BASIC STATISTICS

1. Samples, Random Sampling and Sample Statistics

1.1. **Random Sample.** The random variables $X_1, X_2, ..., X_n$ are called a random sample of size n from the population f(x) if $X_1, X_2, ..., X_n$ are mutually independent random variables and the marginal probability density function of each X_i is the same function of f(x). Alternatively, $X_1, X_2, ..., X_n$ are called independent and identically distributed random variables with pdf f(x). We abbreviate independent and identically distributed as iid.

Most experiments involve n >1 repeated observations on a particular variable, the first observation is X_1, the second is X_2, and so on. Each X_i is an observation on the same variable and each X_i has a marginal distribution given by f(x). Given that the observations are collected in such a way that the value of one observation has no effect or relationship with any of the other observations, the $X_1, X_2, ..., X_n$ are mutually independent. Therefore we can write the joint probability density for the sample $X_1, X_2, ..., X_n$ as

$$f(x_1, x_2, ..., x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i) \tag{1}$$

If the underlying probability model is parameterized by θ, then we can also write

$$f(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2}$$

Note that the same θ is used in each term of the product, or in each marginal density. A different value of θ would lead to a different properties for the random sample.

1.2. **Statistics.** Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population and let $T(x_1, x_2, ..., x_n)$ be a real valued or vector valued function whose domain includes the sample space of $(X_1, X_2, ..., X_n)$. Then the random variable or random vector $Y = (X_1, X_2, ..., X_n)$ is called a statistic. A statistic is a map from the sample space of $(X_1, X_2, ..., X_n)$ call it **X**, to some space of values, usually R^1 or R^n. T is what we compute when we observe the random variable X take on some specific values in a sample. The probability distribution of a statistic Y = T(X) is called the sampling distribution of Y. Notice that T(·) is a function of sample values only, it does not depend on any underlying parameters, θ.

1.3. **Some Commonly Used Statistics.**

1.3.1. *Sample mean.* The sample mean is the arithmetic average of the values in a random sample. It is usually denoted

$$\bar{X}(X_1, X_2, \cdots, X_n) = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3}$$

The observed value of $\bar{X}$ in any sample is demoted by the lower case letter, i.e., $\bar{x}$.

Date: September 27, 2004.

1.3.2. *Sample variance.* The sample variance is the statistic defined by

$$S^2(X_1, X_2, \cdots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{4}$$

The observed value of S^2 in any sample is demoted by the lower case letter, i.e., s^2.

1.3.3. *Sample standard deviation.* The sample standard deviation is the statistic defined by

$$S = \sqrt{S^2} \tag{5}$$

1.3.4. *Sample midrange.* The sample mid-range is the statistic defined by

$$\frac{max(X_1, X_2, \cdots, X_n) - min(X_1, X_2, \cdots, X_n)}{2} \tag{6}$$

1.3.5. *Empirical distribution function.* The empirical distribution function is defined by

$$\hat{F}(X_1, X_2, \cdots, X_n)(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < x) \tag{7}$$

where $\hat{F}(X_1, X_2, \cdots, X_n)(x)$ means we are evaluating the statistic $\hat{F}(X_1,\ X_2,\ \cdots, X_n)$ at the particular value x. The random sample $X_1, X_2, ..., X_n$ is assumed to come from a probability defined on R^1 and I(A) is the indicator of the event A. This statistic takes values in the set of all distribution functions on R^1. It estimates the function valued parameter F defined by its evaluation at $x \in R^1$

$$F(P)(x) = P[X < x] \tag{8}$$

2. Distribution of Sample Statistics

2.1. Theorem 1 on squared deviations and sample variances.

Theorem 1. *Let $x_1, x_2, \cdots x_n$ be any numbers and let $\bar{x} = \frac{x_1+x_2+...+x_n}{n}$. Then the following two items hold.*

a: $min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$
b: $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Part a says that the sample mean is the value about which the sum of squared deviations is minimized. Part b is a simple identity that will prove immensely useful in dealing with statistical data.

Proof. First consider part a of theorem 1. Add and subtract $\bar{x}$ from the expression on the lefthand side in part a and then expand as follows

$$\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2 \tag{9}$$

Now write out the middle term in 9 and simplify

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) &= \bar{x}\sum_{i=1}^{n} x_i - a\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n}\bar{x} + \bar{x}\sum_{i=1}^{n} a \\ &= n\bar{x}^2 - an\bar{x} - n\bar{x}^2 + n\bar{x}a \\ &= 0\end{aligned} \tag{10}$$

We can then write 9 as

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2 \tag{11}$$

Equation 11 is clearly minimized when $a = \bar{x}$. Now consider part b of theorem 1. Expand the second expression in part b and simplify

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned} \tag{12}$$

□

2.2. **Theorem 2 on expected values and variances of sums.**

Theorem 2. *Let $X_1, X_2, \cdots X_n$ be a random sample from a population and let g(x) be a function such that $Eg(X_1)$ and $Varg(X_1)$exist. Then following two items hold.*

a: $E(\sum_{i=1}^{n} g(X_i)) = n(Eg(X_1))$
b: $Var(\sum_{i=1}^{n} g(X_i)) = n(Varg(X_1))$

Proof. First consider part a of theorem 2. Write the expected value of the sum as the sum of the expected values and then note that $Eg(X_1) = Eg(X_2) = ...Eg(X_i) = ...Eg(X_n)$ because the X_i are all from the same distribution.

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} E(g(X_i)) = n(Eg(X_1)) \tag{13}$$

First consider part b of theorem 2. Write the definition of the variance for a variable z as $E(z-E(z))^2$ and then combine terms in the summation sign.

$$Var\left(\sum_{i=1}^{n} g(X_i)\right) = E\left[\sum_{i=1}^{n} g(X_i) - E\left(\sum_{i=1}^{n} g(X_i)\right)\right]^2 \tag{14}$$

Now write out the bottom expression in equation 14 as follows

$$\begin{aligned}Var\left(\sum_{i=1}^{n} g(X_i)\right) &= E\left[g(X_1) - E(g(X_1))\right]^2 + E\left[g(X_1) - E(g(X_1))\right]\ E\left[g(X_2) - E(g(X_2))\right] \\ &+ E\left[g(X_1) - E(g(X_1))\right]\ E\left[g(X_3) - E(g(X_3))\right] + \cdots \\ &+ E\left[g(X_2) - E(g(X_2))\right]\ E\left[g(X_1) - E(g(X_1))\right] + E\left[g(X_2) - E(g(X_2))\right]^2 \\ &+ E\left[g(X_2) - E(g(X_2))\right]\ E\left[g(X_3) - E(g(X_3))\right] + \cdots \\ &+ \cdots \\ &+ E\left[g(X_n) - E(g(X_n))\right]\ E\left[g(X_1) - E(g(X_1))\right]\ +\ \cdots\ +\ E\left[g(X_n) - E(g(X_n))\right]^2\end{aligned} \tag{15}$$

Each of the squared terms in the summation is a variance, i.e., the variance of $X_i = var(X_1)$. Specifically

$$E\left[g(X_i) - E(g(X_i))\right]^2 = Varg(X_i) = Varg(X_1) \tag{16}$$

The other terms in the summation in 15 are covariances of the form

$$E\left[g(X_i) - E(g(X_i))\right]\ E\left[g(X_j) - E(g(X_j))\right] = Cov\left[g(X_i),\ g(X_j)\right] \tag{17}$$

Now we can use the fact that the X_1 and X_j in the sample $X_1, X_2, \cdots, X_n$ are independent to assert that each of the covariances inn the sum in 15 is zero. We can then rewrite 15 as

$$\begin{aligned} Var\left(\sum_{i=1}^{n} g(X_i)\right) &= E\left[g(X_1) - E(g(X_1))\right]^2 + E\left[g(X_2) - E(g(X_2))\right]^2\ +\ \cdots + E\left[g(X_n) - E(g(X_n))\right]^2 \\ &= Var(g(X_1)) + Var(g(X_2)) + Var(g(X_3)) + \ \cdots \\ &= \sum_{i=1}^{n} Varg(X_i) \\ &= \sum_{i=1}^{n} Varg(X_1) \\ &= n\, Varg(X_1) \end{aligned} \tag{18}$$

□

2.3. **Theorem 3 on expected values of sample statistics.**

Theorem 3. *Let $X_1, X_2, \cdots X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

a: $E\bar{X} = \mu$
b: $Var\bar{X} = \frac{\sigma^2}{n}$
c: $ES^2 = \sigma^2$

Proof of part a. In theorem 2 let $g(X) = g(X_i) = \frac{X_i}{n}$. This implies that $Eg(X_i) = \frac{\mu}{n}$ Then we can write

$$E\bar{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\left(nEX_1\right) = \mu \tag{19}$$

Proof of part b.

In theorem 2 let $g(X) = g(X_i) = \frac{X_i}{n}$. This implies that $Varg(X_i) = \frac{\sigma^2}{n}$ Then we can write

$$Var\bar{X} = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(nVarX_1\right) = \frac{\sigma^2}{n} \tag{20}$$

Proof of part c.

As in part b of theorem 1, write S^2 as a function of the sum of square of X_i minus n times the mean of $X_i$squared and then simplify

$$\begin{aligned} ES^2 &= E\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\ &= \frac{1}{n-1}\ \left(nEX_1^2 - nE\bar{X}^2\right) \\ &= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \sigma^2 \end{aligned} \tag{21}$$

The last line follows from the definition of a random variable, i.e.,

$$\begin{aligned} Var\ X = \sigma_X^2 &= EX^2 - (EX)^2 \\ &= EX^2 - \mu_X^2 \\ \Rightarrow E\ X^2 &= \sigma_X^2 - \mu_X^2 \end{aligned} \tag{22}$$

2.4. **Unbiased Statistics.** We say that a statistic T(X)is an unbiased statistic for the parameter θ of the underlying probability distribution if E T(X) = θ. Given this definition, $\bar{X}$ is an unbiased statistic for μ,and S^2 is an unbiased statistic for σ^2 in a random sample.

3. Methods of Estimation

Let $Y_1, Y_2, \cdots Y_n$ denote a random sample from a parent population characterized by the parameters $\theta_1, \theta_2, \cdots \theta_k$. It is assumed that the random variable Y has an associated density function $f(\,\cdot\,; \theta_1, \theta_2, \cdots \theta_k)$.

3.1. **Method of Moments.**

3.1.1. *Definition of Moments.* If Y is a random variable, the rth moment of Y, usually denoted by μ_r', is defined as

$$\begin{aligned} \mu_r' &= E(Y^r) \\ &= \int_{-\infty}^{\infty} y^r f(y; \theta_1, \theta_2, \cdots \theta_k)\, dy \end{aligned} \tag{23}$$

if the expectation exists. Note that $\mu_1' = E(Y) = \mu_Y$, the mean of Y. Moments are sometimes written as functions of θ.

$$E(Y^r) = \mu_r' = g_r\left(\theta_1, \theta_2, \cdots \theta_k\right) \tag{24}$$

3.1.2. *Definition of Central Moments.* If Y is a random variable, the rth central moment of Y about *a* is defined as $E[(Y-a)^r]$. If a = μ_r, we have the rth central moment of Y about μ_Y, denoted by μ_r, which is

$$\begin{aligned} \mu_r &= E[(Y - \mu_Y)^r] \\ &= \int_{-\infty}^{\infty} (y - \mu_y)^r f(y; \theta_1, \theta_2, \cdots \theta_k)\, dy \end{aligned} \tag{25}$$

Note that$\mu_1 = E[(Y - \mu_Y)] = 0$ and $\mu_2 = E[(Y - \mu_Y)^2] = Var[Y]$. Also note that all odd numbered moments of Y around its mean are zero for symmetrical dsitributions, provided such moments exist.

3.1.3. ***Sample** Moments about the Origin.* The rth sample moment about the origin is defined as

$$\hat{\mu}_r' = \bar{x}_n^r = \frac{1}{n}\sum_{i=1}^{n} y_i^r \tag{26}$$

3.1.4. *Estimation Using the Method of Moments.* In general μ'_r will be a known function of the parameters $\theta_1, \theta_2, \cdots \theta_k$ of the distribution of Y, that is $\mu'_r = g_r(\theta_1, \theta_2, \cdots \theta_k)$. Now let $y_1, y_2, \cdots, y_n$ be a random sample from the density $f(\cdot; \theta_1, \theta_2, \cdots \theta_k)$. Form the K equations

$$\begin{aligned}
\mu'_1 =& g_1(\theta_1, \theta_2, \cdots \theta_k) = \hat{\mu}'_1 = \frac{1}{n}\sum_{i=1}^{n} y_i \\
\mu'_2 =& g_2(\theta_1, \theta_2, \cdots \theta_k) = \hat{\mu}'_2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
&\vdots \\
\mu'_K =& g_K(\theta_1, \theta_2, \cdots \theta_k) = \hat{\mu}'_K = \frac{1}{n}\sum_{i=1}^{n} y_i^K
\end{aligned} \tag{27}$$

The estimators of $\theta_1, \theta_2, \cdots \theta_k$, based on the method of moments, are obtained by solving the system of equations for the K parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \cdots \hat{\theta_K}$.

This principle of estimation is based upon the convention of picking the estimators of θ_1 in such a manner that the corresponding population (theoretical) moments are equal to the sample moments. These estimators are consistent under fairly general regularity conditions, but are not generally efficient. Method of moments estimators may also not be unique.

3.1.5. *Example using density function* $f(y) = (p+1)\, y^p$. Consider a density function given by

$$\begin{aligned}
f(y) &= (p+1)\, y^p \quad 0 \le y \le 1 \\
&= 0 \quad otherwise
\end{aligned} \tag{28}$$

Let $Y_1, Y_2, \cdots Y_n$ denote a random sample from the given population. Express the first moment of Y as a function of the parameters.

$$\begin{aligned}
E(Y) &= \int_{-\infty}^{\infty} y\, f(y)\, dy \\
&= \int_0^1 y\, (p+1)\, y^p\, dy \\
&= \int_0^1 y^{p+1}\, (p+1)\, dy \\
&= \left. \frac{y^{p+2}\, (p+1)}{(p+2)} \right|_0^1 \\
&= \frac{p+1}{p+2}
\end{aligned} \tag{29}$$

Then set this expression of the parameters equal to the first sample moment and solve for p.

$$\begin{aligned} \mu_1' &= E(Y) = \frac{p+1}{p+2} \\ \Rightarrow \frac{p+1}{p+2} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \\ \Rightarrow p+1 &= (p+2)\,\bar{y} = p\bar{y} + 2\bar{y} \\ \Rightarrow p - p\bar{y} &= 2\,\bar{y} - 1 \\ \Rightarrow p(1-\bar{y}) &= 2\,\bar{y} - 1 \\ \Rightarrow \hat{p} &= \frac{2\,\bar{y}-1}{1-\bar{y}} \end{aligned} \tag{30}$$

3.1.6. *Example using the Normal Distribution.* Let $Y_1, Y_2, \cdots Y_n$ denote a random sample from a normal distribution with mean μ and variance σ^2. Let $(\theta_1, \theta_2) = (\mu, \sigma^2)$. Remember that $\mu = \mu_1'$ and $\sigma^2 = E[Y^2] - E^2[Y] = \mu_2' - (\mu_1')^2$.

$$\begin{aligned} \mu_1' &= E(Y) = \mu \\ \mu_2' &= E(Y^2) = \sigma^2 + E^2[Y] = \sigma^2 + \mu^2 \end{aligned} \tag{31}$$

Now set the first population moment equal to its sample analogue to obtain

$$\begin{aligned} \mu &= \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \\ \Rightarrow \hat{\mu} &= \bar{y} \end{aligned} \tag{32}$$

Now set the second population moment equal to its sample analogue

$$\begin{aligned} \sigma^2 + \mu^2 &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 \\ \Rightarrow \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \mu^2 \\ \Rightarrow \sigma &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \mu^2} \end{aligned} \tag{33}$$

Now replace μ in equation 33 with its estimator from equation 32 to obtain

$$\begin{aligned} \hat{\sigma} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2} \\ \Rightarrow \hat{\sigma} &= \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n}} \end{aligned} \tag{34}$$

This is, of course, from the sample standard deviation defined in equations 4 and 5.

3.1.7. *Example using the Gamma Distribution.* Let $X_1, X_2, \cdots X_n$ denote a random sample from a gamma distribution with parameters θ and α. The density function is given by

$$\begin{aligned} f(x;\theta,\alpha) &= \frac{1}{\theta^{\alpha}\Gamma(\alpha)}\, x^{\alpha-1}\, e^{\frac{-x}{\theta}} \qquad 0 \leq x < \infty \\ &= 0 \qquad otherwise \end{aligned} \tag{35}$$

Find the first moment of the gamma distribution by integrating as follows

$$\begin{aligned} E(X) &= \int_0^{\infty} x \frac{1}{\theta^{\alpha}\Gamma(\alpha)}\, x^{\alpha-1}\, e^{\frac{-x}{\theta}}\, dx \\ &= \frac{1}{\theta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} x^{(1+\alpha)-1}\, e^{\frac{-x}{\theta}}\, dx \end{aligned} \tag{36}$$

If we multiply equation 36 by $\theta^{1+\alpha}\ \Gamma(1+\alpha)$ we obtain

$$E(X) = \frac{\theta^{1+\alpha}\ \Gamma(1+\alpha)}{\theta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} \frac{1}{\theta^{1+\alpha}\ \Gamma(1+\alpha)}\, x^{(1+\alpha)-1}\, e^{\frac{-x}{\theta}}\, dx \tag{37}$$

The integrand of equation 37 is a gamma density with parameters θ and $1-\alpha$ This integrand will integrate to one so that we obtain the expression in front of the integral sign as the E(X).

$$\begin{aligned} E(X) &= \frac{\theta^{1+\alpha}\ \Gamma(1+\alpha)}{\theta^{\alpha}\Gamma(\alpha)} \\ &= \frac{\theta\ \Gamma(1+\alpha)}{\Gamma(\alpha)} \end{aligned} \tag{38}$$

The gamma function has the property that $\Gamma(t) = (t-1)\Gamma(t-1)$ or $\Gamma(v+1) = v\Gamma(v)$. Replacing $\Gamma(1+\alpha)$ with $\alpha\,\Gamma(\alpha)$ in equation 38, we obtain

$$\begin{aligned} E(X) &= \frac{\theta\,\Gamma(1+\alpha)}{\Gamma(\alpha)} \\ &= \frac{\theta\,\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\ &= \theta\,\alpha \end{aligned} \tag{39}$$

We can find the second moment by finding $E(X^2)$. To do this we multiply the gamma density in equation 36 by x^2 instead of x. Carrying out the computation we obtain

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2 \frac{1}{\theta^{\alpha}\Gamma(\alpha)}\, x^{\alpha-1}\, e^{\frac{-x}{\theta}}\, dx \\ &= \frac{1}{\theta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} x^{(2+\alpha)-1}\, e^{\frac{-x}{\theta}}\, dx \end{aligned} \tag{40}$$

If we then multiply 40 by $\theta^{2+\alpha}\ \Gamma(2+\alpha)$ we obtain

$$\begin{aligned} E(X^2) =& \frac{\theta^{2+\alpha}\ \Gamma(2+\alpha)}{\theta^{\alpha}\Gamma(\alpha)} \int_0^{\infty} \frac{1}{\theta^{2+\alpha}\ \Gamma(2+\alpha)}\, x^{(2+\alpha)-1}\, e^{\frac{-x}{\theta}}\, dx \\ =& \frac{\theta^{2+\alpha}\,\Gamma(2+\alpha)}{\theta^{\alpha}\Gamma(\alpha)} \\ =& \frac{\theta^2(\alpha+1)\,\Gamma(1+\alpha)}{\Gamma(\alpha)} \\ =& \frac{\theta^2\alpha(\alpha+1)\,\Gamma(\alpha)}{\Gamma(\alpha)} \\ =& \theta^2\alpha(\alpha+1) \end{aligned} \tag{41}$$

Now set the first population moment equal to the sample analogue to obtain

$$\begin{aligned} \theta\,\alpha &= \frac{1}{n}\sum_{i=1}^{n} x_i \ =\ \bar{x} \\ \Rightarrow\ \hat{\alpha} &= \frac{\bar{x}}{\theta} \end{aligned} \tag{42}$$

Now set the second population moment equal to its sample analogue

$$\begin{aligned} \theta^2\ \alpha\,(\,\alpha\ +\ 1\,) \ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ \Rightarrow \theta^2\ &= \ \frac{\sum_{i=1}^{n}\ x_i^2}{n\,\alpha\,(\,\alpha\ +\ 1)} \\ \Rightarrow \theta^2\ &= \ \frac{\sum_{i=1}^{n}\ x_i^2}{n\ \left(\frac{\bar{x}}{\theta}\right)\ \left(\left(\frac{\bar{x}}{\theta}\right)\ +\ 1\right)} \\ \Rightarrow \theta^2\ &= \ \frac{\sum_{i=1}^{n}\ x_i^2}{\left(\frac{n\,\bar{x}^{\,2}}{\theta^2}\right)\ +\ \left(\frac{n\,\bar{x}}{\theta}\right)} \\ \Rightarrow & n\,\bar{x}^{\,2}\ + n\,\bar{x}\ \theta\ =\ \sum_{i=1}^{n} x_i^2 \\ \Rightarrow & n\,\bar{x}\ \theta\ =\ \sum_{i=1}^{n} x_i^2\ -n\,\bar{x}^{\,2} \\ \Rightarrow \theta\ &= \ \frac{\sum_{i=1}^{n}\ x_i^2\ -n\,\bar{x}^2}{n\,\bar{x}} \\ &= \frac{\sum_{i=1}^{n}\ (x_i\ -\bar{x})^{\,2}}{n\,\bar{x}} \end{aligned} \tag{43}$$

3.2. **Method of least squares estimation.** Consider the situation in which the Y_i from the random sample can be written in the form

$$Y_i\ =\ \beta\ +\ \epsilon_i\ =\ \hat{\beta}\ +\ e_i \tag{44}$$

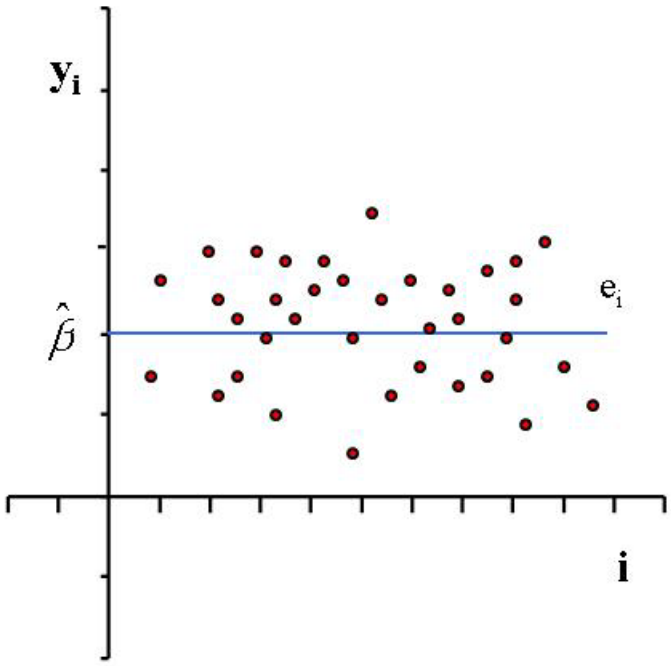
where E(ϵ_i) = 0 and Var(ϵ_i) = σ^2 for all i. This is equivalent to stating that the population from which y_i is drawn has a mean of β and a variance of σ^2.

The least squares estimator of β is obtained by minimizing the sum of squares errors, SSE, defined by

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta})^2 \tag{45}$$

The idea is to pick the value of $\hat{\beta}$ to estimate β which minimizes SSE. Pictorially we select the value of $\hat{\beta}$ which minimizes the sum of squares of the vertical deviations in figure 1.

FIGURE 1. Least Squares Estimation



The solution is obtained by finding the value of β that minimizes equation 45.

$$\begin{aligned} \frac{\partial SSE}{\partial \beta} &= 2 \sum_{i=1}^{n} (y_i - \hat{\beta})(-1) = 0 \\ \Rightarrow \hat{\beta} &= \frac{1}{n} \sum_{i-1}^{n} y_i = \bar{y} \end{aligned} \tag{46}$$

This method chooses values of the parameters of the underlying distribution, θ, such that the distance between the elements of the random sample and "predicted" values are minimized.

3.3. **Method of maximum likelihood estimation (MLE).** Least squares is independent of a specification of a density function for the parent population. Now assume that

$$y_i \sim f\ (\cdot;\ \theta\ =\ (\theta_1, ..., \theta_K)\ )\ \forall i. \tag{47}$$

3.3.1. *Motivation for the MLE method.* If a random variable Y has a probability density function f(·; θ) characterized by the parameters $\theta = (\theta_1, \ldots, \theta_k)$, then the maximum likelihood estimators (MLE) of $\theta_1, \ldots, \theta_k$ are the values of these parameters which would have most likely generated the given sample.

3.3.2. *Theoretical development of the MLE method.* The joint density of a random sample $y_1, y_2, \ldots, y_n$ is given by $L\ =\ g\,(y_1\,, \cdots,\ y_n;\ \theta)\ =\ f(y_1\,;\ \theta)\ \cdot\ f\,(y_2\,;\ \theta)\ \cdot\ f(y_3\,;\ \theta)\ \cdot\ \cdots\ f(y_n\,;\ \theta)$. Given that we have a random sample, the joint density is just the product of the marginal density functions. This is referred to as the *likelihood function*. The MLE of the θ_i are the θ_i which maximize the likelihood function.

The necessary conditions for an optimum are:

$$\frac{\partial L}{\partial \theta_i}\ =\ 0\ i = 1, 2, ..., k \tag{48}$$

This gives k equations in k unknowns to solve for the k parameters $\theta_1, ..., \theta_k$. In many instances it will be convenient to maximize ℓ = ln L rather than L given that the log of a product is the sum of the logs.

3.3.3. *Example 1.* Let the random variable X_i be distributed as a normal N(μ,σ^2) so that its density is given by

$$f(x_i;\ \mu\,,\ \sigma^2)\ =\ \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\ \cdot\ e^{\frac{-1}{2}\left(\frac{x_i\ -\ \mu}{\sigma}\right)^2} \tag{49}$$

Its likelihood function is given by

$$\begin{aligned} L\ &= \Pi_{i=1}^{n}\, f(x_i;\ \mu\,,\ \sigma^2)\ =\ f(x_1)\, f(x_2)\ \cdots\ f(x_n) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\ e^{\frac{-1}{2}\ \left(\ \frac{x_i\ -\ \mu}{\sigma}\ \right)^2} \qquad \cdots\ e^{\frac{-1}{2}\left(\frac{x_n\ -\ \mu}{\sigma}\right)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\ e^{\frac{-1}{2\,\sigma^2}\ \ \sum_{i=1}^{n}(x_i\ -\ \mu)^2} \\ \Rightarrow\ \ln\ L\ &=\ \ell\ =\ \frac{-n}{2}\ \ln(2\,\pi\,\sigma^2)\ -\ \frac{1}{2\,\sigma^2}\sum_{i=1}^{n}\,(x_i\ -\ \mu\,)^2 \end{aligned} \tag{50}$$

The MLE of μ and σ^2 are obtained by taking the partial derivatives of equation 50

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{-n}{2} \left[\frac{2\pi}{2\pi\hat{\sigma}^2} \right] - \left(\frac{1}{2} \right) (-1)(\hat{\sigma}^2)^{-2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0$$

$$\Rightarrow \frac{n}{2\,\hat{\sigma}^2} = \frac{1}{(2\hat{\sigma}^2)^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

$$\Rightarrow n = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{51}$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \left(\frac{n-1}{n} \right) s^2$$

The MLE of σ^2 is equal to the sample variance and not S^2; hence, the MLE is not unbiased as can be seen from equation 21. The MLE of μ is the sample mean.

3.3.4. *Example 2 - Poisson.* The random variable X_i is distributed as a Poisson if the density of X_i is given by

$$f(x_i;\ \lambda) = \begin{cases} \frac{e^{-\lambda}\,\lambda^{x_i}}{x_i\,!} & x_i\ a\ non-negative\ integer \\ 0 & otherwise \end{cases}$$
$$mean\ (X) = \lambda \tag{52}$$
$$Var\ (X) = \lambda$$

The likelihood function is given by

$$L = \left[\frac{e^{-\lambda}\lambda^{x_1}}{x_1\,!} \right] \cdots \left[\frac{e^{-\lambda}\lambda^{x_n}}{x_n\,!} \right]$$

$$= \frac{e^{-\lambda n}\,\lambda^{\Sigma_{i=1}^{n} x_i}}{\pi_{i=1}^{n}\, x_i\,!} \tag{53}$$

$$\Rightarrow \ln L = \ell = -\lambda n + \sum_{i=1}^{n} x_i \ln \lambda - \ln (\pi_{i=1}^{n}\, x_i\,!)$$

To obtain a MLE of λ, differentiate ℓ with respect to λ:

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda} &= -n + \sum_{i=1}^{n} x_i \frac{1}{\lambda} = 0 \\ \Rightarrow \hat{\lambda} &= \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \end{aligned} \tag{54}$$

3.3.5. *Example 3.* Consider the density function

$$\begin{aligned} f(y) &= (p+1)\, y^p \; 0 \leq y \leq 1 \\ &= 0 \quad otherwise \end{aligned} \tag{55}$$

The likelihood function is given by

$$\begin{aligned} L &= \Pi_{i=1}^{n} (p+1)\, y_i^p \\ \ln L = \ell &= \sum_{i=1}^{n} \ln [(p+1)\, y_i^p] \\ &= \sum_{i=1}^{n} (\ln (p+1) + p \ln y_i) \end{aligned} \tag{56}$$

To obtain the MLE estimator differentiate 56 with respect to p

$$\begin{aligned} \frac{\partial \ell}{\partial p} &= \sum_{i=1}^{n} \left(\frac{1}{p+1} + \ln y_i \right) = 0 \\ \Rightarrow \sum_{i=1}^{n} \frac{1}{\hat{p}+1} &= \sum_{i=1}^{n} (-\ln y_i) \\ \Rightarrow \frac{n}{\hat{p}+1} &= \sum_{i=1}^{n} (-\ln y_i) \\ \Rightarrow \hat{p}+1 &= \frac{-n}{\sum_{i=1}^{n} \ln y_i} \\ \Rightarrow \hat{p} &= \frac{-n}{\sum_{i=1}^{n} \ln y_i} - 1 \end{aligned} \tag{57}$$

3.3.6. *Example 4.* Consider the density function

$$f(y) = p^{y_i} (1-p)^{1-y_i} \qquad 0 \leq p \leq 1 \tag{58}$$

The likelihood function is given by

$$
\begin{aligned}
L &= \Pi_{i=1}^{n}\, p^{y_i}\, (1 - p)^{1 - y_i} \\
&= p^{\sum_{i=1}^{n} y_i}\, (1 - p)^{n - \sum_{i=1}^{n} y_i} \\
\ln L = \ell &= \sum_{i=1}^{n} y_i \ln p + \left(n - \sum_{i=1}^{n} y_i \right) \ln (1 - p)
\end{aligned}
\tag{59}
$$

To obtain the MLE estimator differentiate 59 with respect to p where we assume that $0 < p < 1$.

$$
\begin{aligned}
\frac{\partial \ell}{\partial p} &= \frac{\sum_{i=1}^{n} y_i}{p} - \frac{(n - \sum_{i=1}^{n} y_i)}{1 - p} = 0 \\
\Rightarrow \frac{\sum_{i=1}^{n} y_i}{p} &= \frac{(n - \sum_{i=1}^{n} y_i)}{1 - p} \\
\Rightarrow \sum_{i=1}^{n} y_i - p \sum_{i=1}^{n} y_i &= n p - p \sum_{i=1}^{n} y_i \\
\Rightarrow \sum_{i=1}^{n} y_i &= n p \\
\Rightarrow \frac{\sum_{i=1}^{n} y_i}{n} &= \hat{p}
\end{aligned}
\tag{60}
$$

3.4. **Principle of Best Linear Unbiased Estimation (BLUE).**

3.4.1. *Principle of Best Linear Unbiased Estimation.* Start with some desired properties and deduce an estimator satisfying them. For example suppose that we want the estimator to be linear in the observed random variables. This means that if the observations are $y_1, \ldots, y_n$, an estimator of θ must satisfy

$$
\hat{\theta} = \sum_{i=1}^{n} a_i\, y_i \tag{61}
$$

where the a_i are to be determined.

3.4.2. *Some required properties of the estimator (arbitrary).*

1: $E(\hat{\theta}) = \theta\ (unbiased)$

2: $Var(\hat{\theta}) \leq VAR(\tilde{\theta})\ (minimum\ variance)$ where $\tilde{\theta}$ is any other *linear* combination of the y_i that also produces an unbiased estimator.

3.4.3. *Example.* Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample drawn from a population having a mean μ and variance σ^2. Now derive the best linear unbiased estimator (BLUE) of μ.

Let the proposed estimator be denoted by $\hat{\theta}$. It is linear so we can write it as follows.

$$
\hat{\theta} = \sum_{i=1}^{n} a_i\, y_i \tag{62}
$$

If the estimator is to be unbiased, there will be restrictions on the a_i. Specifically

$$
\begin{aligned}
Unbiasedness \quad \Rightarrow \; E(\hat{\theta}) &= E\left(\sum_{i=1}^{n} a_i \, y_i\right) \\
&= \sum_{i=1}^{n} a_i \, E(y_i) \\
&= \sum_{i=1}^{n} a_i \, \mu \\
&= \mu \sum_{i=1}^{n} a_i \\
&=> \sum_{i=1}^{n} a_i \; = \; 1
\end{aligned}
\tag{63}
$$

Now consider the variance of $\hat{\theta}$.

$$
\begin{aligned}
Var\,(\,\hat{\theta}\,) &= Var\left[\sum_{i=1}^{n} a_i \, y_i\right] \\
&= \sum \; a_i^2 \; Var(y_i) \; + \; \Sigma \, \Sigma_{i \neq j} \; a_i \, a_j \; Cov\,(y_i \, y_j) \\
&= \sum_{i=1}^{n} a_i^2 \, \sigma^2
\end{aligned}
\tag{64}
$$

because the covariance between y_i and y_j ($i \neq j$) is equal to zero due to the fact that the y's are drawn from a random sample.

The problem of obtaining a BLUE of μ becomes that of minimizing $\sum_{i=1}^{n} \; a_i^2$ subject to the constraint $\sum_{i=1}^{n} \; a_i \; = \; 1$. This is done by setting up a Lagrangian

$$
L(a, \lambda) \; = \; \sum_{i=1}^{n} a_i^2 \; - \; \lambda(\sum_{i=1}^{n} a_i \; - \; 1) \tag{65}
$$

The necessary conditions for an optimum are

$$
\begin{aligned}
\frac{\partial L}{\partial a_1} &= 2a_1 \; - \; \lambda \; = \; 0 \\
&\;\;\vdots \\
\frac{\partial L}{\partial a_n} &= 2a_n \; - \; \lambda \; = \; 0 \\
\frac{\partial L}{\partial \lambda} &= \; - \sum_{i=1}^{n} a_i \; + \; 1 \; = \; 0
\end{aligned}
\tag{66}
$$

The first n equations imply that $a_1 = a_2 = a_3 = \ldots a_n$ so that the last equation implies that

$$\begin{aligned}
&\sum_{i=1}^{n} a_i \; - \; 1 \; = 0 \\
&\Rightarrow \; na_i \; - \; 1 \; = 0 \\
&\Rightarrow \; na_i \; = 1 \\
&\Rightarrow \; a_i \; = \frac{1}{n} \\
&\Rightarrow \; \hat{\theta} \; = \sum_{i=1}^{n} a_i \, y_i \; = \; \frac{1}{n} \sum_{i=1}^{n} y_i \; = \bar{y}
\end{aligned} \tag{67}$$

Note that equal weights are assigned to each observation.

4. Finite Sample Properties of Estimators

4.1. **Introduction to sample properties of estimators.** In section 3 we discussed alternative methods of estimating the unknown parameters in a model. In order to compare the estimating techniques we will discuss some criteria which are frequently used in such a comparison. Let θ denote an unknown parameter and let $\hat{\theta}\, and\, \tilde{\theta}$ be alternative estimators. Now define the bias, variance and mean squared error of $\hat{\theta}$ as

$$\begin{aligned}
Bias\,(\hat{\theta}) \; &= E\,(\hat{\theta}) \; - \; \theta \\
Var\,(\hat{\theta}) \; &= E \; \left(\hat{\theta} \; - E\,(\hat{\theta})\right)^2 \\
MSE\,(\hat{\theta}\,) \; &= E \; \left(\hat{\theta} \; - \; \theta\right)^2 \\
&= Var\,(\hat{\theta}) \; + \; \left(Bias \; \left(\hat{\theta}\right)\right)^2
\end{aligned} \tag{68}$$

The result on mean squared error can be seen as follows

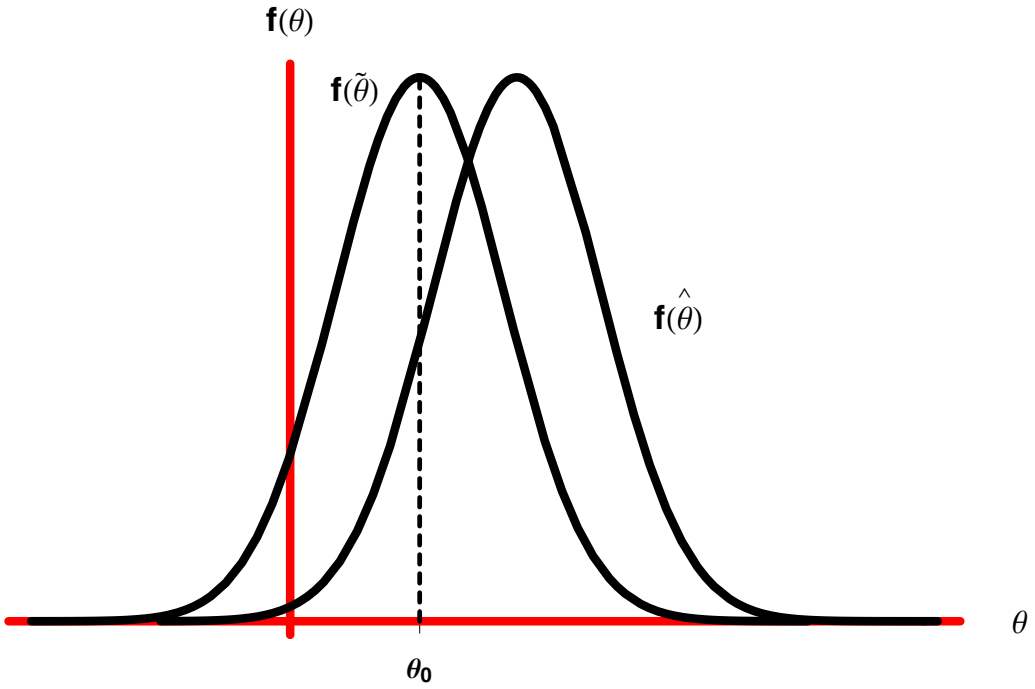
$$\begin{aligned}
MSE(\theta) \; &= E \; \left(\hat{\theta} - \theta\right)^2 \\
&= E \left(\hat{\theta} - E \; \left(\hat{\theta}\right) + E \; \left(\hat{\theta}\right) - \theta\right)^2 \\
&= E \left(\left(\hat{\theta} - E \; \left(\hat{\theta}\right)\right) + \left(E \; \left(\hat{\theta}\right) - \theta\right)\right)^2 \\
&= E \left(\hat{\theta} - E \; \left(\hat{\theta}\right)\right)^2 \; + \; 2 \; \left[E \; \left(\hat{\theta}\right) - \theta\right) \; E \; \left(\hat{\theta} \; - \; E \left(\hat{\theta}\right)\right] \; + \; \left(E \left(\hat{\theta}\right) - \theta\right)^2 \\
&= E \left(\hat{\theta} \; - \; E\,(\hat{\theta})\right)^2 \; + \; \left(E\,(\hat{\theta}) \; - \; \theta\right)^2 \; since \, E \left(\hat{\theta} \; - \; E \left(\hat{\theta}\right)\right) \; = \; 0 \\
&= Var \; \left(\hat{\theta}\right) \; + \; \left(Bias(\hat{\theta})\right)^2
\end{aligned} \tag{69}$$

4.2. **Specific properties of estimators.**

4.2.1. *Unbiasedness.* $\hat{\theta}$ is said to be an *unbiased estimator* of θ if $E\left(\hat{\theta}\right) = \theta$.

In figure 2, $\hat{\theta}$ is an unbiased estimator of θ, while $\tilde{\theta}$ is a biased estimator.

FIGURE 2. Unbiased Estimator



4.2.2. *Minimum variance.* $\hat{\theta}$ is said to be a *minimum variance estimator* of θ if

$$Var\left(\hat{\theta}\right) \leq Var\left(\tilde{\theta}\right) \tag{70}$$

where $\tilde{\theta}$ is any other estimator of θ. This criterion has its disadvantages as can be seen by noting that $\hat{\theta}$ = constant has zero variance and yet completely ignores any sample information that we may have. In figure 3, $\tilde{\theta}$ has a lower variance than $\hat{\theta}$.

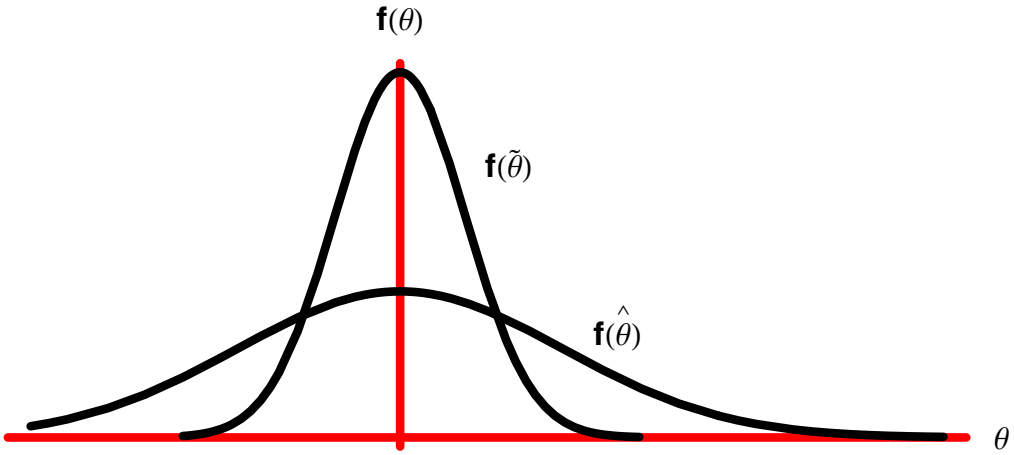
4.2.3. *Mean squared error efficient.* $\hat{\theta}$ is said to be a MSE *efficient* estimator of θ if

$$MSE\left(\hat{\theta}\right) \leq MSE\left(\tilde{\theta}\right) \tag{71}$$

where $\tilde{\theta}$ is any other estimator of θ. This criterion takes into account both the variance and bias of the estimator under consideration. Figure 4 shows three alternative estimators of θ.

4.2.4. *Best linear unbiased estimators.* $\hat{\theta}$ is the best linear unbiased estimator (BLUE) of θ if

FIGURE 3. Estimators with the Same Mean but Different Variances



$$\begin{aligned} \hat{\theta} &= \sum_{i=1}^{n} a_i \; y_i \; linear \\ E(\hat{\theta}) &= \theta \quad unbiased \\ Var(\hat{\theta}) &\leq Var(\tilde{\theta}) \end{aligned} \tag{72}$$

where $\tilde{\theta}$ is any other linear unbiased estimator of θ.

For the class of unbiased estimators of $\hat{\theta}$, the efficient estimators will also be minimum variance estimators.

4.2.5. *Example.* Let $X_1, X_2, \ldots, X_n$ denote a random sample drawn from a population having a population mean equal to μ and a population variance equal to σ^2. The sample mean (estimator of μ) is calculated by the formula

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \tag{73}$$

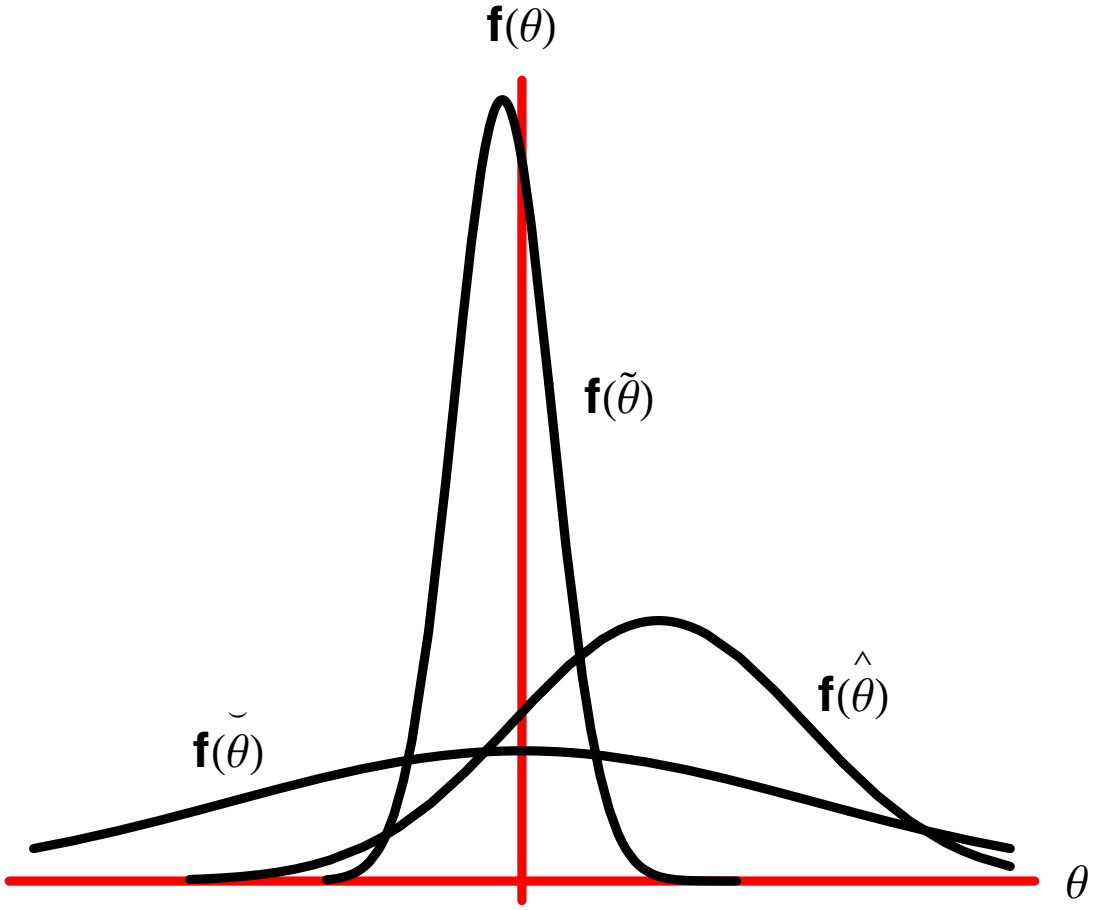
and is an unbiased estimator of μ from theorem 3 and equation 19.

Two possible estimates of the population variance are

$$\begin{aligned} \hat{\sigma}^2 &= \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n} \\ S^2 &= \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \end{aligned}$$

We have shown previously in theorem 3 and equation 21 that $\hat{\sigma}^2$ is a biased estimator of σ^2; whereas S^2 is an unbiased estimator of σ^2. Note also that

FIGURE 4. Three Alternative Estimators



$$\begin{aligned} \hat{\sigma}^2 &= \left(\frac{n-1}{n}\right) S^2 \\ E\left(\hat{\sigma}^2\right) &= \left(\frac{n-1}{n}\right) E\left(S^2\right) \\ &= \left(\frac{n-1}{n}\right) \sigma^2 \end{aligned} \tag{74}$$

Also from theorem 3 and equation 20, we have that

$$Var\left(\bar{X}\right) = \frac{\sigma^2}{n} \tag{75}$$

Now consider the mean square error of the two estimators $\bar{X}$ and S^2 where $X_1, X_2, \ldots X_n$ are a random sample from a normal population with a mean of μ and a variance of σ^2.

$$\begin{aligned} E\left(\bar{X} - \mu\right)^2 &= Var\left(\bar{X}\right) = \frac{\sigma^2}{n} \\ E\left(S^2 - \sigma^2\right)^2 &= Var\left(S^2\right) = \frac{2\,\sigma^4}{n-1} \end{aligned} \tag{76}$$

The variance of S^2 was derived in the lecture on sample moments. The variance of $\hat{\sigma}^2$ is easily computed given the variance of S^2. Specifically,

$$\begin{aligned} Var\hat{\sigma}^2 &= Var\left(\left(\frac{n-1}{n}\right)s^2\right) \\ &= \left(\frac{n-1}{n}\right)^2 Var\left(S^2\right) \\ &= \left(\frac{n-1}{n}\right)^2 \frac{2\,\sigma^4}{n-1} \\ &= \frac{2\,(n-1)\,\sigma^4}{n^2} \end{aligned} \tag{77}$$

We can compute the MSE of $\hat{\sigma}^2$ using equations 68, 74, and 77 as follows

$$\begin{aligned} MSE\,\hat{\sigma}^2 &= E\left(\hat{\sigma}^2 - \sigma\right)^2 = \frac{2\,(n-1)\,\sigma^4}{n^2} + \left[\left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2\right]^2 \\ &= \frac{2\,(n-1)\,\sigma^4}{n^2} + \left(\frac{n-1}{n}\right)^2\sigma^4 - 2\left(\frac{n-1}{n}\right)\sigma^4 + \sigma^4 \\ &= \sigma^4\left(\frac{2\,(n-1)}{n^2} + \frac{(n-1)^2}{n^2} - \frac{2\,n\,(n-1)}{n^2} + \frac{n^2}{n^2}\right) \\ &= \sigma^4\left(\frac{2\,n - 2 + n^2 - 2\,n + 1 - 2\,n^2 + 2\,n + n^2}{n^2}\right) \\ &= \sigma^4\left(\frac{2\,n-1}{n^2}\right) \end{aligned} \tag{78}$$

Now compare the MSE's of S^2 and $\hat{\sigma}^2$.

$$MSE\hat{\sigma}^2 = \sigma^4\left(\frac{2\,n-1}{n^2}\right) < \sigma^4\left(\frac{2}{n-1}\right) = MSE\;S^2 \tag{79}$$

So $\hat{\sigma}^2$ is a biased estimator of S^2 but has lower mean square error.