LARGE SAMPLE THEORY

1. Asymptotic Expectation and Variance

1.1. **Asymptotic Expectation.** Let $\{X^n\} = X_1, \ldots, X_n, \ldots$ be a sequence of random variables and let $\{E\,X_n\} = E(X_1), \ldots, E(X_n), \ldots$ be the sequence of their expectations. Suppose

$$\lim_{n\to\infty} E(X_n) = \mu \tag{1}$$

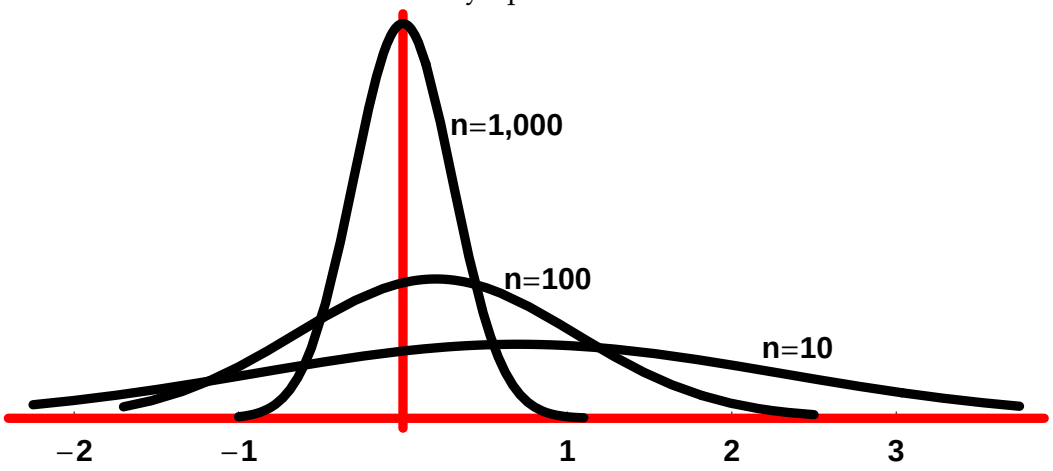
where μ is a finite constant. Then μ is said to be the asymptotic expectation of the sequence $\{X_n\}$ and we write $AE\,X_n = \mu$ or simply $AE\,X = \mu$. For example suppose $E\,X_n = \mu, \forall\, n$. The $AE\,X_n = \mu$. Or suppose that $E\,X_n = \mu + n^{-1}c_1 + n^{-1}c_2 + \cdots$, where the c's are finite constants. Then

$$AE\,X_n = \lim_{n\to\infty}\left(\mu + \frac{1}{n}c_1 + \frac{1}{n^2}\,c_2 + \cdots\right) = \mu$$

1.2. **Asymptotic Unbiasedness.** $\hat{\theta}$ is said to be an *asymptotically unbiased* estimator of θ if

$$AE\,\hat{\theta} = \theta$$

FIGURE 1. Asymptotic Unbiasedness



Note that unbiased estimators are asymptotically unbiased, but asymptotically unbiased estimators are not necessarily unbiased.

Date: February 18, 2008.

1.3. **An Example of Asymptotic Unbiasedness.**

1.3.1. *Sample Variance.* One estimate of the variance of a population is the sample variance S^2. The sample variance is given by

$$\begin{aligned} S^2(X_1, X_2, \cdots, X_n) &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \left\{ \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right\} \end{aligned} \tag{2}$$

Taking the expected value we obtain

$$\begin{aligned} E\left\{S^2(X_1, X_2, \cdots, X_n)\right\} &= \frac{1}{n-1} \left\{ E\left[\sum_{i=1}^{n} X_i^2 \right] - n\, E\left[\bar{X}_n^2 \right] \right\} \\ &= \frac{1}{n-1} \left\{ E\left[\sum_{i=1}^{n} X_i^2 \right] - n \left[\left(E\left[\bar{X}_n \right] \right)^2 + Var(\bar{X}_n) \right] \right\} \\ &= \frac{1}{n-1} \left\{ \sum_{i=1}^{n} E\, X_i^2 - n \left(E\left[\bar{X}_n \right] \right)^2 - n\, Var(\bar{X}_n) \right\} \end{aligned} \tag{3}$$

Now consider the relationship between the expected values in equation 3 and raw population moments.

$$E\left(X_i^2\right) = \mu'_{i,2} = \int_{-\infty}^{\infty} x_i^2 f_i(x_i)\, dx \tag{4}$$

and

$$E\left[\bar{X}_n^r\right] = \frac{1}{n} \sum_{i=1}^{n} E\, X_i^r = \frac{1}{n} \sum_{i=1}^{n} \mu'_{i,1} \tag{5}$$

Substituting equation 4 and equation 5 into equation 3 we obtain

$$\begin{aligned} E\left\{S^2(X_1, X_2, \cdots, X_n)\right\} &= \frac{1}{n-1} \left\{ \sum_{i=1}^{n} E\, X_i^2 - n \left(E\left[\bar{X}_n \right] \right)^2 - n\, Var(\bar{X}_n) \right\} \\ &= \frac{1}{n-1} \left\{ \sum_{i=1}^{n} \mu'_{i,2} - n \left(\frac{1}{n} \sum_{i=1}^{n} \mu'_{i,1} \right)^2 - n\, Var(\bar{X}_n) \right\} \end{aligned} \tag{6}$$

Given that we have a random sample,

$$
\begin{aligned}
E\left\{S^2(X_1, X_2, \cdots, X_n)\right\} &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} \mu'_{i,2} - n\left(\frac{1}{n}\sum_{i=1}^{n} \mu'_{i,1}\right)^2 - n\,Var(\bar{X}_n)\right\} \\
&= \frac{1}{n-1}\left\{n\,\mu'_2 - n\left(\mu'_1\right)^2 - n\frac{\sigma^2}{n}\right\} \\
&= \frac{1}{n-1}\left\{n\left(\mu'_2 - \left(\mu'_1\right)^2\right) - \sigma^2\right\} \\
&= \frac{1}{n-1}\left\{n\sigma^2 - \sigma^2\right\} \\
&= \frac{1}{n-1}\left\{(n-1)\sigma^2\right\} \\
&= \sigma^2
\end{aligned}
\tag{7}
$$

The estimator S^2 is clearly unbiased.

1.3.2. *Maximum Likelihood Estimate of Population Variance.* The maximum likelihood estimate of the variance of a population is $\hat{\sigma}^2$. The estimate is given by

$$
\begin{aligned}
\hat{\sigma}^2(X_1, X_2, \cdots, X_n) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right\}
\end{aligned}
\tag{8}
$$

Taking the expected value we obtain

$$
\begin{aligned}
E\left\{\hat{\sigma}^2(X_1, X_2, \cdots, X_n)\right\} &= \frac{1}{n}\left\{E\left[\sum_{i=1}^{n} X_i^2\right] - n\,E\left[\bar{X}_n^2\right]\right\} \\
&= \frac{1}{n}\left\{E\left[\sum_{i=1}^{n} X_i^2\right] - n\left[\left(E\left[\bar{X}_n\right]\right)^2 + Var(\bar{X}_n)\right]\right\} \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n} E\,X_i^2 - n\left(E\left[\bar{X}_n\right]\right)^2 - n\,Var(\bar{X}_n)\right\}
\end{aligned}
\tag{9}
$$

Substituting equation 4 and equation 5 into equation 9 we obtain

$$
\begin{aligned}
E\left\{\hat{\sigma}^2(X_1, X_2, \cdots, X_n)\right\} &= \frac{1}{n}\left\{\sum_{i=1}^{n} E\,X_i^2 - n\left(E\left[\bar{X}_n\right]\right)^2 - n\,Var(\bar{X}_n)\right\} \\
&= \frac{1}{n}\left\{\sum_{i=1}^{n} \mu'_{i,2} - n\left(\frac{1}{n}\sum_{i=1}^{n} \mu'_{i,1}\right)^2 - n\,Var(\bar{X}_n)\right\}
\end{aligned}
\tag{10}
$$

Given that we have a random sample,

$$
\begin{aligned}
E\left\{\hat{\sigma}^2(X_1, X_2, \cdots, X_n)\right\} &= \frac{1}{n}\left\{\sum_{i=1}^{n} \mu'_{i,2} - n\left(\frac{1}{n}\sum_{i=1}^{n} \mu'_{i,1}\right)^2 - n\,Var(\bar{X}_n)\right\} \\
&= \frac{1}{n}\left\{n\,\mu'_2 - n\left(\mu'_1\right)^2 - n\,\frac{\sigma^2}{n}\right\} \\
&= \frac{1}{n}\left\{n\left(\mu'_2 - \left(\mu'_1\right)^2\right) - \sigma^2\right\} \\
&= \frac{1}{n}\left\{n\sigma^2 - \sigma^2\right\} \\
&= \frac{1}{n}\left\{(n-1)\sigma^2\right\} \\
&= \frac{n-1}{n}\,\sigma^2
\end{aligned}
\tag{11}
$$

This is estimator is not unbiased. But as $n \to \infty$, the estimator converges to σ^2.

1.4. **Asymptotic Variance.** Let $\{X^{(n)}\} = X^{(1)}, X^{(2)}, \ldots, X^{(n)} \ldots$ be a sequence of random variables and let $\{EX^{(n)}\} = EX^{(1)}, EX^{(2)}, \ldots EX^{(n)} \ldots$ be the sequence of their expectations; and let $\{E(X^{(n)} - EX^{(n)})^2\} = E(X^{(1)} - EX^{(1)})^2, \ldots E(X^{(n)} - EX^{(n)})^2 \ldots$ be the sequence of their variances. Suppose the asymptotic expectation of the sequence exists, $AE\,X^{(n)} = AE\,X$. Suppose further that

$$\lim_{n\to\infty} E[\sqrt{n}(X^{(n)} - EX^{(n)})]^2 = v$$

where v is a finite constant. Then

$$\sigma^2 = \frac{v}{n}$$

is said to be the asymptotic variance of the sequence $\{X^{(n)}\}$ and we write

$$AE(X^{(n)} - EX^{(n)})^2 = \sigma^2$$

or simply

$$AE\big(X - AE(X)\big)^2 = \sigma^2$$

1.5. **Vector Generalizations.**

1.5.1. *Expectation.* Let $\{X^{(n)}\}$ be a sequence of random vectors and let $\{EX^{(n)}\}$ be the sequence of expectation vectors. Suppose

$$\lim_{n\to\infty} EX^{(n)} = \mu$$

(a vector of constants) then μ is asymptotic expectation of the sequence and is written

$$AE\,X^{(n)} = \mu$$

or

$$AE\,X = \mu$$

1.5.2. *Variance.* Let $\{X^{(n)}\}$ be a sequence of random vectors with expectation vectors $\{EX^{(n)}\}$ and let $\{E(X^{(n)} - EX^{(n)})(X^{(n)} - EX^{(n)})'\}$ be the sequence of their covariance matrices. Suppose that $AE\,X$ exists. Suppose further that

$$\lim_{n\to\infty} E[\sqrt{n}(X^{(n)} - EX^{(n)})][\sqrt{n}(X^{(n)} - EX^{(n)})]' = V$$

where V is a matrix of finite constants. Then $\sum = n^{-1}V$ is said to be the asymptotic covariance matrix of the sequence $X^{(n)}$ and we write

$$AE(X^{(n)} - EX^{(n)})(X^{(n)} - EX^{(n)})' = \Sigma$$

or simply

$$AE\{(X - AE(X))(X - AE(X))'\} = \Sigma$$

2. Convergence of Random Variables

2.1. **Chebyshev's Inequality.**

2.1.1. *General Form.* Let X be a random variable and let $g(x)$ be a non-negative function. Then for $r > 0$,

$$P\left[g(X) \geq r\right] \leq \frac{Eg(X)}{r} \tag{12}$$

Proof:

$$\begin{aligned} Eg(X) &= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \\ &\geq \int_{[x:g(x)\geq r]} g(x) f_X(x)\, dx \qquad \text{(g is non-negative)} \\ &\geq r \int_{[x:g(x)\geq r]} f_X(x)\, dx \qquad (g(x) \geq r) \\ &= rP\left[g(X) \geq r\right] \\ \Rightarrow P\left[g(X) \geq r\right] &\leq \frac{Eg(X)}{r} \end{aligned} \tag{13}$$

2.1.2. *Common Use Form.* Let X be a random variable with mean μ and variance σ^2. Then for any $\delta > 0$ or any $\varepsilon > 0$

$$\begin{aligned} P\big[|X-\mu| \geq \delta\sigma\big] &\leq \frac{1}{\delta^2} \\ P\big[|X-\mu| \geq \ \ \varepsilon\big] &\leq \frac{\sigma^2}{\varepsilon^2} \end{aligned} \tag{14}$$

Proof: Let $g(x) = (x-\mu)^2/\sigma^2$, where $\mu = E(X)$ and $\sigma^2 = \text{Var}(X)$. Then let $r = \delta^2$. Then

$$P\left[\frac{(X-\mu)^2}{\sigma^2} \geq \delta^2\right] \leq \frac{1}{\delta^2}\, E\left(\frac{X-\mu)^2}{\sigma^2}\right) = \frac{1}{\delta^2} \tag{15}$$

because $E(X-\mu)^2 = \sigma^2$. We can then rewrite equation 15 as follows

$$\begin{aligned} P\left[\frac{(X-\mu)^2}{\sigma^2} \geq \delta^2\right] &\leq \frac{1}{\delta^2} \\ \Rightarrow P\left[(X-\mu)^2 \geq \delta^2\sigma^2\right] &\leq \frac{1}{\delta^2} \\ \Rightarrow P\left[|X-\mu| \geq \delta\sigma\right] &\leq \frac{1}{\delta^2} \end{aligned} \tag{16}$$

2.2. **Convergence In Probability.** A sequence of random variables $\{\theta_n\} = (\theta_1, \theta_2, \ldots, \theta_n)$ is said to converge in probability to a random variable θ in probability if

$$\lim_{n\to\infty} Pr(|\theta_n - \theta| > \varepsilon) = 0 \tag{17}$$

for arbitrarily small $\varepsilon > 0$. This can be abbreviated as

$$\theta_n \xrightarrow{P} \theta \quad \text{or} \quad \operatorname{plim} \theta_n = \theta \tag{18}$$

We say that an estimator of a parameter θ is consistent if

$$\operatorname{plim} \hat{\theta}_n = \theta \tag{19}$$

and θ is the true parameter in question.

2.3. **Convergence In Mean Square Error.** A sequence $\{X_n\}$ is said to converge to X in mean square if

$$\lim_{n\to\infty} E(X_n - X)^2 = 0 \tag{20}$$

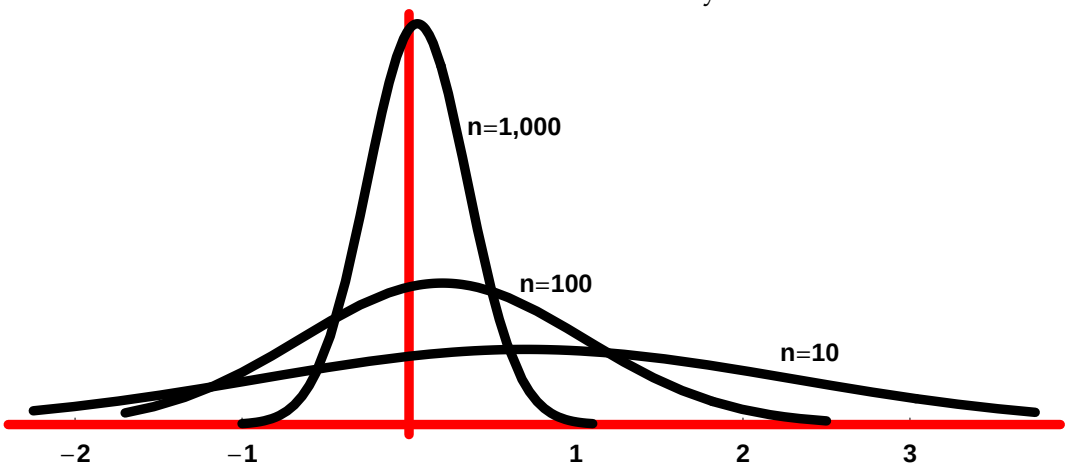
We write

$$X_n \xrightarrow{M} X \tag{21}$$

We say that an estimator is MSE consistent if

$$\hat{\theta}_n \xrightarrow{M} \theta \tag{22}$$

FIGURE 2. MSE Consistency



2.4. **Convergence In Distribution (Law).** A sequence $\{X_n\}$ is said to converge to X in distribution if the distribution function F_n of X_n converges to the distribution function F of X at every continuity point of F. We write

$$X_n \xrightarrow{d} X \tag{23}$$

and we call F the limit distribution of $\{X_n\}$. If $\{X_n\}$ and $\{Y_n\}$ have the same limit distribution we write

$$X_n \overset{LD}{=} Y_n \tag{24}$$

2.5. **Almost Sure Convergence.** A sequence of random variables $\{X_n\}$ is said to converge almost surely (certainly) to the random variable X if there exists a null set N (a set with probability zero) such that

$$\forall\, \omega\, \varepsilon\, \Omega \setminus N : \lim_{n\to\infty} X_n(\omega) = X(\omega) \tag{25}$$

Here ω is some sample point of the sample space Ω, $X(\omega)$ is a random variable defined at the sample point ω, and the set $\Omega \setminus N$ is the set Ω minus the set N. We can write this in a number of ways

$$\begin{aligned} P\left[\lim_{n\to\infty} X_n = X\right] &= 1 \\ \lim_{m\to\infty} P\left[|X_n - X| \le \varepsilon\, \forall n \ge m\right] &= 1 \\ \lim_{m\to\infty} P\left[\sup_{n\ge m}|X_n - X| > \varepsilon\right] &= 0 \\ P\left[\omega|\lim_{n\to\infty} X_n(\omega) = X(\omega)\right] &= 1 \end{aligned} \tag{26}$$

We write

$$X_n \xrightarrow{as} X \tag{27}$$

We say that an estimator $\hat{\theta}$ of θ is strongly consistent if

$$\hat{\theta}_n \xrightarrow{as} \theta \tag{28}$$

2.5.1. *Examples of almost sure convergence.*

Let the sample space S be the closed interval [0,1] with the uniform probability distribution. Define random variables $X_n(s) = s + s^n$ and $X(s) = s$. For every $s \in [0,1)$, $s^n \to 0$ as $n \to \infty$ and $X_n(s) \to X(s)$. However, $X_n(1) = 2$, for every n, so $X_n(1)$ does not converge to $1 = X(1)$. But because convergence occurs on the set [0,1) and $P([0,1)) = 1$, $X_n \to X$ almost surely.

Let the sample space S be the closed interval [0,1] with the uniform probability distribution. Define random variables $X_n(s)$ as follows:

$$X_n(s) = \begin{cases} n, & \text{if } 0 \le s \le \frac{1}{n}, \\ 0, & \text{if } \frac{1}{n} < s \le 1. \end{cases}$$

For this example $X_n(s) \xrightarrow{as} 0$. To see this let $N = \{0\}$. Then $s \in N^c$ implies $X_n(s) \to 0$. It is not true, however, that $X_n(s) \to 0$ for all $s \in [0,1]$, because $X_n(0) = n \to \infty$.

2.6. **Relationships Between Different Types Of Convergence.**

2.6.1. *Convergence Theorem 1.*

Theorem 1.

$$E\left(X_n^2\right) \to 0 \Rightarrow X_n \xrightarrow{P} 0 \tag{29}$$

Proof:

$$\begin{aligned} E\left(X_n^2\right) &= \int_{-\infty}^{\infty} x^2 \, dF_n(x) \\ &= \int_{x|x^2 \geq \varepsilon^2} x^2 \, dF_n(x) + \int_{x|x^2 < \varepsilon^2} x^2 \, dF_n(x) \\ &\geq \epsilon^2 \int_S dF_n(x) \end{aligned} \tag{30}$$

where $S = \{x|x^2 \geq \varepsilon^2\}$. But

$$\begin{aligned} \int_S dF_n(x) &= \int_{-\infty}^{-\varepsilon} dF_n(x) + \int_{\varepsilon}^{\infty} dF_n(x) \\ &= F_n(-\varepsilon) + [1 - F_n(\varepsilon)] \\ &= P\left(X_n < -\varepsilon\right) + P\left(X_n \geq \varepsilon\right) \\ &\geq P\left(X_n^2 > \varepsilon^2\right) \end{aligned} \tag{31}$$

where the last step follows because the inequality is strict. We can see this in figure 3.

Combining (30) and (31) will yield

$$P\left(X_n^2 > \varepsilon^2\right) \leq \frac{E\left(X_n^2\right)}{\varepsilon^2} \tag{32}$$

Thus if $E\, X_n^2$ goes to zero, X_n converges in probability to zero.

2.6.2. *Convergence Theorem 2.*

Theorem 2.

$$X_n \xrightarrow{M} X \Rightarrow X_n \xrightarrow{P} X \tag{33}$$

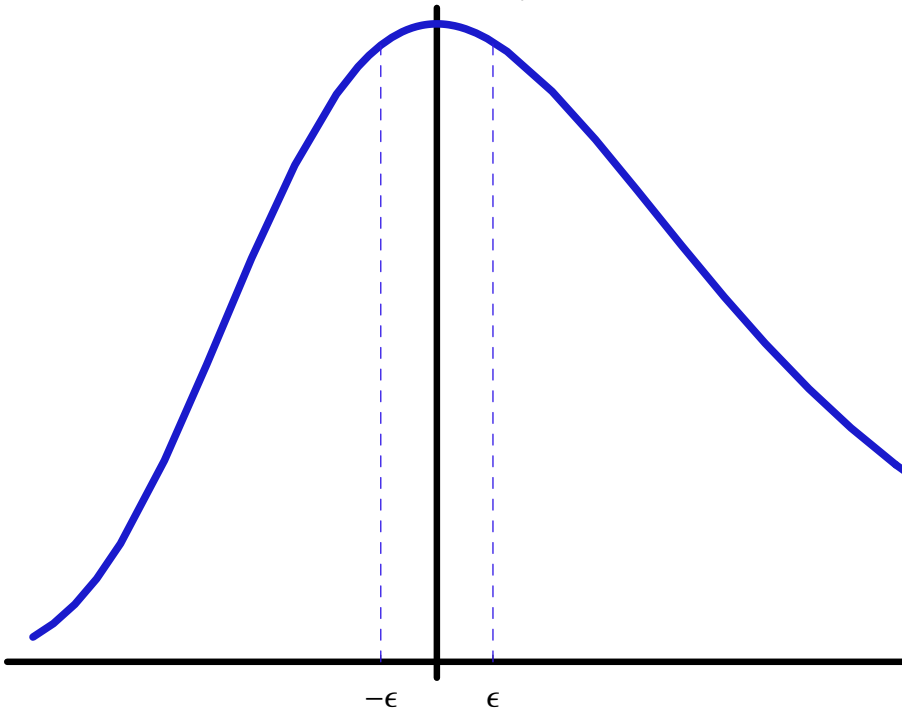
Proof: Let $X_n - X$ be X_n of Theorem 1.

2.6.3. *Convergence Theorem 3.*

Theorem 3. *Let* $\{X_n, Y_n\}$ $n = 1, 2, \ldots$ *be a sequence of pairs of variables. Then*

$$|X_n - Y_{n|} \xrightarrow{P} 0,\ Y_n \xrightarrow{d} Y \Rightarrow X_n \xrightarrow{d} Y \tag{34}$$

FIGURE 3. Probability for $X_n < -\varepsilon$ and $X_n > \varepsilon$



Proof: Let F_n be the distribution function of X and G_n the distribution function of Y. Now define the following sets

$$\begin{aligned} A_n &= [\omega : |X_n(\omega) - Y_n(\omega)| < \varepsilon] \\ B_n &= [\omega : \ X_n(\omega) < x, \, x \epsilon R] \\ C_n &= [\omega : \ Y_n(\omega) < x + \varepsilon] \\ D_n &= [\omega : \ Y_n(\omega) \geq x - \varepsilon] \end{aligned}$$

Clearly

$$\begin{aligned} F_n(x) = P(B_n) &= P(B_n \cap A_n) + P(B_n \cap A_n^c) \\ 1 - F_n(x) = P(B_n^c) &= P(B_n^c \cap A_n) + P(B_n^c \cap A_n^c) \end{aligned}$$

and

$$\begin{aligned} B_n \cap A_n &= [\omega : \ X_n(\omega) < x, \, X_n(\omega) - \varepsilon < Y_n(\omega) < X_n(\omega) + \varepsilon] \subset C_n \\ B_n^c \cap A_n &= [\omega : \ X_n(\omega) \geq x, \, X_n(\omega) - \varepsilon < Y_n(\omega) < X_n(\omega) + \varepsilon] \subset D_n \end{aligned}$$

Now use this information to obtain

$$\begin{aligned} F_n(x) &\leq P(C_n) + P(A_n^c) = G_n(x+\varepsilon) + P(A_n^c) \\ \Rightarrow F_n(x) &\leq G_n(x+\varepsilon) + P(A_n^c) \\ 1 - F_n(x) &\leq P(D_n) + P(A_n^c) = 1 - G_n(x-\varepsilon) + P(A_n^c) \\ \Rightarrow 1 - F_n(x) &\leq 1 - G_n(x-\varepsilon) + P(A_n^c) \\ \Rightarrow G_n(x-\varepsilon) - P(A_n^c) &\leq F_n(x) \\ \Rightarrow G_n(x-\varepsilon) - P(A_n^c) \leq F_n(x) &\leq G_n(x+\varepsilon) + P(A_n^c) \end{aligned}$$

The assumption is that

$$\lim_{n\to\infty} P(A_n^c) = 0$$

and that x is a continuity point of F so that the limits exist and

$$G_n(x-\varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq G_n(x+\varepsilon)$$

Since ε is arbitrary this means that

$$\lim_{n\to\infty} F_n(x) = F(x) = G(x)$$

for all continuity points of F. Thus X_n converges in distribution to Y.

2.6.4. *Corollary To Theorem 3.*

$$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X \tag{35}$$

2.6.5. *Theorem 4.*

Theorem 4.

$$X_n \xrightarrow{as} X \Rightarrow X_n \xrightarrow{P} X \tag{36}$$

Proof: First note that since

$$|X_m - X| > \varepsilon \Rightarrow \sup_{n\geq m} |X_n - X| > \varepsilon$$

then

$$P\left[|X_m - X| > \varepsilon\right] \leq P\left[\sup_{n>m} |X_n - X| > \varepsilon\right] \tag{37}$$

Since by assumption the limit of the rhs is zero, the result follows.

2.6.6. *Convergence in probability does not imply almost sure convergence.* We can show this by an example. Let the sample space S be the closed interval [0,1] with the uniform probability distribution. Define the following set of intervals.

$$A_n^i = \left[\frac{i-1}{n}, \frac{i}{n}\right], i = 1, 2, \ldots, n; n \geq 1.$$

This gives intervals of the following form.

$$
\begin{array}{cccccc}
[0,1] & & & & & \\
\left[0,\frac{1}{2}\right] & \left[\frac{1}{2},1\right] & & & & \\
\left[0,\frac{1}{3}\right] & \left[\frac{1}{3},\frac{2}{3}\right] & \left[\frac{2}{3},1\right] & & & \\
\left[0,\frac{1}{4}\right] & \left[\frac{1}{4},\frac{2}{4}\right] & \left[\frac{2}{4},\frac{3}{4}\right] & \left[\frac{3}{4},1\right] & & \\
\left[0,\frac{1}{5}\right] & \left[\frac{1}{5},\frac{2}{5}\right] & \left[\frac{2}{5},\frac{3}{5}\right] & \left[\frac{3}{5},\frac{4}{5}\right] & \left[\frac{4}{5},1\right] & \\
\vdots & \vdots & \vdots & \vdots & &
\end{array}
$$

For example $A_3^2 = \left[\frac{1}{3}, \frac{2}{3}\right]$ and $A_5^3 = \left[\frac{2}{5}, \frac{3}{5}\right]$. Now define an indicator random variable $\xi_n^i = I_{A_n^i}(\omega)$ which takes the value 1 if ω is in the interval A_n^i and 0 otherwise. For example, $\xi_n^i = 1$ if ω is between 0 and 1. So the random variable ξ_3^2 takes the value one if ω is $\in A_3^2$. Now construct a sequence of random variables as follows

$$
\begin{array}{llllll}
X_1(\omega) = \xi_1^1 = I_{[0,1]}(\omega) & & & & & \\
X_2(\omega) = \xi_2^1 = I_{\left[0,\frac{1}{2}\right]}(\omega), & X_3(\omega) = \xi_2^2 = I_{\left[\frac{1}{2},1\right]}(\omega) & & & & \\
X_4(\omega) = \xi_3^1 = I_{\left[0,\frac{1}{3}\right]}(\omega) & X_5(\omega) = \xi_3^2 = I_{\left[\frac{1}{3},\frac{2}{3}\right]}(\omega) & X_6(\omega) = \xi_3^3 = I_{\left[\frac{2}{3},1\right]}(\omega) & & & \\
X_7(\omega) = \xi_4^1 = I_{\left[0,\frac{1}{4}\right]}(\omega) & X_8(\omega) = \xi_4^2 = I_{\left[\frac{1}{4},\frac{2}{4}\right]}(\omega) & \ldots & & &
\end{array}
$$

As n $\to \infty$,

$$\lim_{n\to\infty} Pr(|X_n(\omega) - 0| > \varepsilon) = 0$$

for arbitrarily small $\varepsilon > 0$ because the length of the intervals is going to zero. But for any $\omega \in$ [0,1],$X_n(\omega)$ does not converge almost surely to zero because $X_n(\omega) = 1$ for infinitely many values of n. Alternatively if we defined $Z_n(\omega) = \omega + X_n(\omega)$, for every value of ω, the value $Z_n(\omega)$ alternates between ω and ω +1 infinitely often.

2.6.7. *Convergence of subsequences of random variables.* If $X_n(\omega) \xrightarrow{P} X$, then there exists a subsequence of $X_{n_j}(\omega)$ such that $X_{n_j}(\omega) \xrightarrow{as} X$.

Convergence in probability allows more erratic behavior in the converging sequence than almost sure convergence, and by simply disregarding the erratic elements of the sequence,we can obtain an almost surely convergent subsequence.

2.6.8. *Theorem 5.*

Theorem 5. *Consider a constant k, then*

$$X_n \xrightarrow{d} k \Rightarrow X_n \xrightarrow{P} k \tag{38}$$

Proof: Since X_n converges in distribution to a constant, the distribution function will be degenerate, i.e.,

$$F(x) = \begin{cases} 0 & x < k \\ 1 & x \geq k \end{cases}$$

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0 & x < k \\ 1 & x \geq k \end{cases}$$

Now write convergence in probability in terms of distribution functions as follows

$$\begin{aligned} P\left[|X_n - k| < \varepsilon\right] &= P\left[k - \varepsilon < X_n < k + \varepsilon\right] \\ &= P\left[k - \varepsilon < X_n \leq k + \varepsilon\right] - P\left[X_n = k + \varepsilon\right] \\ &= P\left[X_n \leq k + \varepsilon\right] - P\left[X_n \leq k - \varepsilon\right] - P\left[X_n = k + \varepsilon\right] \\ &= F_n(k + \varepsilon) - F_n(k - \varepsilon) - P\left[X_n = k + \varepsilon\right] \end{aligned}$$

Now take the limit on both sides to obtain

$$\begin{aligned} \lim_{n\to\infty} P\left[|X_n - k| < \varepsilon\right] &= \lim_{n\to\infty} F_n(k + \varepsilon) - \lim_{n\to\infty} F_n(k - \varepsilon) - \lim_{n\to\infty} P\left[X_n = k + \varepsilon\right] \\ &= 1 - 0 - 0 \\ &= 1 \end{aligned}$$

2.6.9. *Convergence in distribution does not imply convergence in probability.*

Consider the following example. Let D be a N(0,1) random varaible for that the distribution function is symmetric. Define for n ≥ 1

$$X_n \;=\; (-1)^n D.$$

Then $X_n \xrightarrow{d}$ D. But of course, $\{X_n\}$ neigher converges almost surely nor in probability.

2.6.10. *Theorem 6 (Helly-Bray).*

Theorem 6. *$F_n \to F \Rightarrow \int g\, dF_n \to \int g\, dF$ for every bounded continuous function g.*

Proof: Rao [8, p. 117].

2.6.11. *Theorem 7.*

Theorem 7. *Let $X_1, \ldots X_n$ be a sequence of random variables and let $\phi_1(t)$, $\phi_2(t)$ ... be the sequence of characteristic functions of $X_1 \ldots X_n$. Suppose $\phi_n(t)$ converges to $\phi(t)$ as $n \to \infty \,\forall\, t$, then the sequence $X_1, \ldots X_n$ converges in distribution to the limiting distribution whose characteristic function is $\phi(t)$ provided that $\phi(t)$ is continuous at $t = 0$.*

Proof: Rao [8, p. 119].

2.7. **Examples.**

2.7.1. *Convergence in Distribution.* Let $\{X_1, \ldots X_n\}$ be a sequence of random variables having an exponential distribution with $\lambda = 1 + 1/n$.

$$F_n(x) = 1 - e^{-\left(1+\frac{1}{n}\right)x} \quad x > 0$$
$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \left[1 - e^{-\left(1+\frac{1}{n}\right)x}\right] = 1 - e^{-x} \tag{39}$$

which is an exponential distribution with $\lambda = 1$.

2.7.2. *Convergence in Probability.* Consider the sample mean given by

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{40}$$

We want to show that this has expected value equal to the sample mean μ, and converges in limit to the sample mean μ. Remember that each x_i has mean μ. First find the expected value and variance of $\bar{X}_n$.

$$\begin{aligned} E\left(\bar{X}_n\right) =& E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu \\ \operatorname{Var}\left(\bar{X}_n\right) &= \operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(x_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \end{aligned} \tag{41}$$

Now apply Chebyshev's inequality using $\bar{X}_n$ with its mean μ and variance σ^2/n.

$$P\left[|\bar{X}_n - \mu| \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2} \tag{42}$$

Now use the definition of convergence in probability on the above equation

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| > \varepsilon\right) = 0$$

since

$$\lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0 \tag{43}$$

2.8. **Functions of Random Variables.**

2.8.1. *Theorem 8 (Mann and Wald).*

Theorem 8. *If g is a continuous function and*

$$X_n \xrightarrow{d} X \quad \textit{then} \quad g(X_n) \xrightarrow{d} g(X)$$

Proof: Use the characteristic function as follows

$$\begin{aligned} E\left[\mathrm{e}^{\mathrm{i}tg(X_n)}\right] &= \int \mathrm{e}^{\mathrm{i}tg(x)}\, dF_n(x) \\ &= \int \cos\big(tg(x)\big)\, dF_n(x) + \mathrm{i}\int \sin\big(tg(x)\big)\, dF_n(x) \end{aligned} \tag{44}$$

Now since $\cos\big(tg(x)\big)$ and $\sin\big(tg(x)\big)$ are bounded continuous functions then Theorem 6 implies that (44) converges to

$$\int \cos\big(tg(x)\big)\, dF(x) + \mathrm{i}\int \sin\big(tg(x)\big)\, dF(x) = E[\mathrm{e}^{\mathrm{i}tg(x)}] \tag{45}$$

Combining Theorems 6 and 7 on convergence then implies the result.

2.8.2. *Theorem 9.*

Theorem 9. *If g is a continuous function and*

$$X_n \xrightarrow{P} X \quad \textit{then} \quad g(X_n) \xrightarrow{P} g(X)$$

Proof: Let I be a finite interval such that $P(X \epsilon I) = 1 - \eta/2$ and $n > m$, such that $P[|X_n - X| < \delta] > 1 - \eta/2$. Now by the continuity of g we have $|g(X_n) - g(X)| < \varepsilon$, if $|X_n - X| < \delta$ for any X in I. The point is that we have defined the interval such that it contains points where X_n being close to X means $g(X_n)$ is close to $g(X)$. Now define the following sets

$$\begin{aligned} E_1 &= [X : X \epsilon I] \\ E_2 &= [X_n : |X_n - X| < \delta] \\ E_3 &= [X_n : |g(X_n) - g(X)| < \varepsilon] \end{aligned}$$

Because $|g(X_n) - g(X)| < \varepsilon$, if $|X_n - X| < \delta$ for any X in I, it is clear that

$$\begin{aligned} E_3 = [X_n : |g(X_n) - g(X)| < \varepsilon] &\supset E_1 \cap E_2 = [X : |X_n - X| < \delta,\ X \epsilon I] \\ \Rightarrow \quad P\left[|g(X_n) - g(X)| < \varepsilon\right] &\geq P\left[|X_n - X| < \delta,\ X \epsilon I\right] \end{aligned}$$

Also

$$\begin{aligned} P(E_2) &= P(E_2 \cap E_1) + P(E_2 \cap E_1^c) \\ &\leq P(E_2 \cap E_1) + P(E_1^c) \\ \Rightarrow \quad P(E_2 \cap E_1) &\geq P(E_2) - P(E_1^c) \\ &= P(E_2) - P(X \notin I) \end{aligned}$$

We then have

$$\begin{aligned} P\left[|g(X_n) - g(X)| < \varepsilon\right] &\geq P\left[|X_n - X| < \delta,\ X \epsilon I\right] \\ &\geq P\left[|X_n - X| < \delta\right] - P(X \notin I)\, \forall\, n \geq m \end{aligned} \tag{46}$$

Combining all the information we have

$$\begin{aligned} & P[|X_n - X| < \delta] > 1 - \eta/2 \\ & P[X \notin I] \geq \eta/2 \\ \Rightarrow \quad & P\left[|X_n - X| < \delta\right] - P(X \notin I) \geq 1 - \eta \\ \Rightarrow \quad & P\left[|g(X_n) - g(X)| < \varepsilon\right] \geq 1 - \eta \, \forall \, n \geq m \end{aligned} \tag{47}$$

Taking the limit of both sides gives the result.

2.8.3. *Theorem 10.*

Theorem 10. *Let X_n be a vector of random variables with a fixed finite number of elements. Let g be a real-valued function continuous at a constant vector point α. Then*

$$\begin{aligned} X_n \xrightarrow{P} \alpha \Rightarrow g(X_n) \xrightarrow{P} g(\alpha) \\ \operatorname{plim} X_n = \alpha \Rightarrow \operatorname{plim} g(X_n) = g(\alpha) \end{aligned} \tag{48}$$

Proof: Continuity at α means that for any $\varepsilon > 0$, we can find a δ such that $||X_n - \alpha|| < \delta$ implies that $|g(X_n) - g(\alpha)| < \varepsilon$. Therefore

$$P\left[||X_n - \alpha|| < \delta\right] \leq P\left[|g(X_n) - g(\alpha)| < \varepsilon\right] \tag{49}$$

Because the lhs will converge to one by assumption, the result follows.

2.8.4. *Theorem 11.*

Theorem 11.
Let $\{ X_n, Y_n \}$, be a sequence of pairs of random variables. If

$$X_n \xrightarrow{d} X \quad \textit{and} \quad Y_n \xrightarrow{P} 0$$

then

$$X_n Y_n \xrightarrow{P} 0$$

Proof: Rao [8, p. 122].

2.8.5. *Theorem 12 (Slutsky).*

Theorem 12.

Let $\{ X_n, Y_n \}$, be a sequence of pairs of random variables. If

$$X_n \xrightarrow{d} X \quad \textit{and} \quad Y_n \xrightarrow{P} k$$

then

$$\begin{aligned} \mathbf{1}: \quad & X_n + Y_n \xrightarrow{d} X + k \\ \mathbf{2}: \quad & X_n Y_n \xrightarrow{d} kX \\ \mathbf{3}: \quad & \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{k}, \; k \neq 0 \end{aligned} \tag{50}$$

Proof: Rao [8, p. 123].

2.8.6. *Theorem 13.*

Theorem 13. *Suppose $g_n(\theta)$ converges in probability to a non-stochastic function $g(\theta)$ uniformly in θ in an open neighborhood $N(\theta_0)$ of θ_0. Then*

$$\operatorname{plim} g_n\left(\hat{\theta}_n\right) = g\left(\theta_0\right)$$

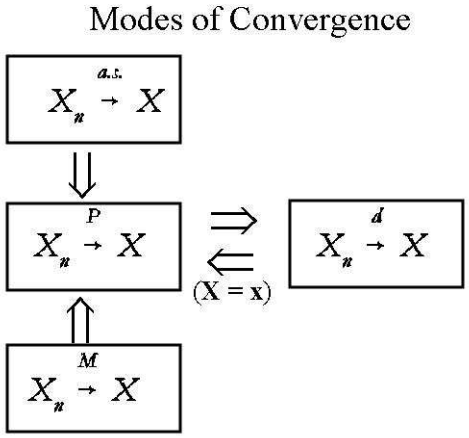
if $\operatorname{plim}\left(\hat{\theta}_n\right) = \theta_0$ *and $g(\theta)$ is continuous at θ.*

Proof: Amemiya [1, p. 113]. The important fact about Theorem 13 as compared to the previous ones is that g is allowed to depend on n.

2.9. **Some Notes.**

(i) Unbiasedness does not imply consistency.
(ii) Consistency does not imply unbiasedness nor asymptotic unbiasedness.
(iii) MSE consistent implies asymptotic unbiasedness, but not unbiasedness.
(iv) $E\left(\hat{\theta}\right) = \theta$ does not imply $E\left(g(\hat{\theta})\right) = g(\theta)$.
(v) An estimator which converges in probability can be asymptotically biased (mean not defined) but consistent. An estimator which converges in mean square is asymptotically unbiased and has a variance which approaches zero as the sample size approaches infinity. If $E(\hat{\theta})$ and $E(\hat{\theta}^2)$ exist, then both definitions will be equivalent.

FIGURE 4



3. Laws Of Large Numbers

3.1. **Introduction.** Laws of large numbers have to do with sequences of random variables X_n and the behavior of

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$$

In particular they relate to how well such statistics approximate the moments of the distribution.

3.2. Khintchine's Theorem (Weak Law of Large Numbers).

Theorem 14. *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $E(X_i) = \mu < \infty$. Then*

$$\frac{1}{n}\sum_{t=1}^{n} X_t = \bar{X}_n \xrightarrow{P} \mu \tag{51}$$

Proof: Rao [8, p. 113]. As an example consider the error terms in the linear regression model.

$$\begin{aligned} \varepsilon_t &\sim iid(0, \sigma^2) \\ E(\varepsilon_t^2) &= \sigma^2 \\ \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t^2 = \frac{\varepsilon\acute{\varepsilon}}{n} &\xrightarrow{P} \sigma^2 \end{aligned} \tag{52}$$

3.3. Kolmogorov Law of Large Numbers Theorem 1 (Strong Law of Large Numbers).

Theorem 15. *Let $\{X_i\}$, $i = 1, 2, \ldots$ be a sequence of independent random variables such that $E(X_i) = \mu_i$ and*

$$\mathrm{Var}(X_i) = \sigma_i^2$$

If

$$\sum_{t=1}^{\infty} \frac{\sigma_t^2}{t^2} < \infty \quad \textit{then} \quad \bar{X}_n - E(\bar{X}_n) \xrightarrow{a.s.} 0 \tag{53}$$

Proof: Rao [8, p. 114].

3.4. Kolmogorov Law of Large Numbers Theorem 2.

Theorem 16. *Let $\{X_i\}$, $i = 1, 2, \ldots$ be a sequence of independent and identically distributed random variables. Then a necessary and sufficient condition that*

$$\bar{X}_n \xrightarrow{as} \mu$$

is that $E(X_i)$ exists and is equal to μ.

Proof: Rao [8, p. 115].

4. CENTRAL LIMIT THEOREMS

4.1. **Central Limit Theorem (Lindberg-Levy).**

Theorem 17. *Let $X_1,\ X_2,\ \ldots\ X_n$ be a sequence of independent and identically distributed random variables with finite mean μ and finite variance σ^2. Then the random variable*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2)$$
$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} \xrightarrow{d} N(0,1) \tag{54}$$

We sometimes say that if

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{55}$$

then asymptotically

$$\bar{x}_n \sim N\left[\mu,\ \frac{\sigma^2}{n}\right] \quad \textit{or} \quad \bar{x}_n \xrightarrow{a} N\left[\mu,\ \frac{\sigma^2}{n}\right] \tag{56}$$

In general for a vector of parameters θ with finite mean vector μ and covariance matrix Σ, the following holds

$$\sqrt{n}(\bar{\theta} - \mu) \xrightarrow{d} N(0,\ \Sigma)$$
$$\bar{\theta} \xrightarrow{a} N\left[\mu,\ \frac{1}{n}\Sigma\right] \tag{57}$$

We say that $\bar{\theta}$ is asymptotically normally distributed with mean vector μ and covariance matrix $(1/n)\Sigma$.

Proof: Rao [8, p. 127] or Theil [11, pp. 368-9].

4.2. **Central Limit Theorem (Lindberg-Feller).**

Theorem 18. *Let $X_1,\ X_2,\ \ldots\ X_n$ be a sequence of independent random variables. Suppose for every t, X_t has finite mean μ_t and finite variance σ_t^2. Let F_t be the distribution function of X_t. Define C_n as follows:*

$$C_n = \left(\sum_{t=1}^{n} \sigma_t^2\right)^{\frac{1}{2}} \tag{58}$$

If

$$\lim_{n\to\infty} \quad \frac{1}{C_n^2} \sum_{t=1}^{n} \int_{|\omega - \mu_t| > \varepsilon C_n} (\omega - \mu_t)^2\, dF_t(\omega) = 0 \,\forall\, \varepsilon > 0$$

then

$$Z_n = \sum_{t=1}^{n} \frac{x_t - \mu_t}{C_n} \sim N(0,1) \tag{59}$$

where ω is a variable of integration.

Proof: Billingsley [3, Theorem 27.2 p. 359-61].

4.3. **Multivariate Central Limit Theorem** (see Rao [8, p. 128]).

Theorem 19. *Let $X_1,\ X_2,\ \ldots$ be a sequence of k dimensional vector random variables. Then a necessary and sufficient condition for the sequence to converge to a distribution F is that the sequence of scalar random variables $\lambda' X_t$ converge in distribution for any k-vector λ.*

Corollary 1. Let $X_1,\ X_2,\ \ldots$ be a sequence of vector random variables. Then if the sequence $\lambda' X_t$ converges in distribution to $N(0,\ \lambda'\Sigma\lambda)$ for every λ, the sequence X_t converges in distribution to $N(0,\ \Sigma)$.

5. Sufficient Statistics

5.1. **Definition of Sufficient Statistic.** Let $X_1,\ X_2,\ \ldots\ , X_n$ be a random sample from the density $f(\cdot : \theta)$, where θ may be a vector. A statistic $T(X) = t(X_1,\ \ldots\ ,\ X_n)$ is defined to be a sufficient statistic if the conditional distribution of $X_1,\ \ldots\ , X_n$ given $T(x) = t$ does not depend on θ for any value t of T. The idea of a sufficient statistic is that it condenses n random variables into one. If no information s lost in this process, the statistic effectively explains the data. The idea is that if we know the value f the sufficient statistic, we do not need to know θ to get the conditional distribution of the X's.

5.2. **Definition of Jointly Sufficient Statistics.** Let $X_1,\ X_2,\ \ldots\ ,\ X_n$ be a random sample from the density $f(\cdot\ :\ \theta)$. The statistics $T_1(X),\ \ldots\ T_r(X)$ are defined to jointly sufficient if the conditional distribution of $X_1,\ \ldots\ ,\ X_n$ given $T_1(x) = t_1,\ \ldots\ ,\ T_r(X) = t_r$ does not depend on θ for any values t.

5.3. **Factorization Theorem (Single Sufficient Statistic) (Neyman-Fisher).**

Theorem 20. *Let $X_1,\ X_2,\ \ldots\ , X_n$ be a random sample from the density $f(\cdot\ :\ \theta)$, where θ may be a vector. A statistic $T(X) = t(X_1,\ \ldots\ , X_n)$ is sufficient if the joint distribution of $X_1,\ \ldots\ X_n$ can be factored as*

$$f_{X_1,\ldots,X_n}(x_1,\ \ldots\ , x_n;\ \theta) = g(t(x),\ \theta)h(x) \tag{60}$$

where g depends on X only through the value of $T(X)$ and h does not depend on θ and g and h are both non-negative. In many cases $h(x)$ may simply be a constant not depending on x, or just the identity $h(x) = 1$.

5.4. **Factorization Theorem (Jointly Sufficient Statistics).**

Theorem 21. *Let $X_1,\ X_2,\ \ldots\ , X_n$ be a random sample from the density $f(\cdot : \theta)$, where θ may be a vector. A set of statistics $T_1(X) = t_1(X_1,\ \ldots\ ,\ X_n),\ \ldots\ , \ldots, T_r(X_1,\ \ldots\ ,\ X_n) = t_r(X_1,\ \ldots\ ,\ X_n)$ is jointly sufficient if the joint distribution of $X_1,\ \ldots\ ,\ X_n$ can be factored as*

$$f_{X_1,\ldots,X_n}(x_1,\ \ldots\ , x_n;\theta) = g(t_1(x),\ldots,t_r(x);\theta)h(x) \tag{61}$$

where g depends on X only through the value of $T_1(X),\ \ldots\ ,\ T_r(X)$ and h does not depend on θ and g and h are both non-negative.

5.5. **Theorem on Efficiency (Rao-Blackwell).**

Theorem 22. *Let $X_1,\ X_2\ \dots\ , X_n$ be a random sample from the density $f(\cdot : \theta)$, where θ may be a vector. Let $T_1(X) = t_1(X_1, \dots , X_n), \dots , T_r = t_r(X_1, \dots , X_n)$ be a set of jointly sufficient statistics. Let the statistic $T = t(X_1, \dots , X_n)$ be an unbiased estimator of some function $\tau(\theta)$.*

Define $\hat{T}$ as $E(T|T_1, \dots , T_r)$. Then the following are true

$$\begin{gathered} \hat{T} \quad \textit{is a statistic a function of the sufficient statistics} \\ E(\hat{T}) = \tau(\theta) \\ \mathrm{Var}(\hat{T}) \leq \mathrm{Var}(T)\, \forall\, \theta, \quad \mathrm{Var}(\hat{T}) < \mathrm{Var}(T) \quad \textit{for some } \theta \end{gathered} \tag{62}$$

The last expression will not hold if $T = \hat{T}$ with probability 1. The point is that an estimator which is unbiased and a function of sufficient statistics will be efficient. If, as is usually the case, there is only one unbiased estimator which is a function of the sufficient statistics, then that estimator is efficient.

5.6. **Example.** Let $X_1,\ X_2, \dots , X_n$ be iid random variables each having a normal distribution with mean μ and variance σ. The density is given by

$$\begin{aligned} f(x; \mu, \sigma^2) &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \cdot \mathrm{e}^{-\frac{n\mu^2}{2\sigma^2}} \cdot \mathrm{e}^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i\right)} \end{aligned} \tag{63}$$

This density only depends on the sufficient statistics

$$\sum_{i=1}^n x_i^2 \quad \text{and} \quad \sum_{i=1}^n x_i$$

6. The Information Matrix and the Cramer-Rao Bound

6.1. **The Information Matrix.** Consider a random sample $(X_1, \dots , X_n)$ from some population characterized by the parameter θ and density function $f(x; \theta)$. The distribution is assumed to be continuous and so the joint density which is the same as the likelihood function is given by

$$L(x_1, x_2, \dots , x_n) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) \tag{64}$$

The following assumptions, called regularity conditions, are used.

$$\begin{aligned} &\text{(i)} \quad f(\cdot; \theta) \text{ and } L(\cdot; \theta) \text{ are } C^2 \text{ w.r.t. } \theta. \\ &\text{(ii)} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L(x_1, \dots , x_n; \theta)\, dx_1\, dx_2 \dots dx_n = 1. \\ &\text{(iii)} \quad \text{The limits of integration don't depend on } \theta. \\ &\text{(iv)} \quad \text{Differentiation under the integral sign is allowed.} \end{aligned} \tag{65}$$

The notation C^2 means that the function is twice continuously differentiable. The regularity conditions imply the following theorem

Theorem 23. *If a likelihood function is regular then*

$$E\left[\frac{\partial log L(\cdot;\theta)}{\partial \theta_i}\right] = 0 \tag{66}$$

Proof: Take the derivative of the condition in (ii) and then multiply and divide by $L(\cdot;\theta)$ inside the integral.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L(x_1,\ \ldots\ ,\ x_n;\ \theta)\, dx_1\, dx_2\, \cdots\cdots dx_n = 1$$

$$\begin{aligned}
&\frac{\partial}{\partial \theta_i}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L(x_1,\ \ldots\ ,\ x_n;\ \theta)\, dx_1\ \ldots\ dx_n \\
&\qquad = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\partial L(x_1,\ \ldots\ ,\ x_n;\ \theta)}{\partial \theta_i}\, dx_1\ \ldots\ dx_n = 0 \\
&\qquad = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\partial L(x_1,\ \ldots\ ,\ x_n;\theta)}{\partial \theta_i}\frac{L(\cdot;\theta)}{L(\cdot;\theta)}\, dx_1\ \ldots\ dx_n = 0 \qquad (67) \\
&\qquad = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{\partial \log L(x_1,\ \ldots\ ,\ x_n;\ \theta)}{\partial \theta_i} L(\cdot;\theta)\, dx_1\ \ldots\ dx_n = 0 \\
&\qquad \Rightarrow E\left[\frac{\partial \log L(\cdot;\ \theta)}{\partial \theta_i}\right] = 0
\end{aligned}$$

This is the condition that is used to define the Fisher information number or matrix. The Fisher information matrix is given by

$$R(\theta) = -E\left[\frac{\partial^2 \log L(X,\theta)}{\partial \theta\, \partial \theta'}\right] \tag{68}$$

We can show that this matrix is the variance of the derivative of the log likelihood function with respect to θ, i.e.,

$$\text{Var}\left[\frac{\partial \log L(X;\theta)}{\partial \theta}\right]$$

This can be shown by differentiating the last expression in (67) with respect to θ (using the product rule) and remembering that the expected value of the derivative of the log

likelihood function is zero, so that its variance is just the expected value of its square.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial \log L(x_1,...,x_n;\theta)}{\partial\theta_i}L(\cdot;\theta)\,dx_1\,\ldots\,dx_n=0$$

$$\begin{aligned}
&\frac{\partial}{\partial\theta_k}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial \log L(\cdot;\theta)}{\partial\theta_i}L(\cdot;\theta)\,dx_1\,\ldots\,dx_n \\
&\qquad=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial^2 \log L(\cdot;\theta)}{\partial\theta_i\partial\theta_k}L(\cdot;\theta)\,dx_1\,\ldots dx_n \\
&\qquad+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial L(\cdot;\theta)}{\partial\theta_k}\frac{\partial \log L(\cdot;\theta)}{\partial\theta_i}\,dx_1\,\ldots\,dx_n=0 \qquad (69)\\
&\qquad=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial^2 logL(\cdot;\theta)}{\partial\theta_i\,\partial\theta_k}L(\cdot;\theta)\,dx_1\,\ldots\,dx_n \\
&\qquad+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial \log L(\cdot;\theta)}{\partial\theta_k}\frac{\partial \log L(\cdot;\theta)}{\partial\theta_i}L(\cdot;\theta)\,dx_1\,\ldots\,dx_n=0 \\
&\qquad=E\left[\frac{\partial^2 logL(\cdot;\theta)}{\partial\theta\,\partial\theta'}\right]+\mathrm{Var}\left[\frac{\partial logL(X;\theta)}{\partial\theta}\right]=0 \\
&\qquad\Rightarrow \mathrm{Var}\left[\frac{\partial logL(X;\theta)}{\partial\theta}\right]=-E\left[\frac{\partial^2 \log L(\cdot;\theta)}{\partial\theta\,\partial\theta'}\right]
\end{aligned}$$

6.2. **Cramer-Rao Theorem.**

Theorem 24. *Consider an unbiased estimator $\hat{\theta}$ of a parameter vector θ. Suppose that this estimator is unbiased with covariance matrix Σ. Then $\Sigma - R(\theta)^{-1}$ is a positive semi definite matrix. Thus $R(\theta)^{-1}$ is a lower bound for the variance of unbiased estimators.*

Proof: A general proof is given in Rao [8, Chapter 5]. We will present here a simpler proof for the case for a single parameter θ. The theorem is slightly restated as follows with the conditions being the same as in the general case.

Theorem 25. *Let $T(X)$ be any statistic such that $\mathrm{Var}_\theta\big(T(X)\big) < \infty\,\forall\,\theta$ where $T(X)$ is a statistic generated from a random sample $(X_1), \ldots, X_n)$ and θ is the population parameter. Let the expected value of $T(X)$ be $E_\theta\big(T(X)\big)$. Suppose the conditions (65) hold and $0 < R(\theta) < \infty$, then for all θ, $E_\theta\big(T(X)\big)$ is differentiable and*

$$\mathrm{Var}_\theta\big(T(X)\big) \geq \frac{\left[\frac{\partial E_\theta\left(T(X)\right)}{\partial \theta}\right]^2}{R(\theta)} \tag{70}$$

Proof: First it is clear that the following integral relationships hold.

$$\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty}\cdots\int_{-\infty}^{\infty} L(x,\theta)\,dx_1\,dx_2\,\ldots\,dx_n = 1$$

$$\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty}\cdots\int_{-\infty}^{\infty} T(x)L(x,\theta)\,dx_1\,dx_2\,\ldots\,dx_n = E(T(x)) \tag{71}$$

Now differentiate (71) with respect to θ to note that:

$$\begin{aligned}
&\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial L(x,\theta)}{\partial\theta}\,dx_1\,dx_2\,\ldots\,dx_n \\
&\qquad = \frac{\partial}{\partial\theta}\left[\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty}\cdots\int_{-\infty}^{\infty} L(x,\theta)\,dx_1\,dx_2\,\ldots dx_n\right] \\
&\qquad = 0 \\
&\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty}\cdots\int_{-\infty}^{\infty} T(x)\frac{\partial L(x,\theta)}{\partial\theta}\,dx_1\,dx_2\,\ldots\,dx_n \\
&\qquad = \frac{\partial}{\partial\theta}\left[\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty}\cdots\int_{-\infty}^{\infty} T(x)L(x,\theta)\,dx_1\,dx_2\,\ldots\,dx_n\right] \\
&\qquad = \frac{\partial E\big(T(x)\big)}{\partial\theta}
\end{aligned} \tag{72}$$

From (66) we already know that

$$E\left[\frac{\partial \log L(\cdot;\theta)}{\partial\theta}\right] = 0 \tag{73}$$

Similarly, using the logarithmic derivative relationship that

$$\frac{d\log f(x)}{dx} = \frac{1}{f(x)}\,\frac{df(x)}{dx} \tag{74}$$

and multiplying and dividing (72) by $L(X,\theta)$ we can show that

$$\int_{-\infty}^{-\infty}\int_{-\infty}^{-\infty}\cdots\int_{-\infty}^{\infty}T(x)\frac{\partial \log L(x,\theta)}{\partial\theta}L(x,\theta)\,dx_1\,dx_2\,\ldots\,dx_n$$

$$= E\left[T(x)\frac{\partial \log L(x,\theta)}{\partial\theta}\right] \tag{75}$$

$$= \frac{\partial E\big(T(x)\big)}{\partial\theta}$$

Now remember $\mathrm{Cov}(X_1X_2) = E(X_1X_2) - E(X_1)E(X_2)$ and realize that

$$\mathrm{Cov}\left[\frac{\partial \log L(x,\theta)}{\partial\theta}, T(X)\right] = \frac{\partial E(T(X))}{\partial\theta} \tag{76}$$

since the expected value of the derivative of the log $L(x,\,\theta)$ is zero.

Now remember that $|\mathrm{Cov}(X_1X_2)| < (\mathrm{Var}(X_1)\mathrm{Var}(X_2))^{1/2}$ from the Cauchy-Schwarz inequality and therefore

$$\left|\frac{\partial E\big(T(X)\big)}{\partial\theta}\right| \leq \left[\mathrm{Var}\big(T(X)\big)\,\mathrm{Var}\left(\frac{\partial \log L(X,\theta)}{\partial\theta}\right)\right]^{1/2} \tag{77}$$

But

$$\mathrm{Var}\left[\frac{\partial \log L(X,\,\theta)}{\partial\theta}\right] = R(\theta) \tag{78}$$

by (69) and the definition of $R(\theta)$. Substituting (78) into (77) will yield:

$$\left|\frac{\partial E\big(T(X)\big)}{\partial\theta}\right| \leq \sqrt{\mathrm{Var}\big(T(X)\big)R(\theta)} \tag{79}$$

which implies that

$$\left[\frac{\partial E\big(T(X)\big)}{\partial\theta}\right]^2 \leq \mathrm{Var}\big((X)\big)R(\theta) \tag{80}$$

and

$$\mathrm{Var}\big(T(X)\big) \geq \frac{\left[\frac{\partial E\left(T(X)\right)}{\partial\theta}\right]^2}{R(\theta)} \tag{81}$$

Corollary to theorem:

$$\mathrm{Var}_\theta\big(T(X)\big) \geq \frac{1}{R(\theta)} \tag{82}$$

6.3. **Example 1.** Let $X_1, \ldots, ...X_n$ be a random sample from a normal population with unknown mean μ and variance 1. So θ here is a scalar.

$$\begin{aligned} \log\ L &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ \frac{\partial \log L}{\partial \mu} &= \sum_{i=1}^{n}(x_i-\mu) \\ \frac{\partial^2 \log L}{\partial \mu^2} &= -n \end{aligned} \tag{83}$$

This then implies that $\mathrm{Var}(T(X)) \geq 1/n$. This implies that the sample mean which has variance $1/n$ is the minimum variance unbiased estimator.

6.4. **Example 2.** Consider a normal population with unknown mean μ and variance σ^2, $\theta = [\mu,\ \sigma^2]$

$$\log\ L(X,\ \mu,\ \sigma^2) = -\frac{n}{2}\log\ 2\pi - \frac{n}{2}\ \log\ \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{84}$$

$$\begin{bmatrix} \frac{\partial^2 \log\ L}{\partial \mu^2} & \frac{\partial^2 \log\ L}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \log\ L}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \log\ L}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(X_i-\mu) \\ -\frac{1}{\sigma^4}\sum_{n=1}^{n}(X_i-\mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(X_i-\mu)^2 \end{bmatrix} \tag{85}$$

$$R(\mu,\ \sigma^2) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \tag{86}$$

$$R(\mu,\ \sigma^2)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix} \tag{87}$$

Remember that

$$V\begin{bmatrix} \bar{X} \\ s^2 \end{bmatrix} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n-1) \end{bmatrix} \tag{88}$$

In this case the Cramer-Rao lower bound cannot be attained.

Note: We can show that the variance of $S^2 = 2\sigma^4/(n-1)$ as follows. We utilize the fact that the chi-square distribution is defined in terms of the normal distribution as follows

$$\begin{aligned} \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 &\sim \chi^2(n-1) \\ \Rightarrow \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2 &\sim \frac{\sigma^2\chi^2(n-1)}{n-1} \\ \Rightarrow S^2 &\sim \frac{\sigma^2\chi^2(n-1)}{n-1} \end{aligned} \tag{89}$$

For a chi-square random variable we have

$$\begin{aligned} E(\chi^2(\nu)) &= \nu \\ \text{Var}(\chi^2(\nu)) &= 2\nu \end{aligned} \tag{90}$$

Therefore we can compute the variance of S^2 as follows.

$$\begin{aligned} \text{Var}\left(S^2\right) &= \frac{\sigma^4}{(n-1)^2}\,\text{Var}\left(\chi^2(n-1)\right) \\ &= \frac{\sigma^4}{(n-1)^2}(2)(n-1) \\ &= \frac{2\sigma^4}{(n-1)} \end{aligned} \tag{91}$$

REFERENCES

[1] Amemiya, T. *Advanced Econometrics.* Cambridge: Harvard University Press, 1985.

[2] Bickel P.J., and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol 1).* 2nd Edition. Upper Saddle River, NJ: Prentice Hall, 2001.

[3] Billingsley, P. *Probability and Measure.* 3nd edition. New York: Wiley, 1995.

[4] Casella, G. And R.L. Berger. *Statistical Inference*. Pacific Grove, CA: Duxbury, 2002.

[5] Cramer, H. *Mathematical Methods of Statistics.* Princeton: Princeton University Press, 1946.

[6] Goldberger, A.S. *Econometric Theory.* New York: Wiley, 1964.

[7] Lukacs, E. *Stochastic Convergence.* 2nd edition. New York: Academic Press, 1975.

[8] Rao, C.R. *Linear Statistical Inference and its Applications.* 2nd edition. New York: Wiley, 1973.

[9] Resnick, S.I. *A Probability Path.* Boston: Birkhauser, 1998.

[10] Shiryaev, A.N. *Probability.* 2nd edition. New York: Springer-Verlag, 1996.

[11] Theil, H. *Principles of Econometrics*. New York: Wiley, 1971.