than or equal to $m$. Notice that there are 20 observations in the sample and when we ordered the data from smallest to the largest (which are given in the second table above), the median is

$$m=[x_{(10)}+x_{(11)}]/2=(3.54+3.62]/2=(7.16)/2=3.58.$$

The mode is the most repeated observation. If we check the data, the observations 3.45 and 3.62 have been observed twice and therefore anyone of these measurements can be considered as the mode of the sample. From the output (in the third part), the mode is given by 3.45. Standard deviation and the variance have been given in the second part (same as in the second part of the output). In that part of the table, the range and interquartile range have also been given. The range of the sample is the difference from the largest to the smallest. That is, $x_{(n)}=4.12$ , $x_{(1)}=2.71$ so the difference (range) is $Range=x_{(n)}-x_{(1)}=4.12-2.71=1.41$. The quartiles are given in the fourth part of the table, for the first quartile consider the first 10 smallest observations and find the median. The median of the first part $(x_{(5)}+x_{(6)})/2=(3.27+3.28)/2=6.55/2=3.275$. That is, 25% of all observations are less than or equal to 3.275. That is, $Q_1=3.275$. Similarly, in order to find the third quartile, consider the second part of the ordered data (largest 10 observations) and find the median of the second part. That is, the third quartile is

$Q_3 = (x_{(15)} + x_{(16)}) / 2 = (3.70 + 3.75) / 2 = 7.45 / 2 = 3.725$.

This means that 75% of all observations are less than or equal to 3.725. Therefore the interquartile range is the difference from the third quantile to the firs quartile. That is,

$IQR = Q_3 - Q_1 = 3.725 - 3.275 = 0.45$.

Some of the percentiles (quantiles) are given in the fourth part of the table. Remember that the pth percentile of the sample is a number $a_p$ that p% of all observations are less than or equal to the number $a_p$. For example, let us find the 90th percentile of the sample. The 90th percentile $a_{90}$ is found to be $a_{90} = (x_{(18)} + x_{(19)}) / 2 = (3.86 + 4.05) = 7.91 / 2 = 3.955$. Similarly, we calculate 95th, 10th and 5th percentiles below:

$a_{95} = (x_{(19)} + x_{(20)}) / 2 = (4.05 + 4.12) = 8.17 / 2 = 4.085$

$a_{10} = (x_{(2)} + x_{(3)}) / 2 = (2.90 + 2.94) = 5.84 / 2 = 2.92$

$a_5 = (x_{(1)} + x_{(2)}) / 2 = (2.71 + 2.90) = 5.61 / 2 = 2.805$.

The fifth part of the table gives some extreme values (5 smallest and 5 largest observations). The last part of the table includes some plots about the data. One of these plots is the stem-and-leaf plot which is very similat to the histogram. Another is the Box-plot which gives some information about the normality and skewness or the symmetry about the sample. If the Box-plot is equally divided then we can say that the data is symmetric around the mean. Notice that $\bar{x}_n = 3.5$ and $m = 3.58$ so that mean and the median are close to each other. That is the data is nearly symmetric around the mean. Finally, at the end of the table, there is a normal probability plot which gives some information about the normality. If the plots seem to be linear then we can say that the data come from a normal population. There are some hypotheses testing results (non-parametric test) that we can test whether the data come from a normal population or not (like the values of Kolmororov-Simirnov and Cramer-von Mises test statistics). Since we did not study the hypothesis testing problem yet, we deleted these output results from the table. In this part of the table, it is important to mention that if we observe linearity in the normal probability plot, we can say that the data come from a normal population.

**Some Properties of Estimators**:

The goal of statistic (or any other field of science) is to understand the real world. Understanding means that we want to get some information about the population unknowns (we called parameter or parameters). In order to that we perform experiments and collect the data and based on the observed values we give estimates. Of course, there is no restriction to give an estimate about the parameters. However a good estimate should satisfies some statistical properties. We want estimators with smaller variance. For example, if we use two estimators in order to estimate a parameter, we should use the one that has a smaller variance. If $T_1$ and $T_2$ are two estimators in order to estimate a parameter $\theta$ and if $Var(T_1) \le Var(T_2)$ we should use $T_1$ to estimate the parameter. Of course, we can propose as many estimators as to estimate a parameter. However, a "better" estimator should satisfy some statistical properties (unbiasedness, consistency, sufficiency, efficiency etc.) and if possible, we want to find the "best" estimator. Here, the meaning of "the best" changes according to the criteria.

**Unbiasedness**:

First of all, we should mention that for a parameter $\theta$, there might be more than one unbiased estimator.

**Definition**: An estimator $T$ is said to be unbiased for a parameter $\theta$ if

$$E(T) = \theta \quad \text{for all } \theta.$$

**Example**: Let $X_1, X_2 \ldots, X_n$ be any random sample from a population having a probability density function $f(x;\theta)$ such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. The sample mean and variance are unbiased estimators for the population mean and variance. The estimators are already defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 .$$

Note that

$$E(\bar{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{\mu + \mu + \ldots + \mu}{n} = \frac{n\mu}{n} = \mu$$

which implies that the sample mean ( $\bar{X}_n$ ) is an unbiased estimator for the population mean ( $\mu$ ). Since,

$$E(S_n^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(X_i^2) - nE(\bar{X}_n^2)\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n(\mu^2 + \sigma^2 / n)\right)$$

$$= \frac{1}{n-1}\left(n((\sigma^2 + \mu^2) - n\mu^2 - \sigma^2\right) = \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - n\mu^2 - \sigma^2\right) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

the sample variance ( $S_n^2$ ) is unbiased for the population variance ( $\sigma^2$ ). Note that the estimator

$$T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

is a biased estimator for $\sigma^2$ because

$$E(T_n) = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = E\left(\frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = \frac{n-1}{n}E\left(S_n^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

That is, the estimator $T_n$ is biased for the population variance $\sigma^2$ .

**Definition** (*The Bias*): Let $T$ be any estimator for a parameter $\theta$ . The difference $E(T)-\theta$ is called the bias of $T$ and denoted by $Bias_\theta(T)$ . That is, $Bias_\theta(T) = E(T)-\theta$ .

Note that if $Bias_\theta(T) = 0$ then the estimator $T$ is unbiased for $\theta$ .

In the above example, we showed that the estimator

$$T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

is biased for the population variance. The bias of the estimator is calculated as

$$Bias_{\sigma^2}(T_n) = E(T_n) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

**The Mean Squared Error**: Sometimes it is difficult to find the best estimator for a parameter $\theta$. We may consider a better estimator should be unbiased and have a smaller variance. If it is possible, among all unbiased estimators it there is one which has the smallest variance this is the best estimator. Such an estimator may not exist. In some cases, we may use a biased estimator which have a smaller variance. If an estimator has a smaller mean square error, this controls both the bias and the variance.

**Definition**: Consider an estimator $T$ for a parameter $\theta$. The mean squared error of $T$ is defined as

$$MSE_\theta(T) = E_\theta(T-\theta)^2$$

which is also known as the lost function.

Note that

$$\begin{aligned} MSE_\theta(T) &= E_\theta(T-\theta)^2 = E_\theta(T - E(T) + E(T) - \theta)^2 \\ &= E_\theta(T - E(T))^2 + (E(T)-\theta)^2 + 2(E(T)-\theta)\underbrace{[E(T)-E(T)]}_{=0} \\ &= E_\theta(T - E(T))^2 + (E(T)-\theta)^2 = Var_\theta(T) + Bias_\theta^2(T). \end{aligned}$$

That is, the mean squared error can also be written as

$$MSE_\theta(T) = E_\theta(T-\theta)^2 = Var_\theta(T) + Bias_\theta^2(T).$$

If an estimator $T$ is unbiased ( $Bias_\theta(T) = 0$ ), then the mean squared error is the variance of the estimator ( $MSE_\theta(T) = Var_\theta(T)$ ).

**Example**: Let $X_1, X_2 \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. We have shown that the sample mean $\bar{X}_n$ is unbiased for the population mean ($E(\bar{X}_n) = \mu$) and therefore the bias is zero. The variance of the sample mean $\sigma^2 / n$ and therefore, $MSE_{\mu}(\bar{X}_n) = Var(\bar{X}_n) = \sigma^2 / n$ . Now, consider the problem of estimating the population variance. Consider the following estimators

$$T_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad T_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad T_3 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 .$$

Here, the estimators $T_1$ is the sample variance, $T_2$ is the maximum likelihood (or method of moment) estimator and $T_3$ is the mean squared error estimator of $\sigma^2$ . Note that the estimator $T_1$ is the sample variance it is unbiased for ( $E(T_1) = \sigma^2$ ). Moreover, as we have discussed earlier $Var(T_1) = 2\sigma^4 / (n-1)$ and therefore,

$$MSE(T_1) = Var(T_1) = 2\sigma^4 / (n-1) .$$

Now let us calculate the variance and the mean squared error of the estimator $T_2$. Notice that

$$T_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n}S_n^2$$

and therefore,

$E(T_2) = \frac{n-1}{n}E(S_n^2) = \frac{n-1}{n}\sigma^2$ , $Bias_{\sigma^2}(T_2) = -\frac{\sigma^2}{n}$ and the variance of $T_2$ is

$$Var(T_2) = Var\left(\frac{n-1}{n}S_n^2\right) = \left(\frac{n-1}{n}\right)^2 Var\left(S_n^2\right) = \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2} .$$

Therefore, the mean squared error of $T_2$ is

$$MSE(T_2) = Var(T_2) + Bias_{\sigma^2}^2(T_2) = \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.$$

Similarly, we can calculate the variance and the mean squared error of the estimator $T_3$. Notice that

$$T_3 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n+1}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n+1}S_n^2$$

and therefore,

$E(T_3) = \frac{n-1}{n+1}E(S_n^2) = \frac{n-1}{n+1}\sigma^2$ , $Bias_{\sigma^2}(T_3) = -\frac{2\sigma^2}{n+1}$ and the variance of $T_3$ is

$$Var(T_3) = Var\left(\frac{n-1}{n+1}S_n^2\right) = \left(\frac{n-1}{n+1}\right)^2 Var\left(S_n^2\right) = \frac{(n-1)^2}{(n+1)^2}\frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{(n+1)^2}.$$

Therefore, the mean squared error of $T_3$ is

$$MSE(T_3) = Var(T_3) + Bias_{\sigma^2}^2(T_3) = \frac{2(n-1)\sigma^4}{(n+1)^2} + \frac{4\sigma^4}{(n+1)^2} = \frac{2\sigma^4}{(n+1)}.$$

The bias, variance and mean squared err of the estimators are summarized below. Compare the variances and the mean squared errors of these three estimators.

| Estimator | Mean | Bias | Variance | MSE |
|---|---|---|---|---|
| $T_1$ | $\sigma^2$ | 0 | $\frac{2\sigma^4}{n-1}$ | $\frac{2\sigma^4}{n-1}$ |
| $T_2$ | $\frac{n-1}{n}\sigma^2$ | $-\frac{\sigma^2}{n}$ | $\frac{2(n-1)\sigma^4}{n^2}$ | $\frac{(2n-1)\sigma^4}{n^2}$ |
| $T_3$ | $\frac{n-1}{n+1}\sigma^2$ | $-\frac{2\sigma^2}{n+1}$ | $\frac{2(n-1)\sigma^4}{(n+1)^2}$ | $\frac{2\sigma^4}{(n+1)}$ |

**Consistency**:

As we have already mention an estimator is a function of the sample. Since the sample depends on the sample size, the estimator is also a function of the sample size. We define the estimators in order to get some information about the population unknowns

| | |
|---|---|
| | Let $X_1, X_2 \ldots, X_n$ be a random sample with probability (or probability density) function $f(x;\theta)$ Let $T_n$ be any sequence of estimators. remember that $T_n$ converges to $\theta$ in probability if for every $\varepsilon > 0$<br>$\lim_{n\to\infty} P(\lvert T_n - \theta \rvert > \varepsilon) = 0$. |

**Definition**: Let $X_1, X_2 \ldots, X_n$ be a random sample with probability (or probability density) function $f(x;\theta)$ Let $T_n$ be any sequence of estimators. If the sequence of estimators $T_n$ converge to the parameter $\theta$ in probability ($T_n \xrightarrow{P} \theta$ as $n \to \infty$ ), then $T_n$ is *consistent* for $\theta$.

Let $X_1, X_2 \ldots, X_n$ be a random sample with probability (or probability density ) function $f(x;\theta)$ such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all $i$. Assume that $T_n$ be any sequence of estimators. From the Chebyshev's inequality we have

$$P(|T_n - \theta| > \varepsilon) \leq \frac{E(T_n - \theta)^2}{\varepsilon^2} = \frac{Var(T_n) + Bias_\theta^2(T_n)}{\varepsilon^2} \;.$$

Thus if $Var(T_n) \to 0$ and $Bias_\theta(T_n) \to \theta$ as $n \to \infty$ it is obvious that $P(|T_n - \theta| > \varepsilon) \to 0$ and therefore, the sequence estimators $T_n$ is consistent for the parameter $\theta$.

**Example**: Let $X_1, X_2 \ldots, X_n$ be a random sample with probability (or probability density) function $f(x;\theta)$ such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ . We have shown that $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \sigma^2 / n$. It is obvious that $\bar{X}_n$ is unbiased for $\mu$ and therefore $Bias_\theta(T_n) = 0$ and since $Var(\bar{X}_n) = \sigma^2 / n \to 0$ as $n \to \infty$ which implies that $\bar{X}_n \xrightarrow{P} \mu$ as $n \to \infty$. That is, the sequence of estimators $\bar{X}_n$ is consistent for the population mean $\mu$. Moreover, as we have shown before, $S_n^2 \xrightarrow{P} \sigma^2$ as $n \to \infty$ which implies that the sequence of estimators ($S_n^2$, the sample variance) is consistent for the population variance $\sigma^2$.

**Efficiency**:

For any parameter $\theta$ we may have many different unbiased estimators as well as any consistent estimators. Among those estimators, we may want to have more efficient estimators. If it is possible, we want the most efficient estimator. Actually, among the class of unbiased estimators, it there is one that the smallest variance this is the most efficient estimators. Such an estimator may not exist. In some cases, we may find such estimators but we are not going to discuss these types of estimators.

**Definition**: Let $X_1, X_2 \ldots, X_n$ be a random sample with probability (or probability density) function $f(x;\theta)$ and consider two estimators $T_1$ and $T_2$ to estimate the parameter $\theta$. If

$Var(T_1) \leq Var(T_2)$ for all $\theta$

then the estimator $T_1$ is said to be more efficient that $T_2$.

**Example**: Let $X_1, X_2 \ldots, X_n$ be a random sample from $Poisson(\theta)$ population. We know that the mean and variance for the Poisson distribution are the same. That is, $E(X) = \theta$ and

$Var(X) = \theta$ since the sample mean and sample variance are unbiased for the population mean and variance (population mean $E(X) = \mu = \theta$ and $Var(X) = \sigma^2 = \theta$) we have

$E(\bar{X}_n) = \theta$ and $E(S_n^2) = \theta$.

Moreover, for any fixed reel number $a$ a class of estimators $T_a = a\bar{X}_n + (1-a)S_n^2$ are all unbiased for the population parameter $\theta$. That is for a parameter we can find infinitely many unbiased estimators for the population parameter. The sample mean is unbiased for $\theta$ with the variance $\theta / n$ . That is, $Var(\bar{X}_n) = \theta / n$. It can also be shown that $Var(\bar{X}_n) \leq Var(S_n^2)$ (the calculation of $Var(S_n^2)$ is really difficult and the inequality can be shown from the Cramer-Rao's inequality that we are not going to cover here). That is, the sample mean is more efficient that the sample variance for the Poisson parameter $\theta$. Again, from Cramer-Rao's inequality, it can be shown that the sample mean is the most efficient estimator among all these unbiased estimators (This is known as the Uniformly Minimum variance Unbiased Estimator, the UMVUE).

**<u>Sufficiency</u>**:

Sufficiency property of the estimators are very important especially for statistical inference. In estimation, we usually look for the most efficient unbiased estimators. If we can find a sufficient estimator for a parameter, based on this sufficient estimator we can write and unbiased estimator for the parameter that we are interested in.

**<u>Definition</u>**: Let $X_1, X_2 \ldots, X_n$ be a random sample from a population with probability (or probability density) function $f(x;\theta)$ and $T$ be any estimator for the parameter $\theta$. If the conditional probability density function of $X_1, X_2 \ldots, X_n$ given $T = t$ then we say that the estimator $T$ is sufficient for $\theta$.

A sufficient estimator is the one that summarized all the information in the sample about the parameter. If and estimator $T$ is sufficient for a parameter $\theta$, then any statistical inference can be made based on the value of the sufficient estimator $T$.

Let $X_1, X_2 \ldots, X_n$ be a random sample from a population with a probability function $f(x,\theta)$ and $T$ be any estimator for $\theta$. As we know, the estimator is a function of the sample and it is a random variable. For discrete case, if the conditional probability

$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid T = t)$

does not depend on the parameter $\theta$ the $T$ is sufficient for $\theta$. Similar argument can be carried out for continuous case. For simplicity let us denote $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)'$ and the estimator as $T(\underset{\sim}{X})$. As it is understood from the definition a sufficient estimator is not unique.

Let $T$ be a sufficient estimator for $\theta$ then the conditional probability $P(\underset{\sim}{X} = \underset{\sim}{x} | T(\underset{\sim}{X}) = t)$ does not depend on the parameter $\theta$. On the other hand, since $\{\underset{\sim}{X} = \underset{\sim}{x}\} \subset \{T(\underset{\sim}{X}) = T(\underset{\sim}{x})\}$ the conditional probability $P\left(\underset{\sim}{X} = \underset{\sim}{x} | T = t\right)$ can be calculated as

$$P\left(\underset{\sim}{X} = \underset{\sim}{x} | T = t\right) = \frac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x}\ , T = t)}{P_\theta(T = t)} = \frac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T = t)} = \frac{p(\underset{\sim}{x};\theta)}{q(T(\underset{\sim}{x});\theta)}$$

and therefore if the ratio $p(\underset{\sim}{x};\theta) / q(T(\underset{\sim}{x});\theta)$ does not depend on the parameter the estimator $T$ is sufficient for $\theta$. Therefore we can state the following theorem without the proof.

**<u>Theorem</u>:** An estimator $T$ is sufficient for $\theta$ if and only if the ration $p(\underset{\sim}{x};\theta) / q(T(\underset{\sim}{x});\theta)$ does not depen on the parameter.

**<u>Example 1</u>**. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with the parameter $p$. let us try to check whether the estimator $T = X_1 + X_2 + \ldots + X_n$ is sufficient or not. Note that $T \sim Binom(n, p)$ with the probability function,

$$P_p(T = t) = \binom{n}{t} p^t (1-p)^{n-t}\ ,\ t = 0,1,2,\ldots,n\,.$$

Therefore if the ratio $p(\underset{\sim}{x}; p) / q(T(\underset{\sim}{x}); p)$ for $t = x_1 + x_2 + \ldots + x_n$ does not depen on the parameter $p$ then $T$ will be sufficient. Note that the ratio for $t = x_1 + x_2 + \ldots + x_n$ can be written as

$$\frac{p(\underset{\sim}{x}; p)}{q(T(\underset{\sim}{x}); p)} = \frac{P_p\left(\underset{\sim}{X} = \underset{\sim}{x}\ \right)}{P_p\left(T = t\right)} = \frac{\prod_{i=1}^{n} P_p(X_i = x_i)}{P_p\left(T = t\right)} = \frac{\prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

and which is independent from $\theta$ and therefore the estimator $T = X_1 + X_2 + \ldots + X_n$ is sufficient for $p$.

The following theorem is very important to find a sufficient estimatıor for a parameter. We state the theorem without the proof. The proof of the theorem can be found in any textbook retaled to estimation theory (e.g Casella and Berger, 2002).

**<u>Theorem</u>** *(Factorization Theorem, very important)* Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with probability or probability density function $f(x;\theta)$. The estimator $T$ is sufficient for $\theta$ if and only if there are functions $g(t;\theta)$ and $h(x)$ such that the joint probability or probability density function of $X_1, X_2, \ldots, X_n$ can be written as

$$f(\underset{\sim}{x};\theta) = g(T(\underset{\sim}{x});\theta)\,h(\underset{\sim}{x})\,.$$

Here $g$ is a function of $T(\underset{\sim}{x})$ and $\theta$ and $h$ is only a function of $\underset{\sim}{x}$ which does not depend on the parameter.

**Example**: a) Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution with the parameter $p$. In the previous example we showed that $T = X_1 + X_2 + \ldots + X_n$ is sufficient for $p$. By using the factorization theorem the joint probability distribution of $X$ 's can be written as

$$f(\underset{\sim}{x};p) = P_p\left(\underset{\sim}{X} = \underset{\sim}{x}\right) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} = p^{T(\underset{\sim}{x})}(1-p)^{n-T(\underset{\sim}{x})}$$

and therefore for the functions $g(T(\underset{\sim}{x});p) = p^{T(\underset{\sim}{x})}(1-p)^{n-T(\underset{\sim}{x})}$ and $h(\underset{\sim}{x}) = 1$ the joint probability function can be written as $f(\underset{\sim}{x};p) = g(T(\underset{\sim}{x});p)\,h(\underset{\sim}{x})$ and thus by the fuctorization theorem $T$ is sufficient for $p$.

**b)** let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu,\sigma^2)$ distribution. The joint probability density function of $X_1, X_2, \ldots, X_n$ can be written as for $\underset{\sim}{T}\left(\underset{\sim}{x}\right) = \left(\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} x_i^2\right)$,

$$f\left(\underset{\sim}{x};\mu,\sigma^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i\right)$$

$$= \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right)\exp\left[\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)\begin{pmatrix}\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i\end{pmatrix}\right]$$

$$= \left(\frac{1}{2\pi}\right)^{n/2}\left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right)\exp\left[\underset{\sim}{w}'\left(\mu,\sigma^2\right)\underset{\sim}{T}(\underset{\sim}{x})\right].$$

Therefore for the functions

$$h(\underset{\sim}{x}) = (1/2\pi)^{n/2} \text{ and } g(T(\underset{\sim}{x});\mu,\sigma^2) = \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right)\exp\left[\underset{\sim}{w}'\left(\mu,\sigma^2\right)\underset{\sim}{T}(\underset{\sim}{x})\right]$$

the joint probability density function can be written as

$$f(\underset{\sim}{x};\mu,\sigma^2)=g(\underset{\sim}{T}(\underset{\sim}{x});\mu,\sigma^2)\,h(\underset{\sim}{x})$$

and thus by the factorization theorem the bivariate estimator $\underset{\sim}{T}\left(\underset{\sim}{X}\right)=\left(\sum_{i=1}^{n}X_i,\sum_{i=1}^{n}X_i^2\right)$ is sufficient for $(\mu,\sigma^2)$.

**c)** Let $X_1,X_2,\ldots,X_n$ be a random sample from a uniform distribution with the parameter $\theta$. The probability density function of the uniform distribution is given by

$$f(x;\theta)=\begin{cases}1/\theta & , \quad 0<x<\theta\\ 0 & , \quad d.y.\end{cases}$$

This probability density function can also be written with the following indicator function

$$I_A(x)=\begin{cases}1 & , \quad x\in A\\ 0 & , \quad x\notin A\end{cases}$$

as $f(x;\theta)=\theta^{-1}I_{(0<x<\theta)}(x)$. Therefore the joint probability density function of the sample can be written for $x_{(n)}=\max\{x_1,x_2,...,x_n\}$ as

$$f(\underset{\sim}{x};\theta)=\prod_{i=1}^{n}f(x_i;\theta)=\frac{1}{\theta}I_{\{0<x_{(1)}<\theta\}}(x_1)\frac{1}{\theta}I_{\{0<x_{(2)}<\theta\}}(x_2)...\frac{1}{\theta}I_{\{0<x_{(n)}<\theta\}}(x_n)=\frac{1}{\theta^n}I_{\{0<x_{(n)}<\theta\}}(\underset{\sim}{x}).$$

By defining the functions $g(T(\underset{\sim}{x});\theta)=\theta^{-n}I_{\{0<x_{(n)}<\theta\}}(\underset{\sim}{x})$ and $h(\underset{\sim}{x})=1$ allows us to write the joint probability density function as $f(\underset{\sim}{x};\theta)=g(T(\underset{\sim}{x});\theta)\,h(\underset{\sim}{x})$ and thus by the factorization theorem $T(\underset{\sim}{X})=X_{(n)}$ is sufficient for $\theta$.

**d)** Let $X_1,X_2,\ldots,X_n$ be a random sample from a population with the following probability density function:

$$f(x;\theta)=\begin{cases}\theta x^{\theta-1} & , \quad 0<x<1\\ 0 & , \quad d.y.\end{cases}$$

Note that the joint probability density function of $X_1,X_2,\ldots,X_n$ can be written as

$$f(\underset{\sim}{x};\theta)=\prod_{i=1}^{n}f(x_i;\theta)=\theta^n\prod_{i=1}^{n}x_i^{\theta-1}=\theta^n\left(\prod_{i=1}^{n}x_i\right)^{\theta}\prod_{i=1}^{n}\frac{1}{x_i}=g\left(T(\underset{\sim}{x});\theta\right)h(\underset{\sim}{x})$$

and thefore by the factorization theorem the estimator $T(\underset{\sim}{X})=\prod_{i=1}^{n}X_i$ is sufficient for $\theta$. Here the functions $g$ and $h$ are defined as

$$g\left(T(\underset{\sim}{x});\theta\right)=\theta^n\left(\prod_{i=1}^{n}x_i\right)^{\theta}\quad\text{and}\quad h(\underset{\sim}{x})=\prod_{i=1}^{n}\frac{1}{x_i}.$$

The application of the factorization theorem does not requires the independence. That is, the theorem is valid for non-independent and identically distributed sample.

As it is noted before, the sufficient estimator is not unique. Any one-to-one function of a sufficient estimator is also sufficient. To show that, let $T$ be any sufficient estimator for a parameter $\theta$ and $r$ be any one-to-one function. then $T^* = r(T)$ is also sufficient for the same parameter $\theta$. Since $T$ is sufficient for $\theta$ by the factorization theorem the joint probability density function can be written as $f(\underset{\sim}{x};\theta = g(T(\underset{\sim}{x});\theta)\,h(\underset{\sim}{x})$ for specified function $g(t;\theta)$ and $h(x)$. Moreover, since $r$ is a one-to-one $T = r^{-1}(T^*)$ and therefore the joint probability density function of the sample can also be written as

$$\begin{aligned} f(\underset{\sim}{x};\theta) &= g(T(\underset{\sim}{x});\theta)\,h(\underset{\sim}{x} = g(r^{-1}(T^*(\underset{\sim}{x}));\theta)\ h(\underset{\sim}{x}) \\ &= (g \circ r^{-1})(T^*(\underset{\sim}{x});\theta)h(\underset{\sim}{x}) = g^*(T^*(\underset{\sim}{x});\theta)\,h(\underset{\sim}{x}) \end{aligned}$$

and thus by the factorization theorem $T^* = r(T)$ is also sufficient for $\theta$. According to this result, since $T = \sum_{i=1}^{n} X_i$ is sufficient for $p$ in the example (a), $\bar{X}_n$ is also sufficient for the same parameter. In the example (b) we showed that $\underset{\sim}{T}(\underset{\sim}{X}) = \left(\sum_{i=1}^{n} X_i,\ \sum_{i=1}^{n} X_i^2\right)$ is sufficient for $(\mu,\sigma^2)$ and for the same reason the estimator $\underset{\sim}{T}^*(\underset{\sim}{X}) = (\bar{X}_n, S_n^2)$ is also sufficent for the same parameter $(\mu,\sigma^2)$. In the example (d) since $T = \prod_{i=1}^{n} X_i$ is sufficient for $\theta$, the estimator $T^* = -\sum_{i=1}^{n} \ln(X_i)$ is also sufficient for the same parameter $\theta$.