MATHEMATICAL STATISTICS

LECTURE NOTES

Prof. Dr. Onur KÖKSOY

MULTIVARIATE DISTRIBUTIONS

- ***Joint and Marginal Distributions***

In this section we shall be concerned first with the *bivariate case*, that is, with situations where we are interested at the same time, in a pair of random variables defined over a joint sample space. Later, we shall extend this discussion to the multivariate case, covering any finite number of random variables.

If X and Y are discrete random variables, we write the probability that X will take on the value x and Y will take on the value y as $P(X = x, Y = y)$. Thus, $P(X = x, Y = y)$ is the probability of the intersection of the events $X = x$ and $Y = y$. We can now, in the bivariate case, display the probabilities associated with all pairs of values X and Y by means of a table.

Theorem: A bivariate function, $p(x, y) = P(X = x, Y = y)$ can serve as the *joint probability distribution* of a pair of discrete random variables X and Y if and only if its values, $p(x, y)$, satisfy the conditions;

1. $p(x, y) \geq 0$ for each pair of values (x, y) within its domain;
2. $\sum_x \sum_y p(x, y) = 1$, where the double summation extends over all possible pairs (x, y) within its domain.

Definition: If X and Y are discrete random variables, the function given by,

$$F(x, y) = P(X \leq x; Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} p(s, t) \quad for - \infty < x < \infty \; ; \; -\infty < y < \infty$$

where $p(s, t)$ is the value of the joint probability distribution of X and Y at (s, t), is called the *joint distribution function*, or the *joint cumulative distribution*, of X and Y.

Example: Determine the value of k for which the function given by,

$$p(x,y) = kxy \quad , \quad for\ x = 1,2,3\ and\ y = 1,2,3$$

can serve as a joint probability distribution?

Solution: Substituting the various values of x and y, we get,

$p(1,1) = k,\ p(1,2) = 2k,\ p(1,3) = 3k, p(2,1) = 2k,\ p(2,2) = 4k,\ p(2,3) = 6k,$
$p(3,1) = 3k,\ p(3,2) = 6k$ and $p(3,3) = 9k$.

To satisfy the first condition of the Theorem, the constant k must be nonnegative, and to satisfy the second condition,

$$k + 2k + 3k + 2k + 4k + 6k + 3k + 6k + 9k = 1$$

So that $36k = 1$ and $k = \frac{1}{36}$.

Example: If the values of the joint distribution of X and Y are as shown in the table,

		x		
		0	1	2
y	0	1/6	1/3	1/12
	1	2/9	1/6	
	2	1/36		

Find $F(1,1)$.

Solution:

$$F(1,1) = P(X \le 1, Y \le 1) = p(0,0) + p(0,1) + p(1,0) + p(1,1) = \frac{1}{6} + \frac{2}{9} + \frac{1}{3} + \frac{1}{6} = \frac{8}{9}$$

Some Remarks: We also get,

$F(-2,1) = P(X \le -2, Y \le 1) = 0$

$F(3.7,4.5) = P(X \le 3.7, Y \le 4.5) = 1$

Definition: A bivariate function with values $f(x,y)$, defined over the xy-plane, is called a *joint probability density function* of the continuous random variables X and Y if and only if,

$$P\big((X,Y) \in A\big) = \iint_A f(x,y)dxdy$$

for any region A in the xy-plane.

Theorem: A bivariate function can serve as a joint probability density function of continuous random variables X and Y if its values, $f(x,y)$, satisfy the conditions,

1. $f(x,y) \geq 0 \;\; for -\infty < x < \infty \;, -\infty < y < \infty$.
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1.$

Definition: If $X \; and \; Y$ are continuous random variables, the function given by,

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)\, dsdt \;, \;\; for -\infty < x < \infty \;;\; -\infty < y < \infty$$

where $f(s,t)$ is the value of the joint probability distribution of X and Y at (s,t), is called the *joint distribution function*, of X and Y.

Note: $f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$

All definitions of this section can be generalized to the *multivariate case* where there are n random variables. Let the values of the joint probability distribution of n discrete random variables $X_1, X_2, \ldots, X_n$ are given by,

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

For each n −tuple$(x_1, x_2, \ldots, x_n)$ within the range of the random variables and so of their joint distribution are given by

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)\,,$$

$for -\infty < x_1 < \infty \;;\; -\infty < x_2 < \infty \,; \ldots; \; -\infty < x_n < \infty$

In the continuous case, probabilities are again obtained by integrating the joint probability density, and the joint distribution function is given by,

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f\,(t_1, t_2, \dots, t_n) dt_1 dt_2 \dots dt_n\,,$$

$for -\infty < x_1 < \infty\ ;\ -\infty < x_2 < \infty\,; \dots;\ -\infty < x_n < \infty$

Also practical differentiation yields,

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n F(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n}$$

wherever these partial derivatives exist.

Example: If the joint probability density of X and Y is given by,

$$f(x,y) = \begin{cases} x + y, & for\ 0 < x < 1, 0 < y < 1 \\ 0\ , & elsewhere \end{cases}$$

Find the joint distribution function of these two random variables.

Solution:

$$F(x,y) = \int_0^y \int_0^x (s+t) ds dt = \frac{1}{2} xy(x+y)\,; \quad 0 < x < 1\,,\ 0 < y < 1$$

Example: Find the joint probability density function of the two random variables X and Y whose joint distribution function is given by,

$$F(x,y) = \begin{cases} (1 - e^{-x})(1 - e^{-y})\ , & for\ 0 < x,\ 0 < y \\ 0\ , & elsewhere \end{cases}$$

Also use the joint probability density to determine $P(1 < X < 3, 1 < Y < 2)$.

Solution: Since partial differentiation yields,

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y} = e^{-(x+y)}$$

for $x > 0\ and\ y > 0$ and 0 elsewhere, we find that the joint probability density X and Y is given by,

$$f(x,y) = \begin{cases} e^{-(x+y)}, & for\ 0 < x, \quad 0 < y \\ 0, & elsewhere \end{cases}$$

Thus, integration yields,

$$P(1 < X < 3, 1 < Y < 2) = \int_1^2 \int_1^3 e^{-(x+y)} dxdy = (e^{-1} - e^{-3})(e^{-1} - e^{-2}) = 0.074$$

Example: If the joint probability distribution of three discrete random variables X, Y and Z is given by,

$$p(x,y,z) = \frac{(x+y)z}{63} \quad for\ x = 1,2\ ;\ y = 1,2,3\ ;\ z = 1,2$$

Find $P(X = 2, Y + Z \leq 3) = p(2,1,1) + p(2,1,2) + p(2,2,1) = \frac{3}{63} + \frac{6}{63} + \frac{4}{63} = \frac{13}{63}$.

Example: If the trivariate probability density of $X_1, X_2\ and\ X_3$ is given by,

$$f(x_1,x_2,x_3) = \begin{cases} (x_1 + x_2)e^{-x_3}, & for\ 0 < x_1 < 1, \quad 0 < x_2 < 1, \quad x_3 > 0 \\ 0, & elsewhere \end{cases}$$

Find $P((x_1,x_2,x_3) \in A)$ where A is the region,

$$\left\{(x_1,x_2,x_3) \middle| 0 < X_1 < \tfrac{1}{2},\ \tfrac{1}{2} < X_2 < 1, X_3 < 1\right\}.$$

Solution:

$$P\big((x_1,x_2,x_3) \in A\big) = P\left(0 < X_1 < \frac{1}{2},\ \frac{1}{2} < X_2 < 1,\ X_3 < 1\right)$$

$$= \int_0^1 \int_{\frac{1}{2}}^1 \int_0^{\frac{1}{2}} (x_1 + x_2)e^{-x_3} dx_1 dx_2 dx_3 = \int_0^1 \int_{\frac{1}{2}}^1 \left(\frac{1}{8} + \frac{x_2}{2}\right) e^{-x_3} dx_2 dx_3$$

$$= \int_0^1 \frac{1}{4} e^{-x_3} dx_3 = \frac{1}{4}(1 - e^{-1}) = 0.158$$

- ***Marginal Distributions***

Definition: If X and Y are discrete random variables and $p(x,y)$ is the value of their joint probability distribution at(x,y):

$g(x) = \sum_y p(x,y)$ for each x within the range of X : *The marginal distribution of X.*

$g(y) = \sum_x p(x,y)$ for each y within the range of Y : *The marginal distribution of Y.*

Definition: If X and Y are continuous random variables and $f(x,y)$ is the value of their joint probability density at(x,y):

$g(x) = \int_{-\infty}^{\infty} f(x,y)dy$ for each $-\infty < x < \infty$: *The marginal density of X.*

$g(y) = \int_{-\infty}^{\infty} f(x,y)dx$ for each $-\infty < y < \infty$: *The marginal density of Y.*

- ***Joint Marginal Distributions:***

Definition: If the joint probability distribution of the discrete random variables $X_1, X_2, \dots, X_n$ has the values $p(x_1, x_2, \dots, x_n)$, the marginal distribution of X_1 alone is given by,

$$g(x_1) = \sum_{x_2} \sum_{x_3} \dots \sum_{x_n} p(x_1, x_2, \dots, x_n)$$

for all values within the range of X_1, the joint marginal distribution of X_1, X_2, X_3 is given by,

$$g(x_1, x_2, x_3) = \sum_{x_4} \sum_{x_5} \dots \sum_{x_n} p(x_1, x_2, \dots, x_n)$$

for all values within the range of X_1, X_2 *and* X_3, and other marginal distributions defined in the same manner.

For the continuous case, let the continuous random variables be $X_1, X_2, \dots, X_n$ and joint density be $f(x_1, x_2, \dots, x_n)$. Then the marginal density of X_2 alone is given by,

$$h(x_2) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 dx_3 \dots dx_n \,, \quad for -\infty < x_2 < \infty \,.$$

The joint marginal density of X_1 and X_n is given by,

$$h(x_1, x_n) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f\,(x_1, x_2, \dots, x_n) dx_2 dx_3 \dots dx_{n-1}\,,$$

$for -\infty < x_1 < \infty\,, -\infty < x_n < \infty\,,$ and so forth.

Example:

		x			
	$p(x,y)$	0	1	2	$h(y)$
y	0	1/6	1/3	1/12	7/12
	1	2/9	1/6		7/18
	2	1/36			1/36
	$g(x)$	5/12	1/2	1/12	

The column totals are probabilities that X will take on the values 0, 1 and 2. In other words, they are the values

$$g(x) = \sum_{y=0}^{2} p(x,y)\ for\ x = 0,1,2$$

of the probability distribution of X. By the same taken, the row totals are the values,

$$h(y) = \sum_{x=0}^{2} p(x,y)\ for\ y = 0,1,2$$

of the probability distribution of Y.

Example: Given the joint probability density,

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y) & , \quad for\ 0 < x < 1\,, \quad 0 < y < 1 \\ 0 & , \quad elsewhere \end{cases}$$

Find the marginal densities of X and Y.

Solution: Performing the necessary integrations, we get

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{0}^{1} \frac{2}{3}(x+2y)dy = \frac{2}{3}(x+1), \qquad 0 < x < 1$$

Likewise,

$$h(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_{0}^{1} \frac{2}{3}(x+2y)dx = \frac{1}{3}(1+4y), \qquad 0 < y < 1$$

Example: Given the joint trivariate probability density,

$$f(x_1,x_2,x_3) = \begin{cases} (x_1+x_2)e^{-x_3} & , \ for\ 0 < x_1 < 1, \quad 0 < x_2 < 1\ , \quad x_3 > 0 \\ 0 & , \qquad elsewhere \end{cases}$$

Find the joint marginal density of $X_1\ and\ X_3$ and the marginal density of X_1 alone.

Solution: Performing the necessary integration, we find that the joint marginal density of $X_1\ and\ X_3$ is given by,

$$m(x_1,x_3) = \int_{0}^{1} (x_1+x_2)e^{-x_3}dx_2 = \left(x_1 + \frac{1}{2}\right)e^{-x_3},$$

$for\ 0 < x_1 < 1\ and\ x_3 > 0\ and\ m(x_1,x_3) = 0\ elsewhere.$

Using this result, we find that the marginal density of X_1 alone is given by,

$$g(x_1) = \int_{0}^{\infty}\int_{0}^{1} f(x_1,x_2,x_3)dx_2dx_3 = \int_{0}^{\infty} m(x_1,x_3)\,dx_3 = \int_{0}^{\infty} \left(x_1 + \frac{1}{2}\right)e^{-x_3}\,dx_3 = x_1 + \frac{1}{2}$$

$for\ 0 < x_1 < 1\ and\ g(x_1) = 0\ elsewhere.$

- ***Covariance and Correlation:***

Given any two random variables (X_1, X_2), the *covariance* of X_1 and X_2 is defined as,

$$Cov(X_1, X_2) = E\big((X_1 - \mu_1)(X_2 - \mu_2)\big) = E(X_1 X_2) - \mu_1 \mu_2$$

$$Cov(X_1, X_2) = \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \mu_1)(x_2 - \mu_2) f(x_1, x_2) dx_1 dx_2 \,, & if\ (X_1, X_2)\ continuous\ r.v. \\ \sum_{x_1} \sum_{x_2} (x_1 - \mu_1)(x_2 - \mu_2) p(x_1, x_2) \quad , & if\ (X_1, X_2)\ discrete\ r.v \end{cases}$$

$Here, \mu_i$ is the expected value of X_i.

In the continuous case,

$$\mu_i = \int_{-\infty}^{\infty} x_i f_i(x) dx_i$$

In the discrete case,

$$\mu_i = \sum_{x_i} x_i p_i(x_i)$$

The *correlation* between X_1 and X_2 is defined as,

$$\rho_{12} = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2}$$

where, σ_i $(i = 1,2)$ is the standard deviation of X_i.

The correlation between X_1 and X_2 is known as a measure of association and $-1 \leq \rho_{12} \leq 1$.

Example:

		x			
	$p(x,y)$	0	1	2	
y	0	1/6	1/3	1/12	7/12
	1	2/9	1/6		7/18
	2	1/36			1/36
		5/12	1/2	1/12	

Find the covariance of X and Y.

Solution:

$$E(XY) = \left(0 * 0 * \frac{1}{6}\right) + \left(0 * 1 * \frac{2}{9}\right) + \left(0 * 2 * \frac{1}{36}\right) + \left(1 * 0 * \frac{1}{3}\right) + \left(1 * 1 * \frac{1}{6}\right) + \left(2 * 0 * \frac{1}{12}\right) = \frac{1}{6}$$

$$\mu_X = E(X) = \left(0 * \frac{5}{12}\right) + \left(1 * \frac{1}{2}\right) + \left(2 * \frac{1}{12}\right) = \frac{2}{3}$$

$$\mu_Y = E(Y) = \left(0 * \frac{7}{12}\right) + \left(1 * \frac{7}{18}\right) + \left(2 * \frac{1}{36}\right) = \frac{4}{9}$$

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y = \frac{1}{6} - \left(\frac{2}{3} * \frac{4}{9}\right) = -\frac{7}{54}$$

Example: Given,

$$f(x,y) = \begin{cases} 2\,, & for\ x > 0,\ \ y > 0\,,\ \ x + y < 1 \\ 0\,, & elsewhere \end{cases}$$

Find the covariance of the random variables.

Solution: $\mu_X = \int_0^1 \int_0^{1-x} 2x dy dx = \frac{1}{3}$ *and* $\mu_Y = \int_0^1 \int_0^{1-x} 2y dy dx = \frac{1}{3}$

$$E(XY) = \int_0^1 \int_0^{1-x} 2xy dy dx = \frac{1}{12}$$

$$Cov(X,Y) = \sigma_{XY} = \frac{1}{12} - \left(\frac{1}{3} * \frac{1}{3}\right) = -\frac{1}{36}$$

✓ The following are some properties of covariance:

1. $|Cov(X,Y)| \leq \sigma_X \sigma_Y$ where σ_X and σ_Y are the standard deviations of X and Y, respectively.

2. If c is any constant, then $Cov(X,c) = 0$.

3. For any constant a and b, $Cov(aX,bY) = abCov(X,Y)$.

4. For any constants a, $b, c, and\ d$,

$$Cov(aX_1 + bX_2, cX_3 + dX_4) = acCov(X_1,X_3) + adCov(X_1,X_4) + bcCov(X_2,X_3) + bdCov(X_2,X_4)$$

Property 4 can be generalized to,

$$Cov\left(\sum_{i=1}^{m} a_i X_i\ , \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{b} a_i b_j Cov(X_i\ , Y_j)$$

From Property 1, we deduce that $-1 \leq \rho_{12} \leq 1$. The correlation obtains the values ± 1 only if the two variables are linearly dependent.

Definition: Random variable $X_1, \ldots X_k$ are said to be *mutually independent* if, for every $(x_1, \ldots, x_k)$,

$$f(x_1, \ldots, x_k) = \prod_{i=1}^{k} f_i(x_i)$$

where $f_i(x_i)$ is the marginal p.d.f of X_i.

Remark : If two random variables are independent, their correlation (or covariance) is zero. The converse is generally not true. *Zero correlation does not imply independence.*

Example: If the joint probability distribution of X and Y is given by,

		x			
	$p(x,y)$	-1	0	1	
y	-1	1/6	1/3	1/6	2/3
	0	0	0	0	0
	1	1/6	0	1/6	1/3
		1/3	1/3	1/3	

Show that their covariance is zero even though the two random variables are not independent.

Solution:

$$\mu_X = E(X) = \left((-1) * \frac{1}{3}\right) + \left(0 * \frac{1}{3}\right) + \left(1 * \frac{1}{3}\right) = 0$$

$$\mu_Y = E(Y) = \left((-1) * \frac{2}{3}\right) + (0 * 0) + \left(1 * \frac{1}{3}\right) = -\frac{1}{3}$$

$$E(XY) = \left((-1) * (-1) * \frac{1}{6}\right) + \left(0 * (-1) * \frac{1}{3}\right) + \left(1 * (-1) * \frac{1}{6}\right) + \left((-1) * 1 * \frac{1}{6}\right)\left(1 * 1 * \frac{1}{6}\right) = 0$$

Thus,$\sigma_{XY} = Cov(X, Y) = 0 - \left(0 * \left(-\frac{1}{3}\right)\right) = 0$, the covariance is zero, but the two random variables are not independent. For instance, $(x, y) \neq p(x) * p(y)\ for\ x = -1$ and $y = -1$.

The following result is very important for independent random variables.

- ✓ If $X_1, \dots X_k$ are mutually independent, then, for any integrable functions, $g_1(X_1), \dots, g_k(X_k)$,

$$E\left(\prod_{i=1}^{k} g_i(X_i)\right) = \prod_{i=1}^{k} E\big(g_i(X_i)\big)$$

Indeed,

$$\begin{aligned} E\left(\prod_{i=1}^{k} g_i(X_i)\right) &= \int \dots \int g_1(x_1) \dots g_k(x_k) f(x_1, \dots, x_k) dx_1 \dots dx_k \\ &= \int \dots \int g_1(x_1) \dots g_k(x_k) f_1(x_1) \dots f_k(x_k) dx_1 \dots dx_k \\ &= \int g_1(x_1) f_1(x_1) dx_1 \dots \int g_k(x_k)\, f_k(x_k) dx_k \\ &= \prod_{i=1}^{k} E\big(g_i(X_i)\big) \end{aligned}$$

- ***Conditional Distributions:***

Continuous Case: (X_1, X_2) continuous r.v.'s with a joint p.d.f $f(x_1, x_2)$ and also marginal p.d.f ' s are $f_1(.)$ and $f_2(.)$, respectively. Then, the conditional p.d.f of X_2 , given $(X_1 = x_1)$ where $f_1(x_1) > 0$, is defined to be,

$$f_{2.1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$$

Remark : $f_{2.1}(x_2|x_1)$ is a p.d.f.

Proof:

$$\int_{-\infty}^{\infty} f_{2.1}(x_2|x_1)dx_2 = \frac{\int_{-\infty}^{\infty} f(x_1, x_2)dx_2}{f_1(x_1)} = \frac{f_1(x_1)}{f_1(x_1)} = 1$$

The conditional expectation of X_2, given $(X_1 = x_1)$ such that $f_1(x_1) > 0$, is the expected value of X_2 with respect to the conditional p.d.f $f_{2.1}(x_2|x_1)$, that is,

$$E(X_2|X_1 = x_1) = \int_{-\infty}^{\infty} x_2 f_{2.1}(x_2|x_1)dx_2 = \mu_{X_2|x_1}$$

The conditional variance of X_2, given $(X_1 = x_1)$ such that $f_1(x_1) > 0$ is

$$Var(X_2|X_1 = x_1) = E\left(\left(X_2 - \mu_{X_2|X_1}\right)^2|X_1 = x_1\right) = E\left({X_2}^2|X_1 = x_1\right) - \left(E(X_2|X_1 = x_1)\right)^2$$

Remark: X_1 *and* X_2 independent if and only if,

$$f_{2.1}(x_2|x_1) = f_2(x_2)\ for\ all\ x_2 \quad \text{and} \quad f_{1.2}(x_1|x_2) = f_1(x_1)\ for\ all\ x_1$$

Remark: $E_X\big(E(Y|X)\big) = E(Y)$: *The law of iterated expectation*

Proof:

$$E_X\big(E(Y|X)\big) = \int E(Y|X=x)\, f_X(x)dx$$

$$= \int \left\{ \int y f_{Y.X}(y|x)dy \right\} f_X(x)dx$$

$$= \int \int y \frac{f(x,y)}{f_X(x)}\, f_X(x)dydx$$

$$= \int y \left\{ \int f(x,y)dx \right\} dy$$

$$= \int y f_Y(y)dy$$

$$= E(Y)$$

Discrete Case:

$$E(u(X)|Y=y) = \sum_X u(x)p(\,x|y)$$

$$\mathrm{P}(X|Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

provided that $P(Y=y) \neq 0$

Example: If the joint density of X and Y is given by,

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y)\,, & for\ 0<x<1, \quad 0<y<1 \\ 0 \quad , & elsewhere \end{cases}$$

Find the conditional mean and the conditional variance X given $Y = \frac{1}{2}$.

Solution: The marginal densities,

$$g(x) = \int_0^1 \frac{2}{3}(x + 2y)dy = \frac{2}{3}(x + 1) \quad for\ 0 < x < 1$$

$$h(y) = \int_0^1 \frac{2}{3}(x + 2y)dx = \frac{1}{3}(4y + 1) \quad for\ 0 < y < 1$$

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{\frac{2}{3}(x + 2y)}{\frac{1}{3}(1 + 4y)} = \frac{2x + 4y}{1 + 4y} \qquad for\ 0 < x < 1$$

So that,

$$f\left(x\middle|\frac{1}{2}\right) = \begin{cases} \frac{2}{3}(x + 1)\,, & for\ 0 < x < 1 \\ 0\ , & elsewhere \end{cases}$$

Thus,

$$E\left(X\middle|\frac{1}{2}\right) = \mu_{X|\frac{1}{2}} = \int_0^1 \frac{2}{3}x(x + 1)dx = \frac{5}{9}$$

$$E\left(X^2\middle|\frac{1}{2}\right) = \mu_{X|\frac{1}{2}} = \int_0^1 \frac{2}{3}x^2(x + 1)dx = \frac{7}{18}$$

$$Var\left(X\middle|\frac{1}{2}\right) = \frac{7}{8} - \left(\frac{5}{9}\right)^2 = \frac{13}{162}$$

- ***Linear Combinations of Random Variables:***

Let $X_1, \dots X_n$ be random variables having a joint distribution, with joint p.d.f $f(x_1, \dots, x_n)$. Let $\alpha_1, \dots \alpha_n$ be given constants. Then,

$$W = \sum_{i=1}^{n} \alpha_i X_i$$

is the linear combination of X' s. The expected value and the variance of a linear combination are as follows:

$$E(W) = \sum_{i=1}^{n} \alpha_i E(X_i)$$

and

$$Var(W) = \sum_{i=1}^{n} {\alpha_i}^2 Var(X_i) + \underbrace{\sum_{i \neq j} \sum \alpha_i \alpha_j Cov(X_i, X_j)}_{=2\sum_{i<j}\sum \alpha_i \alpha_j Cov(X_i, X_j)}$$

Example: Let $X_1, \dots X_n$ be i.i.d random variables, with common expectations μ and common finite variances σ^2. The sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a particular linear combination, with $\alpha_1 = \alpha_2 = \cdots = \alpha_n = \frac{1}{n}$. Hence,

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

(We have shown that in a random sample of n i.i.d random variables, the sample mean has the same expectation as that of the individual variables.) and, because $X_1, \dots X_n$ are mutually independent, $Cov(X_i, X_j) = 0, all\ i \neq j$. Hence,

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

(the sample variance is reduced by a factor of $\frac{1}{n}$.)

Moreover, from Chebyshev's inequality, for any $\varepsilon > 0$,

$$Pr(|\bar{X}_n - \mu| > \varepsilon) < \frac{\sigma^2}{n\varepsilon^2}$$

Therefore, because $\lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$,

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

This property is called the ***CONVERGENCE IN PROBABILITY*** of $\bar{X}_n \; to \; \mu$.

Example: If the random variable X, Y and Z have the means $\mu_X = 2$, $\mu_Y = -3$, $\mu_Z = 4$, and the variances ${\sigma_X}^2 = 1$, ${\sigma_Y}^2 = 5$, ${\sigma_Z}^2 = 2$ and the covariances $Cov(X,Y) = -2$, $Cov(X,Z) = -1$, $Cov(Y,Z) = 1$. Find the mean and variance of $W = 3X - Y + 2Z$.

Solution:

$$E(W) = E(3X - Y + 2Z) = 3E(X) - E(Y) + 2E(Z) = (3 * 2) - (-3) + (2 * 4) = 17$$

$$\begin{aligned} Var(W) &= 9Var(X) + Var(Y) + 4Var(Z) - 6Cov(X,Y) + 12Cov(X;Z) - 4Cov(Y,Z) \\ &= (9 * 1) + 5 + (2 * 4) - \big(6 * (-2)\big) + \big(12 * (-1)\big) - (4 * 1) \; = 18 \end{aligned}$$

Theorem: If $X_1, \dots X_n$ are r.v. and $Y_1 = \sum_{i=1}^{n} \alpha_i X_i$ and $Y_2 = \sum_{i=1}^{n} b_i X_i$, where $\alpha_1, \dots \alpha_n$ and $b_1, \dots b_n$ are constant, then,

$$Cov(Y_1, Y_2) = \sum_{i=1}^{n} \alpha_i \, b_i Var(X_i) + \sum_{i<j} \sum \big(\alpha_i b_j + \alpha_j b_i\big) Cov(X_i, X_j)$$

Corollary: If the r.v.'s $X_1, \dots X_n$ are independent, $Y_1 = \sum_{i=1}^{n} \alpha_i X_i$ and $Y_2 = \sum_{i=1}^{n} b_i X_i$, then,

$$Cov(Y_1, Y_2) = \sum_{i=1}^{n} \alpha_i \, b_i Var(X_i)$$

Example: If the random variables X, Y and Z have the means $\mu_X = 3$, $\mu_Y = 5$, $\mu_Z = 2$, and the variances ${\sigma_X}^2 = 8$, ${\sigma_Y}^2 = 12$, ${\sigma_Z}^2 = 18$ and the covariances $Cov(X,Y) = 1$, $Cov(X,Z) = -3, Cov(Y,Z) = 2$. Find the covariance of $W = X + 4Y + 2Z$ and $V = 3X - Y - Z$.

Solution:

$$Cov(W,V) = Cov(X + 4Y + 2Z,\ 3X - Y - Z)$$

$$= 3Var(X) - 4Var(Y) - 2Var(Z) + 11Cov(X,Y) + 5Cov(X,Z) - 6Cov(Y,Z)$$

$$= (3 * 8) - (4 * 12) - (2 * 18) + (11 * 1) + \big(5 * (-3)\big) - (6 * 2) = -76$$

- ✓ The following is a useful result:

If $X_1, \dots X_n$ are mutually independent, then the m.g.f of $T_n = \sum_{i=1}^{n} X_i$ is,

$$M_{T_n} = \prod_{i=1}^{n} M_{X_i}(t)$$

where $M_{X_i}(t) = E(e^{tX_i}) = \sum_{x_i} e^{tx_i} f(x_i)$.

Proof:

$$M_{T_n} = E\big(e^{t\sum_{i=1}^{n} X_i}\big) = E\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} E(e^{tX_i}) = \prod_{i=1}^{n} M_{X_i}(t)$$

Example: If $X_1, \dots X_n$ be independent random variables having Poisson distribution with parameters λ_i , $i = 1,2, \dots, n$, then show that their sum $T_n = \sum_{i=1}^{n} X_i$ has the Poisson distribution with parameter $\mu_n = \sum_{i=1}^{n} \lambda_i$.

Solution:

$$M_{X_i}(t) = \sum_{x_i=0}^{\infty} e^{tx_i} \frac{e^{-\lambda_i} {\lambda_i}^{x_i}}{x_i!} = e^{-\lambda_i} \sum_{x_i=0}^{\infty} \frac{(e^t \lambda_i)^{x_i}}{x_i!} = e^{-\lambda_i(1-e^t)} = \exp\big(-\lambda_i(1-e^t)\big)$$

Then,

$$M_{T_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \exp\left(-\lambda_i(1-e^t)\right) = \exp\left(-\sum_{i=1}^{n} \lambda_i(1-e^t)\right) = \exp(-\mu_n(1-e^t))$$

Example: If $X_1, \dots X_k$ be independent random variables each having binomial distribution like $B(n_i, p), i = 1,2, \dots, k$, then their sum T_k has the binomial distribution. To show this,

$$M_{T_k} = \prod_{i=1}^{k} M_{X_i}(t) = (pe^t + 1 - p)^{\sum_{i=1}^{k} n_i}$$

Solution:

$$M_{X_i}(t) = \sum_{x_i=0}^{n_i} e^{tx_i} \frac{n_i!}{x_i!\,(n_i - x_i)!} p^{x_i}(1-p)^{n_i - x_i}$$

$$= \sum_{x_i=0}^{n_i} (pe^t)^{x_i} \frac{n_i!}{x_i!\,(n_i - x_i)!} (1-p)^{n_i - x_i} = (pe^t + 1 - p)^{n_i}$$

Then,

$$M_{T_k} = \prod_{i=1}^{k} M_{X_i}(t) = (pe^t + 1 - p)^{\sum_{i=1}^{k} n_i}$$

That is, T_k is distributed like $B\left(\sum_{i=1}^{k} n_i, p\right)$.

Example: Suppose $X_1, \dots X_n$ are independent random variables and the distribution of X_i is normal $N(\mu_i, {\sigma_i}^2)$, then the distribution of $W = \sum_{i=1}^{n} \alpha_i X_i$ is normal, like that of

$$N\left(\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} {\alpha_i}^2 {\sigma_i}^2\right)$$

Example: Suppose $X_1, \dots X_n$ are independent random variables and the distribution of X_i is gamma like $G(v_i, \beta)$, then the distribution of $T_n = \sum_{i=1}^{n} X_i$ is gamma, like that of $G(\sum_{i=1}^{n} v_i, \beta)$.

$$M_{T_k} = \prod_{i=1}^{k} M_{X_i}(t) = \prod_{i=1}^{k} (1-\beta t)^{-v_i} = (1-\beta t)^{-\sum_{i=1}^{n} v_i}$$

EXERCISES

1. Let X and Y have the joint probability function

$$p(x,y)=\frac{x+y}{21}\ ,x=1,2,3\ ,y=1,2$$

a. Find the conditional probability function of X, given that $Y=y$.

b. $P(X=2|Y=2)$=?

c. Find the conditional probability function of Y, given that $X=x$.

d. Find the conditional mean $\mu_{Y|x}$ and the conditional variance ${\sigma^2}_{Y|x}$ when $x=3$.

e. Graph the joint, marginal and conditional probablity functions.

Solution:

$$p_1(x)=\sum_{y=1}^{2}\frac{x+y}{21}=\frac{2x+3}{21}\ ,\qquad x=1,2,3$$

$$p_2(y)=\sum_{x=1}^{3}\frac{x+y}{21}=\frac{3y+6}{21}\ ,\qquad y=1,2$$

a. $g(x|y)=\frac{p(x,y)}{p_2(y)}=\frac{\frac{x+y}{21}}{\frac{3y+6}{21}}=\frac{x+y}{3y+6}\ ,\ x=1,2,3,when\ y=1\ or\ 2.$

b. $P(X=2|Y=2)=g(2|2)=\frac{4}{12}=\frac{1}{3}$

c. $h(y|x)=\frac{p(x,y)}{p_1(x)}=\frac{x+y}{2x+3}\ ,\ y=1,2,\ when\ x=1,2\ or\ 3.$

d. $\mu_{Y|x}=E(Y|X=3)=\sum_{y=1}^{2}y\,h(y|3)=\sum_{y=1}^{2}y\frac{3+y}{9}=\left(1*\frac{4}{9}\right)+\left(2*\frac{5}{9}\right)=\frac{14}{9}$

$${\sigma^2}_{Y|3}=E\left(\left(Y-\frac{14}{9}\right)^2\middle|X=3\right)=\sum_{y=1}^{2}\left(y-\frac{14}{9}\right)^2\left(\frac{3+y}{9}\right)=\left(\frac{25}{81}*\frac{4}{9}\right)+\left(\frac{16}{81}*\frac{5}{9}\right)=\frac{20}{81}$$

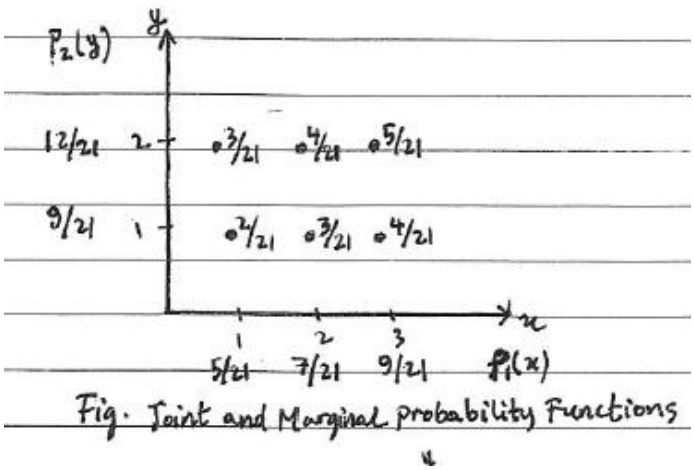
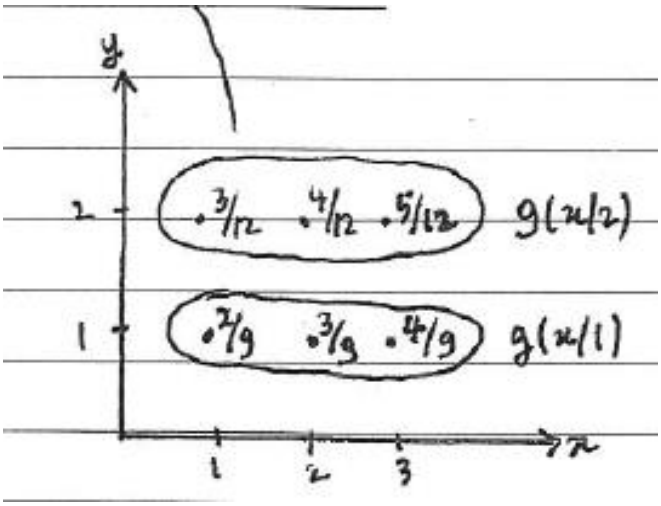
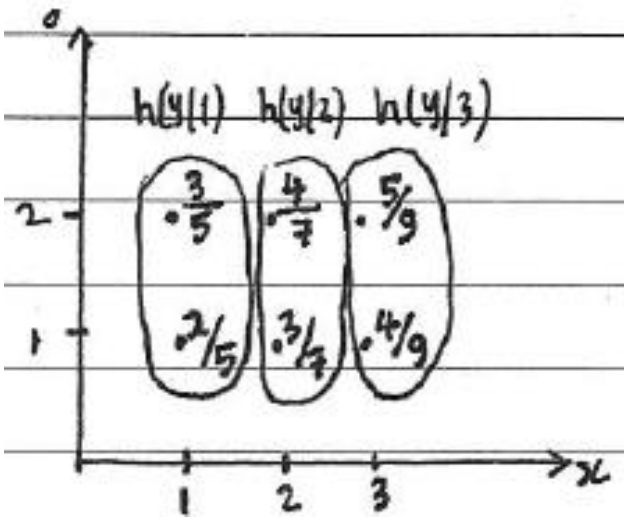


Fig. Joint and Marginal probability Functions

conditional probablity function of X, given y.



conditional probablity function of Y, given x.



2. Let X and Y have the joint probability density function

$$f(x,y) = 2\ , \qquad 0 \le\ x \le y \le 1$$

a. Find the marginal density functions.

b. $h(y|x)$=?

c. Find the conditional mean $\mu_{Y|x}$ and the conditional variance ${\sigma^2}_{Y|x}$

d. Calculate $P\left(\frac{3}{4} \le Y \le \frac{7}{8}\middle|X = \frac{1}{4}\right)$.

Solution:

a. $f_1(x) = \int_x^1 2dy = 2(1-x)\ ,\ 0 \le x \le 1$

$f_2(y) = \int_0^y 2dx = 2y \qquad ,\ 0 \le y \le 1$

b. $h(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x} \quad , x \le y \le 1\ ,\ 0 \le x \le 1.$

c. $\mu_{Y|x} = E(Y|x) = \int_x^1 y\left(\frac{1}{1-x}\right)dy = \frac{y^2}{2(1-x)}\Big|_x^1 = \frac{1+x}{2}\ ,\ 0 \le x \le 1$

${\sigma^2}_{Y|x} = E\left(\left(Y - \mu_{Y|x}\right)^2\middle|x\right) = \int_x^1 \left(y - \frac{1+x}{2}\right)^2 \frac{1}{1-x}dy = \frac{1}{3(1-x)}\left(y - \frac{1+x}{2}\right)^3\Big|_x^1 = \frac{(1-x)^2}{12}$

d. $P\left(\frac{3}{4} \le Y \le \frac{7}{8}\middle|X = \frac{1}{4}\right) = \int_{\frac{3}{4}}^{\frac{7}{8}} h\left(y\middle|\frac{1}{4}\right)dy = \int_{\frac{3}{4}}^{\frac{7}{8}} \frac{1}{3/4}\ dy = \frac{1}{6}.$

3. Let X and Y have the joint probability function

$$p(x,y) = \frac{x+2y}{18} \ , \qquad x = 1{,}2 \ , \qquad y = 1{,}2$$

a. Show that X and Y are dependent.

b. Find $Cov(X,Y)$.

c. Find the correlation coefficient ρ.

Solution:

a. The marginal probability functions are, respectively,

$p(x) = \sum_{y=1}^{2} \frac{x+2y}{18} = \frac{2x+6}{18} \ , \quad x = 1{,}2 \qquad p(y) = \sum_{x=1}^{2} \frac{x+2y}{18} = \frac{3+4y}{18} \ , \quad y = 1{,}2$

Since $(x,y) \neq p(x)p(y)$, X and Y are dependent.

b. $\mu_X = E(X) = \sum_{x=1}^{2} x \frac{2x+6}{18} = \left(1 * \frac{8}{18}\right) + \left(2 * \frac{10}{18}\right) = \frac{14}{9}$

$\mu_Y = E(Y) = \sum_{y=1}^{2} y \frac{3+4y}{18} = \left(1 * \frac{7}{18}\right) + \left(2 * \frac{11}{18}\right) = \frac{29}{18}$

$E(XY) = \sum_{y=1}^{2} \sum_{x=1}^{2} xy \frac{x+2y}{18} = \left(1 * 1 * \frac{3}{18}\right) + \left(2 * 1 * \frac{4}{18}\right) + \left(1 * 2 * \frac{5}{18}\right) + \left(2 * 2 * \frac{6}{18}\right) = \frac{45}{18}$

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y = \frac{45}{18} - \left(\frac{14}{9} * \frac{29}{18}\right) = -\frac{1}{162}$$

c. $\sigma_X^2 = Var(X) = \sum_{x=1}^{2} x^2 \frac{2x+6}{18} - \left(\frac{14}{9}\right)^2 = \frac{20}{81}$

$\sigma_Y^2 = Var(Y) = \sum_{y=1}^{2} y^2 \frac{3+4y}{18} - \left(\frac{29}{18}\right)^2 = \frac{77}{324}$

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = -\frac{\frac{1}{162}}{\sqrt{\frac{20}{81} * \frac{77}{324}}} = -0.025$$

4. "Independence implies zero correlation, but zero correlation does not necessarily imply independence."

$$X\ and\ Y\ are\ independent \rightarrow Cov(X,Y) = 0$$

$$(But\ the\ converse\ of\ this\ not\ necessarily\ true)$$

To show this, let X and Y have the joint probability function,

$$p(x,y) = \frac{1}{3}\ ,\quad (x,y) = (0,1),(1,0),(2,1)$$

Show that $Cov(X,Y) = 0$, but X and Y are dependent.

Solution:

$\mu_X = E(X) = 1$ *and* $\mu_Y = E(Y) = \frac{2}{3}$.

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y = \left(0 * 1 * \frac{1}{3}\right) + \left(1 * 0 * \frac{1}{3}\right) + \left(2 * 1 * \frac{1}{3}\right) - \left(1 * \frac{2}{3}\right) = 0$$

That is ρ=0 but X and Y are dependent.

(Show that $p(x,y) \neq p(x)p(y)$)

5. . Let X and Y have the joint probability function

$$p(x,y) = \frac{x+y}{21} \ , x = 1,2,3 \ , y = 1,2$$

Show that X and Y are dependent.

Solution:

$$p(x) = \sum_{y=1}^{2} \frac{x+y}{21} = \frac{2x+3}{21} \ , \qquad x = 1,2,3$$

$$p(y) = \sum_{x=1}^{3} \frac{x+y}{21} = \frac{6+3y}{21} \ , \qquad y = 1,2$$

Since $p(x)p(y) \neq p(x,y)$, X and Y are dependent.

6. Let the joint probability function of X and Y are,

$$p(x,y) = \frac{xy^2}{30}, \qquad x = 1,2,3 \ , \qquad y = 1,2$$

Show that X and Y are independent.

Solution:

$$p(x) = \sum_{y=1}^{2} \frac{xy^2}{30} = \frac{x}{6} \ , \qquad x = 1,2,3$$

$$p(y) = \sum_{x=1}^{3} \frac{xy^2}{30} = \frac{y^2}{5} \ , \qquad y = 1,2$$

Since $p(x)p(y) = p(x,y) \ for \ x = 1,2,3 \ and \ y = 1,2$, X and Y are independent.

7. There are eight similar chips in a bowl: three marked (0,0), two marked (1,0), two marked (0,1) and one marked (1,1). A player selects a chip at random and is given the sum of the two coordinates in dollars. If X and Y represent those two coordinates respectively, their joint probability function is

$$p(x,y) = \frac{3-x-y}{8}, \qquad x = 0,1 \ , \qquad y = 0,1$$

Find the expected pay off.

Solution:

$$E(X+Y) = \sum_{y=0}^{1}\sum_{x=0}^{1}(x+y)\frac{3-x-y}{8} = \left(0*\frac{3}{8}\right) + \left(1*\frac{2}{8}\right) + \left(1*\frac{2}{8}\right) + \left(2*\frac{1}{8}\right) = \frac{3}{4}$$

That is, the expected payoff is 75 cents.

8. Let X and Y have the joint probability density function,

$$f(x,y) = \frac{3}{2}x^2(1-|y|) \ , \qquad -1 < x < 1 \ , -1 < y < 1$$

Let $A = \{(x,y): 0 < x < 1, 0 < y < x\}$. Find the probability that (X,Y) falls into A.

Solution:

$$P\big((X,Y) \in A\big) = \int_0^1 \int_0^x \frac{3}{2}x^2(1-y)\,dydx = \int_0^1 \frac{3}{2}x^2\left(y - \frac{y^2}{2}\right)\Big|_0^x dx = \int_0^1 \frac{3}{2}\left(x^3 - \frac{x^4}{2}\right)dx$$

$$= \frac{3}{2}\left(\frac{x^4}{4} - \frac{x^5}{10}\right)\Big|_0^1 = \frac{9}{40}.$$

9. Let X and Y have the joint probability density function,

$$f(x,y) = 2\ , \qquad 0 \le x \le y \le 1$$

a. Find $P\left(0 \le X \le \frac{1}{2}, 0 \le Y \le \frac{1}{2}\right)$.

b. Find the marginal densities $f(x)$ and $f(y)$.

c. Find the $E(X)$ and $E(Y)$.

d. Show that X and Y are dependent.

Solution:

a. $P\left(0 \le X \le \frac{1}{2}, 0 \le Y \le \frac{1}{2}\right) = \int_0^{\frac{1}{2}} \int_0^y 2dxdy = \frac{1}{4}$.

b. $f(x) = \int_x^1 2dy = 2(1-x)\ ,\ 0 \le x \le 1$

$f(y) = \int_0^y 2dx = 2y\ ,\ 0 \le y \le 1$

c. $E(X) = \int_0^1 \int_x^1 2xdydx = \int_0^1 2x(1-x)dx = \frac{1}{3}$.

$E(Y) = \int_0^1 \int_0^y 2ydxdy = \int_0^1 2y^2dy = \frac{2}{3}$.

d. $f(x)f(y) \neq f(x,y)$ $\Rightarrow$ *X and Y are dependent because the support S is not a product space, since it is bounded by the diagonal line y=x.*

ORDER STATISTICS

Many functions of random variables of interest in practice depend on the relative magnitudes of the observed variables. For instance, we may be interested in the fastest time in an automobile race or the heaviest mouse among those fed on a certain diet. Thus, we often order observed random variables according to their magnitudes. The resulting ordered variables are called *order statistics*.

Formally, let $Y_1, Y_2, \ldots, Y_n$ denote independent continuous random variables with distribution function $F(y)$ and density function $f(y)$. We denote the ordered random variables Y_i by $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$, where $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$. (Because the random variables are continuous, the equality signs can be ignored.) Using this notation,

$$Y_{(1)} = \min(Y_1, Y_2, \ldots, Y_n)$$

is the minimum of the random variables Y_i, and

$$Y_{(n)} = \max(Y_1, Y_2, \ldots, Y_n)$$

is the maximum of the random variables Y_i.

The probability density functions for $Y_{(1)}$ and $Y_{(n)}$ can be found using the method of distribution functions. We will derive the density function of $Y_{(n)}$ first. Because $Y_{(n)}$ is the maximum of $Y_1, Y_2, \ldots, Y_n$, the event $(Y_{(n)} \leq y)$ will occur if and only if the events $(Y_i \leq y)$ occur for every $i = 1, 2, \ldots, n$. That is,

$$P(Y_{(n)} \leq y) = P(Y_1 \leq y, Y_2 \leq y, \ldots, Y_n \leq y).$$

Because the Y_i are independent and $P(Y_i \leq y) = F(y)$ for $i = 1, 2, \ldots, n$, it follows that the distribution function of $Y_{(n)}$ is given by

$$F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) = P(Y_1 \leq y)P(Y_2 \leq y) \cdots P(Y_n \leq y) = [F(y)]^n.$$

Letting $g_{(n)}(y)$ denote the density function of $Y_{(n)}$, we see that, on taking derivatives of both sides,

$$g_{(n)}(y) = n[F(y)]^{n-1} f(y).$$

The density function for $Y_{(1)}$ can be found in a similar manner. The distribution function of $Y_{(1)}$ is

$$F_{Y_{(1)}}(y) = P(Y_{(1)} \leq y) = 1 - P(Y_{(1)} > y).$$

Because $Y_{(1)}$ is the minimum of $Y_1, Y_2, \ldots, Y_n$, it follows that the event $(Y_{(1)} > y)$ occurs if and only if the events $(Y_i > y)$ occur for $i = 1, 2, \ldots, n$. Because the Y_i are

independent and $P(Y_i > y) = 1 - F(y)$ for $i = 1, 2, \ldots, n$, we see that

$$\begin{aligned} F_{Y_{(1)}}(y) &= P(Y_{(1)} \leq y) = 1 - P(Y_{(1)} > y) \\ &= 1 - P(Y_1 > y, Y_2 > y, \ldots, Y_n > y) \\ &= 1 - [P(Y_1 > y)P(Y_2 > y) \cdots P(Y_n > y)] \\ &= 1 - [1 - F(y)]^n. \end{aligned}$$

Thus, if $g_{(1)}(y)$ denotes the density function of $Y_{(1)}$, differentiation of both sides of the last expression yields

$$g_{(1)}(y) = n[1 - F(y)]^{n-1} f(y).$$

Remark:

joint density of $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$, which

$$g_{(1)(2)\cdots(n)}(y_1, y_2, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2), \ldots, f(y_n), & y_1 \leq y_2 \leq \cdots \leq y_n, \\ 0, & \text{elsewhere.} \end{cases}$$

Example:

Electronic components of a certain type have a length of life Y, with probability density given by

$$f(y) = \begin{cases} (1/100)e^{-y/100}, & y > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

(Length of life is measured in hours.) Suppose that two such components operate independently and in series in a certain system (hence, the system fails when either component fails). Find the density function for X, the length of life of the system.

Solution Because the system fails at the first component failure, $X = \min(Y_1, Y_2)$, where Y_1 and Y_2 are independent random variables with the given density. Then, because $F(y) = 1 - e^{-y/100}$, for $y \geq 0$,

$$\begin{aligned} f_X(y) = g_{(1)}(y) &= n[1 - F(y)]^{n-1} f(y) \\ &= \begin{cases} 2e^{-y/100}(1/100)e^{-y/100}, & y > 0, \\ 0, & \text{elsewhere,} \end{cases} \end{aligned}$$

and it follows that

$$f_X(y) = \begin{cases} (1/50)e^{-y/50}, & y > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

Thus, the minimum of two exponentially distributed random variables has an exponential distribution. Notice that the mean length of life for each component is 100 hours, whereas the mean length of life for the system is $E(X) = E(Y_{(1)}) = 50 = 100/2$. ■

Example:

Suppose that the components in Example 6.16 operate in parallel (hence, the system does not fail until both components fail). Find the density function for X, the length of life of the system.

Solution Now $X = \max(Y_1, Y_2)$, and

$$\begin{aligned} f_X(y) = g_{(2)}(y) &= n[F(y)]^{n-1} f(y) \\ &= \begin{cases} 2(1 - e^{-y/100})(1/100)e^{-y/100}, & y > 0, \\ 0, & \text{elsewhere,} \end{cases} \end{aligned}$$

and, therefore,

$$f_X(y) = \begin{cases} (1/50)(e^{-y/100} - e^{-y/50}), & y > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

We see here that the maximum of two exponential random variables is not an exponential random variable. ■

Theorem:

Let $Y_1, \ldots, Y_n$ be independent identically distributed continuous random variables with common distribution function $F(y)$ and common density function $f(y)$. If $Y_{(k)}$ denotes the kth-order statistic, then the density function of $Y_{(k)}$ is given by

$$g_{(k)}(y_k) = \frac{n!}{(k-1)!\,(n-k)!}[F(y_k)]^{k-1}\,[1-F(y_k)]^{n-k} f(y_k), \qquad -\infty < y_k < \infty.$$

If j and k are two integers such that $1 \le j < k \le n$, the joint density of $Y_{(j)}$ and $Y_{(k)}$ is given by

$$g_{(j)(k)}(y_j, y_k) = \frac{n!}{(j-1)!\,(k-1-j)!\,(n-k)!}[F(y_j)]^{j-1} \times [F(y_k)-F(y_j)]^{k-1-j} \times [1-F(y_k)]^{n-k} f(y_j)\, f(y_k), \qquad -\infty < y_j < y_k < \infty.$$

Example:

Suppose that $Y_1, Y_2, \ldots, Y_5$ denotes a random sample from a uniform distribution defined on the interval (0, 1). That is,

$$f(y) = \begin{cases} 1, & 0 \le y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the density function for the second-order statistic. Also, give the joint density function for the second- and fourth-order statistics.

Solution The distribution function associated with each of the Y's is

$$F(y) = \begin{cases} 0, & y < 0, \\ y, & 0 \le y \le 1, \\ 1, & y > 1. \end{cases}$$

The density function of the second-order statistic, $Y_{(2)}$, can be obtained directly from Theorem 6.5 with $n = 5,\ k = 2$. Thus, with $f(y)$ and $F(y)$ as noted,

$$g_{(2)}(y_2) = \frac{5!}{(2-1)!\,(5-2)!}[F(y_2)]^{2-1}[1 - F(y_2)]^{5-2} f(y_2), \quad -\infty < y_2 < \infty,$$

$$= \begin{cases} 20y_2(1 - y_2)^3, & 0 \le y_2 \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

The preceding density is a beta density with $\alpha = 2$ and $\beta = 4$. In general, the kth-order statistic based on a sample of size n from a uniform (0, 1) distribution has a beta density with $\alpha = k$ and $\beta = n - k + 1$.

The joint density of the second- and fourth-order statistics is readily obtained from the second result in Theorem . With $f(y)$ and $F(y)$ as before, $j = 2,\ k = 4$, and $n = 5$,

$$g_{(2)(4)}(y_2, y_4) = \frac{5!}{(2-1)!\,(4-1-2)!\,(5-4)!}[F(y_2)]^{2-1}[F(y_4) - F(y_2)]^{4-1-2}$$

$$\times [1 - F(y_4)]^{5-4} f(y_2)\, f(y_4), \quad -\infty < y_2 < y_4 < \infty$$

$$= \begin{cases} 5!\, y_2(y_4 - y_2)(1 - y_4), & 0 \le y_2 < y_4 \le 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Of course, this joint density can be used to evaluate joint probabilities about $Y_{(2)}$ and $Y_{(4)}$ or to evaluate the expected value of functions of these two variables. ■

FUNCTIONS OF RANDOM VARIABLES

- ***One-to-one Transformation***

Frequently in statistics, one encounters the need to derive the probability distribution of a function of one or more random variables. For example, suppose that X is a discrete random variable with probability distribution$p(x)$, and suppose further that $Y = u(X)$ defines a one-to-one transformation between the values of X and Y. We wish to find the probability distribution of Y. It is important to note that the one-to-one transformation implies that each value x is related to one, and only one, value $y = u(x)$ and that each value y is related to one, and only one, value $x = w(y)$, where $w(y)$ is obtained by solving $y = u(x)$ for x in terms of y. The random variable Y assumes the value y when X assumes the value $w(y)$. Consequently, the probability distribution of Y is given by

$$g(y) = P(Y = y) = P[X = w(y)] = p[w(y)].$$

Theorem: Suppose that X is a *discrete random variable* with probability distribution $p(x)$. Let $Y = u(X)$ define a one-to-one transformation between the values of X and Y so that the equation $Y = u(x)$ can be uniquely solved for x in terms of y, say $x = w(y)$. Then the probability distribution of Y is

$$g(y) = p[w(y)].$$

Example

Let X be a geometric random variable with probability distribution

$$f(x) = \frac{3}{4}\left(\frac{1}{4}\right)^{x-1}, \qquad x = 1, 2, 3, \ldots.$$

Find the probability distribution of the random variable $Y = X^2$.

Solution:

Since the values of X are all positive, the transformation defines a one-to-one correspondence between the x and y values, $y = x^2$ and $x = \sqrt{y}$. Hence

$$g(y) = \begin{cases} p(\sqrt{y}) = \frac{3}{4}\left(\frac{1}{4}\right)^{\sqrt{y}-1}, & y = 1, 4, 9, \ldots, \\ 0, & \text{elsewhere.} \end{cases}$$

Example:

Let X be a random variable with probability

$$p(x) = \begin{cases} \frac{1}{3}, & x = 1, 2, 3, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability distribution of the random variable $Y = 2X - 1$.

Solution:

From $y = 2x - 1$ we obtain $x = (y + 1)/2$, and given $x = 1, 2$, and 3, then

$$g(y) = p[(y + 1)/2] = 1/3, \qquad \text{for } y = 1, 3, 5.$$

Theorem:

Suppose that X_1 and X_2 are discrete random variables with joint probability distribution $p(x_1, x_2)$. Let $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ define a one-to-one transformation between the points (x_1, x_2) and (y_1, y_2) so that the equations

$$y_1 = u_1(x_1, x_2) \qquad \text{and} \qquad y_2 = u_2(x_1, x_2)$$

may be uniquely solved for x_1 and x_2 in terms of y_1 and y_2, say $x_1 = w_1(y_1, y_2)$ and $x_2 = w_2(y_1, y_2)$. Then the joint probability distribution of Y_1 and Y_2 is

$$g(y_1, y_2) = p[w_1(y_1, y_2), w_2(y_1, y_2)].$$

Theorem is extremely useful for finding the distribution of some random variable $Y_1 = u_1(X_1, X_2)$, where X_1 and X_2 are discrete random variables with joint probability distribution $p(x_1, x_2)$. We simply define a second function, say $Y_2 = u_2(X_1, X_2)$, maintaining a one-to-one correspondence between the points (x_1, x_2) and (y_1, y_2), and obtain the joint probability distribution $g(y_1, y_2)$. The distribution of Y_1 is just the marginal distribution of $g(y_1, y_2)$, found by summing over the y_2 values. Denoting the distribution of Y_1 by $h(y_1)$, we can then write

$$h(y_1) = \sum_{y_2} g(y_1, y_2).$$

Example:

Let X_1 and X_2 be two independent random variables having Poisson distributions with parameters μ_1 and μ_2, respectively. Find the distribution of the random variable $Y_1 = X_1 + X_2$.

Solution:

Since X_1 and X_2 are independent, we can write

$$p(x_1, x_2) = p(x_1)p(x_2) = \frac{e^{-\mu_1}\mu_1^{x_1}}{x_1!}\frac{e^{-\mu_2}\mu_2^{x_2}}{x_2!} = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{x_1}\mu_2^{x_2}}{x_1!x_2!},$$

where $x_1 = 0, 1, 2, \ldots$ and $x_2 = 0, 1, 2, \ldots$. Let us now define a second random variable, say $Y_2 = X_2$. The inverse functions are given by $x_1 = y_1 - y_2$ and $x_2 = y_2$.

Using Theorem , we find the joint probability distribution of Y_1 and Y_2 to be

$$g(y_1, y_2) = \frac{e^{-(\mu_1+\mu_2)}\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!},$$

where $y_1 = 0, 1, 2, \ldots$ and $y_2 = 0, 1, 2, \ldots, y_1$. Note that since $x_1 > 0$, the transformation $x_1 = y_1 - x_2$ implies that y_2 and hence x_2 must always be less than or equal to y_1. Consequently, the marginal probability distribution of Y_1 is

$$\begin{aligned} h(y_1) &= \sum_{y_2=0}^{y_1} g(y_1, y_2) = e^{-(\mu_1+\mu_2)} \sum_{y_2=0}^{y_1} \frac{\mu_1^{y_1-y_2}\mu_2^{y_2}}{(y_1 - y_2)!y_2!} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{y_2!(y_1 - y_2)!}\mu_1^{y_1-y_2}\mu_2^{y_2} \\ &= \frac{e^{-(\mu_1+\mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \binom{y_1}{y_2}\mu_1^{y_1-y_2}\mu_2^{y_2}. \end{aligned}$$

Recognizing this sum as the binomial expansion of $(\mu_1 + \mu_2)^{y_1}$ we obtain

$$h(y_1) = \frac{e^{-(\mu_1+\mu_2)}(\mu_1 + \mu_2)^{y_1}}{y_1!}, \qquad y_1 = 0, 1, 2, \ldots,$$

from which we conclude that the sum of the two independent random variables having Poisson distributions, with parameters μ_1 and μ_2, has a Poisson distribution with parameter $\mu_1 + \mu_2$.

Example

Let X_1 and X_2 be discrete random variables with

$$p(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{18}, & x_1 = 1, 2;\ x_2 = 1, 2, 3, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability distribution of the random variable $Y = X_1 X_2$.

Solution:

Let $W = X_2$. The inverse functions of $y = x_1 x_2$ and $w = x_2$ are $x_1 = y/w$ and $x_2 = w$, where $y/w = 1, 2$. Then

$$g(y, w) = (y/w)(w/18) = y/18, \quad y = 1, 2, 3, 4, 6;\ w = 1, 2, 3,\ y/w = 1, 2.$$

In tabular form the joint distribution $g(y, w)$ and marginal $h(y)$ are given by

		y				
$g(y,w)$		1	2	3	4	6
	1	1/18	2/18			
w	2		2/18		4/18	
	3			3/18		6/18
$h(y)$		1/18	2/9	1/6	2/9	1/3

$P(Y = 1) = g(1,1) = \frac{1}{18}$

$P(Y = 2) = g(2,1) + g(2,2) = \frac{2}{18} + \frac{2}{18} = \frac{2}{9}$

$P(Y = 3) = g(3,3) = \frac{3}{18} = \frac{1}{6}$

$P(Y = 4) = g(4,2) = \frac{4}{18} = \frac{2}{9}$

$P(Y = 6) = g(6,3) = \frac{6}{18} = \frac{1}{3}$

Theorem :

Suppose that X is a **continuous** random variable with probability distribution $f(x)$. Let $Y = u(X)$ define a one-to-one correspondence between the values of X and Y so that the equation $y = u(x)$ can be uniquely solved for x in terms of y, say $x = w(y)$. Then the probability distribution of Y is

$$g(y) = f[w(y)]|J|,$$

where $J = w'(y)$ and is called the **Jacobian** of the transformation.

Example

Let X be a continuous random variable with probability distribution

$$f(x) = \begin{cases} \frac{x}{12}, & 1 < x < 5, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability distribution of the random variable $Y = 2X - 3$.

Solution:

The inverse solution of $y = 2x - 3$ yields $x = (y + 3)/2$, from which we obtain $J = w'(y) = dx/dy = 1/2$. Therefore, using Theorem we find the density function of Y to be

$$g(y) = \begin{cases} \frac{(y+3)/2}{12}\left(\frac{1}{2}\right) = \frac{y+3}{48}, & -1 < y < 7, \\ 0, & \text{elsewhere.} \end{cases}$$

Example: The hospital period, in days, for patients following treatment for a certain type of kidney disorder is a random variable $Y = X + 4$, where X has the density function

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Find the probability density function of the random variable Y .

b. Using the density function of Y , find the probability that the hospital period for a patient following this treatment will exceed 8 days.

Solution:

(a) The inverse of $y = x + 4$ is $x = y - 4$, for $y > 4$, from which we obtain $|J| = 1$. Therefore,

$$g(y) = f(y - 4)|J| = 32/y^3, \quad y > 4.$$

(b) $P(Y > 8) = 32 \int_8^\infty y^{-3}\, dy = -16y^{-2}|_8^\infty = \frac{1}{4}$.

Theorem:

Suppose that X_1 and X_2 are **continuous** random variables with joint probability distribution $f(x_1, x_2)$. Let $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ define a one-to-one transformation between the points (x_1, x_2) and (y_1, y_2) so that the equations $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ may be uniquely solved for x_1 and x_2 in terms of y_1 and y_2, say $x_1 = w_1(y_l, y_2)$ and $x_2 = w_2(y_1, y_2)$. Then the joint probability distribution of Y_1 and Y_2 is

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)]|J|,$$

where the Jacobian is the 2×2 determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

and $\frac{\partial x_1}{\partial y_1}$ is simply the derivative of $x_1 = w_1(y_1, y_2)$ with respect to y_1 with y_2 held constant, referred to in calculus as the partial derivative of x_1 with respect to y_1. The other partial derivatives are defined in a similar manner.

Example:

The random variables X and Y , representing the weights of creams and coffees, respectively, in 1- kilogram boxes of chocolates containing a mixture of creams, toffees, and cordials, have the joint density function

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1, x+y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

a. Find the probability density function of the random variable $Z = X + Y$.

b. Using the density function of Z, find the probability that, in a given box, the sum of the weights of creams and toffees accounts for at least 1/2 but less than 3/4 of the total weight.

Solution:

(a) Let $W = X$. The inverse functions of $z = x + y$ and $w = x$ are $x = w$ and $y = z - w$, $0 < w < z$, $0 < z < 1$, from which we obtain

$$J = \begin{vmatrix} \frac{\partial x}{\partial w} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial w} & \frac{\partial y}{\partial z} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

Then $g(w,z) = f(w, z-w)|J| = 24w(z-w)$, for $0 < w < z$ and $0 < z < 1$. The marginal distribution of Z is

$$f_1(z) = \int_0^z 24(z-w)w\, dw = 4z^3, \qquad 0 < z < 1.$$

(b) $P(1/2 < Z < 3/4) = 4\int_{1/2}^{3/4} z^3\, dz = 65/256.$

Example:

Let X_1 and X_2 be two continuous random variables with joint probability distribution

$$f(x_1, x_2) = \begin{cases} 4x_1x_2, & 0 < x_1 < 1,\ 0 < x_2 < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the joint probability distribution of $Y_1 = X_1^2$ and $Y_2 = X_1X_2$.

Solution:

The inverse solutions of $y_1 = x_1^2$ and $y_2 = x_1x_2$ are $x_1 = \sqrt{y_1}$ and $x_2 = y_2/\sqrt{y_1}$, from which we obtain

$$J = \begin{vmatrix} 1/(2\sqrt{y_1}) & 0 \\ -y_2/2y_1^{3/2} & 1/\sqrt{y_1} \end{vmatrix} = \frac{1}{2y_1}.$$

To determine the set B of points in the y_1y_2 plane into which the set A of points in the x_1x_2 plane is mapped, we write

$$x_1 = \sqrt{y_1} \quad \text{and} \quad x_2 = y_2/\sqrt{y_1}.$$

Then setting $x_1 = 0$, $x_2 = 0$, $x_1 = 1$, and $x_2 = 1$, the boundaries of set A are transformed to $y_1 = 0$, $y_2 = 0$, $y_1 = 1$, and $y_2 = \sqrt{y_1}$, or $y_2^2 = y_1$. The two regions are illustrated in Figure .1. Clearly, the transformation is one-to-one, mapping the set $A = \{(x_1, x_2) \mid 0 < x_1 < 1,\ 0 < x_2 < 1\}$ into the set $B = \{(y_1, y_2) \mid y_2^2 < y_1 < 1,\ 0 < y_2 < 1\}$. From Theorem the joint probability distribution of Y_1 and Y_2 is

$$g(y_1, y_2) = 4(\sqrt{y_1})\frac{y_2}{\sqrt{y_1}}\frac{1}{2y_1} = \begin{cases} \frac{2y_2}{y_1}, & y_2^2 < y_1 < 1,\ 0 < y_2 < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

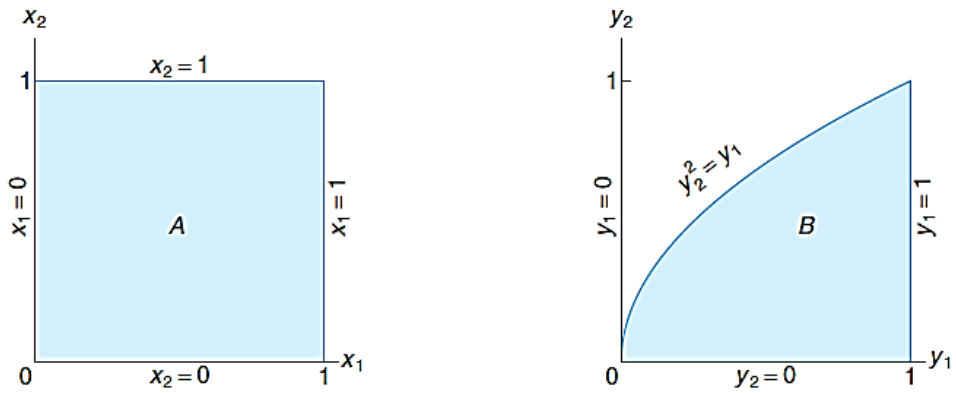


Figure 1: Mapping set A into set B.

- ***Not One-to-one Transformation***

Problems frequently arise when we wish to find the probability distribution of the random variable $Y = u(X)$ when X is a continuous random variable and the transformation is not one-to-one. That is, to each value x there corresponds exactly one value y, but to each y value there corresponds more than one x value. For example, suppose that $f(x)$ is positive over the interval $-1 < x < 2$ and zero elsewhere. Consider the transformation $y = x^2$. In this case, $x = \pm\sqrt{y}$ for $0 < y < 1$ and $x = \sqrt{y}$ for $1 < y < 4$. For the interval $1 < y < 4$, the probability distribution of Y is found as before, using Theorem . That is,

$$g(y) = f[w(y)]|J| = \frac{f(\sqrt{y})}{2\sqrt{y}}, \qquad 1 < y < 4.$$

However, when $0 < y < 1$, we may partition the interval $-1 < x < 1$ to obtain the two inverse functions

$$x = -\sqrt{y}, \qquad -1 < x < 0, \qquad \text{and} \qquad x = \sqrt{y}, \qquad 0 < x < 1.$$

Then to every y value there corresponds a single x value for each partition. From Figure 2 we see that

$$\begin{aligned} P(a < Y < b) &= P(-\sqrt{b} < X < -\sqrt{a}) + P(\sqrt{a} < X < \sqrt{b}) \\ &= \int_{-\sqrt{b}}^{-\sqrt{a}} f(x)\, dx + \int_{\sqrt{a}}^{\sqrt{b}} f(x)\, dx. \end{aligned}$$

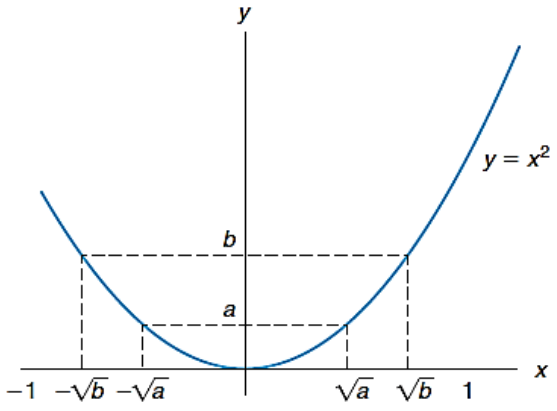


Figure .2: Decreasing and increasing function.

Changing the variable of integration from x to y, we obtain

$$P(a < Y < b) = \int_b^a f(-\sqrt{y})J_1 \, dy + \int_a^b f(\sqrt{y})J_2 \, dy$$
$$= -\int_a^b f(-\sqrt{y})J_1 \, dy + \int_a^b f(\sqrt{y})J_2 \, dy,$$

where

$$J_1 = \frac{d(-\sqrt{y})}{dy} = \frac{-1}{2\sqrt{y}} = -|J_1|$$

and

$$J_2 = \frac{d(\sqrt{y})}{dy} = \frac{1}{2\sqrt{y}} = |J_2|.$$

Hence, we can write

$$P(a < Y < b) = \int_a^b [f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2|] \, dy,$$

and then

$$g(y) = f(-\sqrt{y})|J_1| + f(\sqrt{y})|J_2| = \frac{f(-\sqrt{y}) + f(\sqrt{y})}{2\sqrt{y}}, \qquad 0 < y < 1.$$

The probability distribution of Y for $0 < y < 4$ may now be written

$$g(y) = \begin{cases} \frac{f(-\sqrt{y})+f(\sqrt{y})}{2\sqrt{y}}, & 0 < y < 1, \\ \frac{f(\sqrt{y})}{2\sqrt{y}}, & 1 < y < 4, \\ 0, & \text{elsewhere.} \end{cases}$$

Theorem:

Suppose that X is a **continuous** random variable with probability distribution $f(x)$. Let $Y = u(X)$ define a transformation between the values of X and Y that is not one-to-one. If the interval over which X is defined can be partitioned into k mutually disjoint sets such that each of the inverse functions

$$x_1 = w_1(y), \quad x_2 = w_2(y), \quad \ldots, \quad x_k = w_k(y)$$

of $y = u(x)$ defines a one-to-one correspondence, then the probability distribution of Y is

$$g(y) = \sum_{i=1}^{k} f[w_i(y)]|J_i|,$$

where $J_i = w_i'(y)$, $i = 1, 2, \ldots, k$.

Example

Show that $Y = (X-\mu)^2/\sigma^2$ has a chi-squared distribution with 1 degree of freedom when X has a normal distribution with mean μ and variance σ^2.

Solution:

Let $Z = (X-\mu)/\sigma$, where the random variable Z has the standard normal distribution

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad -\infty < z < \infty.$$

We shall now find the distribution of the random variable $Y = Z^2$. The inverse solutions of $y = z^2$ are $z = \pm\sqrt{y}$. If we designate $z_1 = -\sqrt{y}$ and $z_2 = \sqrt{y}$, then $J_1 = -1/2\sqrt{y}$ and $J_2 = 1/2\sqrt{y}$. Hence, by Theorem , we have

$$g(y) = \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{-1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-y/2} \left| \frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi}} y^{1/2-1} e^{-y/2}, \qquad y > 0.$$

Since $g(y)$ is a density function, it follows that

$$1 = \frac{1}{\sqrt{2\pi}} \int_0^\infty y^{1/2-1} e^{-y/2}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}} \int_0^\infty \frac{y^{1/2-1} e^{-y/2}}{\sqrt{2}\Gamma(1/2)}\, dy = \frac{\Gamma(1/2)}{\sqrt{\pi}},$$

the integral being the area under a gamma probability curve with parameters $\alpha = 1/2$ and $\beta = 2$. Hence, $\sqrt{\pi} = \Gamma(1/2)$ and the density of Y is given by

$$g(y) = \begin{cases} \frac{1}{\sqrt{2}\Gamma(1/2)} y^{1/2-1} e^{-y/2}, & y > 0, \\ 0, & \text{elsewhere}, \end{cases}$$

which is seen to be a chi-squared distribution with 1 degree of freedom.

Example

Let X have the probability distribution

$$f(x) = \begin{cases} \frac{2(x+1)}{9}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability distribution of the random variable $Y = X^2$.

Solution:

The inverse functions of $y = x^2$ are $x_1 = \sqrt{y}$, $x_2 = -\sqrt{y}$ for $0 < y < 1$ and $x_1 = \sqrt{y}$ for $0 < y < 4$. Now $|J_1| = |J_2| = |J_3| = 1/2\sqrt{y}$, from which we get

$$g(y) = f(\sqrt{y})|J_1| + f(-\sqrt{y})|J_2| = \frac{2(\sqrt{y}+1)}{9} \cdot \frac{1}{2\sqrt{y}} + \frac{2(-\sqrt{y}+1)}{9} \cdot \frac{1}{2\sqrt{y}} = \frac{2}{9\sqrt{y}},$$

for $0 < y < 1$ and

$$g(y) = f(\sqrt{y})|J_3| = \frac{2(\sqrt{y}+1)}{9} \cdot \frac{1}{2\sqrt{y}} = \frac{\sqrt{y}+1}{9\sqrt{y}}, \quad \text{for } 1 < y < 4.$$

- ***The CDF Technique***

Let X be a random variable with a known distribution and let g be a one-to-one, continuous function defined over the range of X. Then $Y = g(X)$ is a new random variable with a new range. How can we determine the distribution of Y?

One method is to write the cdf of Y in terms of the cdf of X. Then take the derivative of each to obtain the pdf's which determine the distribution. This method is called the *cdf technique* .

Step 1: $$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P\left(X \le g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$

Step 2: $$f_Y(y) = \frac{d(F_Y(y))}{dy} = \frac{d\left(F_X\left(g^{-1}(y)\right)\right)}{dy} = f_X\left(g^{-1}(y)\right) \times \frac{d(g^{-1}(y))}{dy}$$

(If g is not one-to-one, then we can still use the cdf technique provided we can solve for X within the expression $P(g(X) \le y)$. That is, we must be able to obtain an equation of the form $F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le h(y)) = F_X(h(y))$.)

Example 1. Let $X \sim U[10, 30]$ and let $Y = X^4$. Is Y still uniformly distributed? If not, then what is the pdf of Y? What is $E[Y]$?

Solution. We know the pdf of X is $f_X(x) = \frac{1}{20}$ for $10 \le x \le 30$. Because x^4 is strictly increasing over [10, 30], the range of Y is $[10^4, 30^4] = [10000, 810000]$.

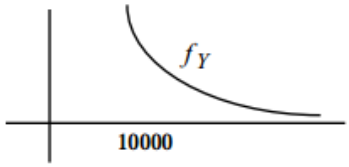
Now for $10{,}000 \le y \le 810{,}000$, we have

$$F_Y(y) = P(Y \le y) = P(X^4 \le y) = P\left(X \le y^{1/4}\right) = F_X\left(y^{1/4}\right).$$

Now we take derivatives:

$$f_Y(y) = \frac{d(F_Y(y))}{dy} = \frac{d\left(F_X\left(y^{1/4}\right)\right)}{dy} = f_X\left(y^{1/4}\right) \times \frac{d(y^{1/4})}{dy} = \frac{1}{20} \times \frac{1}{4y^{3/4}}$$

So the pdf of Y is $f_Y(y) = \frac{1}{80y^{3/4}}$ for $10{,}000 \le y \le 810{,}000$, which is *not* a uniform distribution. The values of Y are much more likely to be at the lower end of the interval.



We can compute the average of $Y = X^4$ using either the pdf of X or the pdf of Y:

$$E[X^4] = \int_{RangeX} x^4 f_X(x)\,dx = \int_{10}^{30} x^4 \frac{1}{20}\,dx = \frac{x^5}{100}\Big|_{10}^{30} = 242,000$$

or

$$E[Y] = \int_{RangeY} y f_Y(y)\,dy = \int_{10000}^{810000} y \frac{1}{80y^{3/4}}\,dy = \int_{10000}^{810000} \frac{y^{1/4}}{80}\,dy = \frac{y^{5/4}}{100}\Big|_{10000}^{810000} = 242,000$$

Example 2. Let $Z \sim N(0,1)$ and let $Y = Z^2$. Derive the pdf of Y.

Solution. We first note that $f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for *all* x. But $Y = Z^2$ has range [0, ∞). Thus for $y \geq 0$

$$F_Y(y) = P(Y \leq y) = P(Z^2 \leq y) = P\left(-\sqrt{y} \leq Z \leq \sqrt{y}\right) = 2\,P\left(0 \leq Z \leq \sqrt{y}\right)$$
$$= 2\left(P\left(Z \leq \sqrt{y}\right) - P(Z \leq 0)\right) = 2\left(P\left(Z \leq \sqrt{y}\right) - 1/2\right) = 2F_Z(\sqrt{y}) - 1$$

We now take derivatives to obtain the pdf of $Y = Z^2$:

$$f_Y(y) = \frac{d(F_Y(y))}{dy} = \frac{d\left(2F_Z(\sqrt{y}) - 1\right)}{dy} = 2f_Z\left(\sqrt{y}\right) \times \frac{d(\sqrt{y})}{dy}$$
$$= 2\frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \times \frac{1}{2y^{1/2}} = \frac{y^{-1/2} e^{-y/2}}{\sqrt{2\pi}} \quad \text{for } y \geq 0$$

Theorem. Let $Z \sim N(0,1)$ and let $Y = Z^2$. Then $Y \sim \Gamma[1/2, 1/2]$.

Proof. Recall that for $X \sim \Gamma[\alpha, \beta]$, the pdf of X is $f_X(x) = \frac{\beta^\alpha}{\Gamma[\alpha]} x^{\alpha-1} e^{-\beta x}$, for $x \geq 0$.

So the pdf of $Y \sim \Gamma[1/2, 1/2]$ is $f_Y(y) = \frac{(1/2)^{1/2}}{\Gamma[1/2]} y^{1/2-1} e^{-y/2} = \frac{y^{-1/2} e^{-y/2}}{\sqrt{2} \times \sqrt{\pi}}$, for $y \geq 0$,

which is the same pdf derived in Example 2 for Z^2 where $Z \sim N(0,1)$.

Example 3 Suppose that $Y \sim \mathcal{U}(0,1)$. Find the distribution of $U = g(Y) = -\ln Y$.

SOLUTION. The cdf of $Y \sim \mathcal{U}(0,1)$ is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y, & 0 < y < 1 \\ 1, & y \geq 1. \end{cases}$$

The support for $Y \sim \mathcal{U}(0,1)$ is $R_Y = \{y : 0 < y < 1\}$; thus, because $u = -\ln y > 0$ (sketch a graph of the log function), it follows that the support for U is $R_U = \{u : u > 0\}$. Using the method of distribution functions, we have

$$\begin{aligned} F_U(u) = P(U \leq u) &= P(-\ln Y \leq u) \\ &= P(\ln Y > -u) \\ &= P(Y > e^{-u}) = 1 - P(Y \leq e^{-u}) = 1 - F_Y(e^{-u}). \end{aligned}$$

Notice how we have written the cdf of U as a function of the cdf of Y. Because $F_Y(y) = y$ for $0 < y < 1$; i.e., for $u > 0$, we have

$$F_U(u) = 1 - F_Y(e^{-u}) = 1 - e^{-u}.$$

Taking derivatives, we get, for $u > 0$,

$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du}(1 - e^{-u}) = e^{-u}.$$

Summarizing,

$$f_U(u) = \begin{cases} e^{-u}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This is an exponential pdf with mean $\beta = 1$; that is, $U \sim$ exponential(1). □

Example 4. Suppose that $X \sim U[-1,1]$. Find the distribution $Y = exp(X)$.

Solution:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } x \in [-1,1] \\ 0 & \text{otherwise} \end{cases}$$
$$F_X(x) = \frac{1}{2} + \frac{1}{2}x, \text{ for } x \in [-1,1].$$

$$\begin{aligned} F_Y(y) &= P(\exp(X) \le y) \\ &= P(X \le \log y) \\ &= F_X(\log y) = \frac{1}{2} + \frac{1}{2}\log y, \quad \text{for } y \in [\frac{1}{e}, e]. \end{aligned}$$

Be careful about the bounds of the support!

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) \\ &= f_X(\log y)\frac{1}{y} = \frac{1}{2y}, \text{ for } y \in [\frac{1}{e}, e]. \end{aligned}$$

Example 5. Suppose that $X \sim U[-1,1]$. Find the distribution $Y = X^2$

Solution:

$$\begin{aligned} F_Y(y) &= P(X^2 \le y) \\ &= P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= 2F_X(\sqrt{y}) - 1, \text{ by symmetry: } F_X(-\sqrt{y}) = 1 - F_X(\sqrt{y}). \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) \\ &= 2f_X(\sqrt{y})\frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}, \text{ for } y \in [0,1]. \end{aligned}$$

Example 6. Suppose that $X \sim U[-1,1]$. Find the distribution $Y = 1 - X^2$

Solution:

$$\begin{aligned} F_Y(y) &= P(1 - X^2 \le y) \\ &= P(X \le -\sqrt{1-y} \cup X \ge \sqrt{1-y}) \\ &= 1 - F_X(\sqrt{1-y}) + F_X(-\sqrt{1-y}) \\ &= 2 - 2F_X(\sqrt{1-y}), \text{ by symmetry: } F_X(-\sqrt{1-y}) = 1 - F_X(\sqrt{1-y}). \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \frac{\partial}{\partial y} F_Y(y) \\ &= 2\frac{f_X(\sqrt{1-y})}{2\sqrt{1-y}} = \frac{1}{2\sqrt{1-y}}, \text{ for } y \in [0,1]. \end{aligned}$$

ESTIMATION

- ***General Remarks :***

Suppose we have a population and a random variable X which results from some measurement in the population.

e.g. : population = all people, X = IQ , population = all trees, X = lifespan

We are interested in knowing something about the distribution of X in the population. If we draw an observation, say X, then X has this distribution. If we draw n observations, X_1, X_2, …, X_n then each X_i's have this distribution. Hence X_i's are *identically distributed*. Furthermore, if the observations are random sample from the population, then the X_i's will be *independent*. Hence X_1, X_2, …, X_n are *i.i.d* random variables. (*identically and independent distributed random variables.)*

Inference : Infer something about the distribution of X in the population, based on a random sample.

Two types of inference:

a. Estimation of a population parameters .

There is ***two types of estimation:***

1. ***Point estimation:*** from the sample produce a "best guess" as to the value of a parameter. Since usually we work with continuous distributions, $P(exactly\ correct) = 0$.
2. ***Interval estimation:*** produce some interval, i.e. region, which we believe includes the true parameter value. Since this gives us an area, we can find the probability that the interval includes the parameter.

b. Hypothesis testing: checking some preconceived notion about a population parameter.

Point Estimation

- ***Introduction:***

Suppose we draw a random sample $X_1, X_2, \ldots, X_n$ from a population, $X_1, X_2, \ldots, X_n$ are iid. Let θ represent some parameter of the pdf $f(x)$. We wish to estimate θ based on the sample $X_1, X_2, \ldots, X_n$. Usually we form some function of the sample:

$$\hat{\theta}\ (X_1, X_2, \ldots, X_n) \text{ as an estimator of } \theta.$$

Now, $\hat{\theta}$ is a function of x_i's and is called the *estimator function* or *simply estimator*. Once we draw a random sample, we compute a numeric value from this function, which we call the *estimate of* θ. For example, $\bar{X}$ is an estimator of μ_X : Once we draw a random sample and compute the sample average we get a numeric value which is the estimate of μ_X.

Definition: An estimator is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample. For example, the sample mean,

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

is one possible point estimator of the population mean μ. Clearly, the expression for $\bar{Y}$ is both a rule and a formula. It tells us to sum the sample observations and divide by the sample size n.

Table 1 Expected values and standard errors of some common point estimators

Target Parameter θ	Sample Size(s)	Point Estimator $\hat{\theta}$	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
μ	n	$\overline{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^{*\dagger}$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}^{\dagger}$

*σ_1^2 and σ_2^2 are the variances of populations 1 and 2, respectively.
†The two samples are assumed to be independent.

1. Techniques for Finding Point Estimators

1.1 Method of Moments

How do we find a good estimator with desirable properties? One of the oldest methods for finding point estimators is the method of moments. This is a very simple procedure for finding an estimator for one or more population parameters. Let $\mu_k' = E[X^k]$ be the kth moment about the origin of a random variable X, whenever it exists. Let $m_k' = (1/n)\sum_{i=1}^{n} X_i^k$ be the corresponding kth sample moment. Then, the estimator of μ_k' by the method of moments is m_k'. The method of moments is based on matching the sample moments with the corresponding population (distribution) moments and is founded on the assumption that sample moments should provide good estimates of the corresponding population moments. Because the population moments $\mu_k' = h_k(\theta_1, \theta_2, \ldots, \theta_l)$ are often functions of the population parameters, we can equate corresponding population and sample moments and solve for these parameters in terms of the moments.

METHOD OF MOMENTS

Choose as estimates those values of the population parameters that are solutions of the equations $\mu_k' = m_k', k = 1, 2, \ldots, l$. Here μ_k' is a function of the population parameters.

For example, the first population moment is $\mu_1' = E(X)$, and the first sample moment is $\overline{X} = \sum_{i=1}^{n} X_i/n$. Hence, the moment estimator of μ_1' is $\overline{X}$. If $k = 2$, then the second population and sample moments are $\mu_2' = E(X^2)$ and $m_2' = (1/n)\sum_{i=1}^{n} X_i^2$, respectively. Basically, we can use the following procedure in finding point estimators of the population parameters using the method of moments.

THE METHOD OF MOMENTS PROCEDURE

Suppose there are l parameters to be estimated, say $\theta = (\theta_1, \ldots, \theta_l)$.

1. Find l population moments, $\mu_k', k = 1, 2, \ldots, l$. μ_k' will contain one or more parameters $\theta_1, \ldots, \theta_l$.
2. Find the corresponding l sample moments, $m_k', k = 1, 2, \ldots, l$. The number of sample moments should equal the number of parameters to be estimated.
3. From the system of equations, $\mu_k' = m_k', k = 1, 2, \ldots, l$, solve for the parameter $\theta = (\theta_1, \ldots, \theta_l)$; this will be a moment estimator of $\hat{\theta}$.

Example: Given that the sample random variables $X_1, X_2, \ldots, X_n$ are i.i.d taken from a Bernoulli distribution with parameter *p*.

a. Find the moment estimator for *p*.

b. Tossing a coin 10 times and equating heads to value 1 and tails to value 0, we obtained the following values:

0-1-1-0-1-0-1-1-1-0

Obtain a moment estimate for *p*, the probability of success (head).

Solution

(a) *For the Bernoulli random variable, $\mu_k' = E[X] = p$, so we can use m_1' to estimate p. Thus,*

$$m_1' = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Let

$$Y = \sum_{i=1}^{n} X_i.$$

Then, the method of moments estimator for p is $\hat{p} = Y/n$. That is, the ratio of the total number of heads to the total number of tosses will be an estimate of the probability of success.

(b) *Note that this experiment results in Bernoulli random variables. Thus, using part (a) with $Y = 6$, we get the moment estimate of p is $\hat{p} = \frac{6}{10} = 0.6$.*
We would use this value $\hat{p} = 0.6$, to answer any probabilistic questions for the given problem. For example, what is the probability of exactly obtaining 8 heads out of 10 tosses of this coin? This can be obtained by using the binomial formula, with $\hat{p} = 0.6$, that is, $P(X = 8) = \binom{10}{8}(0.6)^8(0.4)^{10-8}$.

Example: Given that the sample random variables $X_1, X_2, \ldots, X_n$ are i.i.d taken from a Gamma distribution with parameters α *and* β. Use the method of moments to estimate α and β.

Solution

For the gamma distribution

$$E[X] = \alpha\beta \quad \text{and} \quad E\left[X^2\right] = \alpha\beta^2 + \alpha^2\beta^2.$$

Because there are two parameters, we need to find the first two moment estimators. Equating sample moments to distribution (theoretical) moments, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} = \alpha\beta, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \alpha\beta^2 + \alpha^2\beta^2.$$

Solving for α *and* β *we obtain the estimates as* $\alpha = (\overline{x}/\beta)$ *and* $\beta = \left[\{(1/n)\sum_{i=1}^{n} x_i^2 - \overline{x}^2\}/\overline{x}\right]$.

Therefore, the method of moments estimators for α *and* β *are*

$$\hat{\alpha} = \frac{\overline{X}}{\hat{\beta}}$$

and

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}{\overline{X}} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n\overline{X}},$$

which implies that

$$\hat{\alpha} = \frac{\overline{X}}{\hat{\beta}} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2} = \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

Thus, we can use these values in the gamma pdf to answer questions concerning the probabilistic behavior of the r.v. X.

Example: Given that the sample random variables $X_1, X_2, \ldots, X_n$ are i.i.d taken from $N(\mu, \sigma^2)$. Use the method of moments to estimate μ and σ^2.

Solution

(a) *For the normal distribution,* $E(X) = \mu$, *and because* $Var(X) = EX^2 - \mu^2$, *we have the second moment as* $E(X^2) = \sigma^2 + \mu^2$.

Equating sample moments to distribution moments we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu_1' = \mu$$

and

$$\mu_2' = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2.$$

Solving for μ *and* σ^2, *we obtain the moment estimators as*

$$\hat{\mu} = \overline{X}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Example: Given that the sample random variables $X_1, X_2, \ldots, X_n$ are i.i.d taken from $U(a, b)$. Use the method of moments to estimate a and b.

Solution

Here, a and b are treated as parameters. That is, we only know that the sample comes from a uniform distribution on some interval, but we do not know from which interval. Our interest is to estimate this interval.

The pdf of a uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise.} \end{cases}$$

Hence, the first two population moments are

$$\mu_1 = E(X) = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2} \quad \text{and} \quad \mu_2 = E(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{a^2+ab+b^2}{3}.$$

The corresponding sample moments are

$$\hat{\mu}_1 = \overline{X} \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Equating the first two sample moments to the corresponding population moments, we have

$$\hat{\mu}_1 = \frac{a+b}{2} \quad \text{and} \quad \hat{\mu}_2 = \frac{a^2+ab+b^2}{3}$$

which, solving for a and b, results in the moment estimators of a and b,

$$\hat{a} = \hat{\mu}_1 - \sqrt{3\left(\hat{\mu}_2 - \hat{\mu}_1^2\right)} \quad \text{and} \quad \hat{b} = \hat{\mu}_1 + \sqrt{3\left(\hat{\mu}_2 - \hat{\mu}_1^2\right)}.$$

$$\hat{\mu}_1 = \frac{(a+b)}{2} \Rightarrow 2\hat{\mu}_1 = a + b \Rightarrow \hat{a} = 2\hat{\mu}_1 - \hat{b}$$

$$\hat{\mu}_2 = \frac{a^2+ab+b^2}{3} \Rightarrow 3\hat{\mu}_2 = \hat{a}^2 + \hat{a}\hat{b} + \hat{b}^2 \Rightarrow 3\hat{\mu}_2 = (2\hat{\mu}_1 - \hat{b})^2 + (2\hat{\mu}_1 - \hat{b})\hat{b} + \hat{b}^2$$

$$3\hat{\mu}_2 = 4{\hat{\mu}_1}^2 - 4\hat{\mu}_1\hat{b} + \hat{b}^2 + 2\hat{\mu}_1\hat{b} - \hat{b}^2 + \hat{b}^2$$

$$3\hat{\mu}_2 = 3{\hat{\mu}_1}^2 + \left(\hat{\mu}_1 - \hat{b}\right)^2$$

$$\sqrt{3\hat{\mu}_2 - 3{\hat{\mu}_1}^2} = \hat{\mu}_1 - \hat{b}$$

$$\hat{b} = \hat{\mu}_1 + \sqrt{3\hat{\mu}_2 - 3{\hat{\mu}_1}^2}$$

Example: Given that the sample random variables $X_1, X_2, \dots, X_n$ are i.i.d taken from Poisson distribution with parameter $\lambda > 0$. Show that both $\frac{1}{n}\sum_{i=1}^{n} X_i$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$ are moment estimators of λ.

Solution

We know that $E(X) = \lambda$, from which we have a moment estimator of λ as $(1/n)\sum_{i=1}^{n} X_i$. Also, because we have $Var(X) = \lambda$, equating the second moments, we can see that

$$\lambda = E(X^2) - (EX)^2,$$

so that

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Thus,

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Both are moment estimators of λ. Thus, the moment estimators may not be unique. We generally choose $\overline{X}$ as an estimator of λ, for its simplicity.

1.2 Method of Maximum Likelihood

It is highly desirable to have a method that is generally applicable to the construction of statistical estimators that have "good" properties. In this section we present an important method for finding estimators of parameters proposed by geneticist/statistician Sir Ronald A. Fisher around 1922 called the method of maximum likelihood. Even though the method of moments is intuitive and easy to apply, it usually does not yield "good" estimators. The method of maximum likelihood is intuitively appealing, because we attempt to find the values of the true parameters that would have most likely produced the data that we in fact observed. For most cases of practical interest, the performance of maximum likelihood estimators is optimal for large enough data. This is one of the most versatile methods for fitting parametric statistical models to data.

Definition: Let $f(x_1, x_2, \dots, x_n; \theta)$, be the joint probability(or density) function of n random variables $X_1, X_2, \dots, X_n$ with sample values $x_1, x_2, \dots, x_n$. The likelihood function of the sample is given by,

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n; \theta), (= L(\theta) in\ a\ briefer\ notation)$$

We emphasize that L is function of θ for fixed sample values.

If $X_1, \dots, X_n$ are discrete iid random variables with probability function $p(x, \theta)$, then, the likelihood function is given by

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, \dots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(X_i = x_i), \quad \text{(by multiplication rule for independent random variables)} \\ &= \prod_{i=1}^{n} p(x_i, \theta) \end{aligned}$$

and in the continuous case, if the density is $f(x, \theta)$, then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

It is important to note that the likelihood function, although it depends on the observed sample values $x = (x_1, \dots, x_n)$, is to be regarded as a function of the parameter θ. In the discrete case, $L(\theta; x_1, \dots, x_n)$ gives the probability of observing $x = (x_1, \dots, x_n)$, for a given θ. Thus, the likelihood function is a statistic, depending on the observed sample $x = (x_1, \dots, x_n)$.

Definition: The maximum likelihood estimators (MLEs) are those values of the parameters that maximize the likelihood function with respect to the parameter θ. That is,

$$L(\hat{\theta}; x_1, x_2, \dots, x_n) = max\ L(\theta; x_1, x_2, \dots, x_n) = max \prod_{i=1}^{n} f(x_i; \theta)$$

For computational convenience L will be transformed to,

$$ln\ L(\theta; x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} lnf(x_i; \theta)$$

To find $\hat{\theta}$, we need only maximize $ln\ L$ with respect to θ. In this regard, if L is a twice-differentiable function of θ, then a necessary or first-order condition for $ln\ L$ to attain a maximum at $\theta = \hat{\theta}$ is,

$$\frac{dln\ L}{d\theta}\Big|_{\theta=\hat{\theta}} = 0$$

Hence all we need to do is set $\frac{dln\ L}{d\theta} = 0$ and solve for the value of θ, $\hat{\theta}$, which makes this derivative vanish. If $\hat{\theta} = g(x_1, x_2, \dots, x_n)$ is the value of θ that maximizes *ln L, then* $\hat{\theta}$ *will be termed the maximum likelihood estimate of* θ; *it is realization of the maximum likelihood estimator* $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ and represents the parameter value *most likely* to have generated the sample realizations $x_i, i = 1, \dots, n$.

PROCEDURE TO FIND MLE

1. Define the likelihood function, $L(\theta)$.
2. Often it is easier to take the natural logarithm (ln) of $L(\theta)$.
3. When applicable, differentiate ln $L(\theta)$ with respect to θ, and then equate the derivative to zero.
4. Solve for the parameter θ, and we will obtain $\hat{\theta}$.
5. Check whether it is a maximizer or global maximizer.

Example: Suppose we have a biased coin for which one side is four times as likely to turn up on any given flip as the other. After $n = 3$ tosses we must determine if the coin is biased in favor of heads (H) or in favor of tails (T). If we define a success as getting heads on any flip of the coin, then the probability that heads occurs will be denoted as p (and thus the probability of tails on any single flip is $1 - p$). For this problem $\theta = p$ so that the Bernoulli probability function can be written as,

$$P(X = x; \theta) = p(x; \theta) = p^x(1-p)^{1-x}$$

So if heads occurs, $x = 1$ and $p(1; \theta) = p$; and if tails occurs, $x = 0$ and $p(0; \theta) = 1 - p$. Since one side of the coin is four times as likely to occur as the other, the possible values of $\theta = p$ are $\frac{1}{5}$ or $\frac{4}{5}$.

(In general, if the odds in favor of some event A occurring are a to b, then $P(A) = \frac{a}{a+b}$; and the odds against A must be $P(A') = \frac{b}{a+b}$. For our problem, since the odds of one side are4 to 1 relative to the other, we have $P(A) = \frac{1}{1+4}$. Hence $\frac{1}{5}$ is the probability of one side occurring and $\frac{4}{5}$ is the probability of the other side occurring.)

For each of these $'s$, the associated Bernoulli probability distribution are provided by Table.

	$\theta = \frac{1}{5}$	$\theta = \frac{4}{5}$
X	$P\left(X = x; \frac{1}{5}\right) = p\left(x; \frac{1}{5}\right)$	$P\left(X = x; \frac{4}{5}\right) = p\left(x; \frac{4}{5}\right)$
0	$p\left(0; \frac{1}{5}\right) = \frac{4}{5}$	$p\left(0; \frac{4}{5}\right) = \frac{1}{5}$
1	$p\left(1; \frac{1}{5}\right) = \frac{1}{5}$	$p\left(1; \frac{4}{5}\right) = \frac{4}{5}$

Suppose we toss the combination (H, T, H) so that $x_1 = 1, x_2 = 0\ and\ x_3 = 1$. Let us express the likelihood function for $n = 3$ as,

$$L(\theta; x_1, x_2, x_3) = \prod_{i=1}^{3} p(x_i; \theta) = \prod_{i=1}^{3} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{3} x_i}(1-p)^{3-\sum_{i=1}^{3} x_i}$$

Then,

$$L(\theta; 1,0,1) = p^2(1-p)$$

Clearly, the probability of the observed sample is a function of $\theta = p$.

For $\theta = \frac{1}{5}$, $L\left(\frac{1}{5}; 1,0,1\right) = \frac{1}{5}^2 \left(1 - \frac{1}{5}\right) = \frac{4}{125}$

For $\theta = \frac{4}{5}$, $L\left(\frac{4}{5}; 1,0,1\right) = \frac{4}{5}^2 \left(1 - \frac{4}{5}\right) = \frac{16}{125}$

So, if the coin is biased toward heads, $\theta = p = \frac{4}{5}$ and thus the probability of the event (H,T,H) is $\frac{16}{125}$; and if it is biased toward tails, $\theta = p = \frac{1}{5}$ and thus the probability of the event (H,T,H) is $\frac{4}{125}$. Since the maximum likelihood function is $\frac{16}{125}$, the maximum likelihood estimate of p is $\hat{p} = \frac{4}{5}$, and this estimate yields the largest a priori probability of the given event (H,T,H); that is, it is the value of p that renders the observed sample combination (H,T,H) most likely.

Example: Suppose that in n=5 drawing (with replacement) from a vessel containing a large number of red and black balls we obtain two red and three black balls. What is the best estimate of the proportion of red balls in the vessels?

Solution:

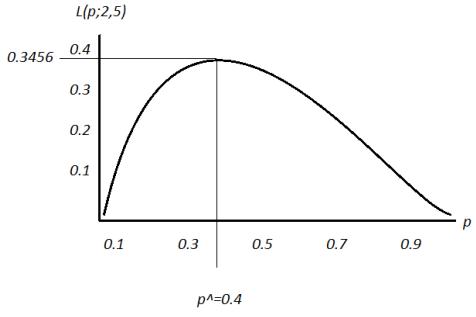
Let p (respectively, $1-p$) denote the probability of getting a red (respectively, black) ball from the vessel on any draw. Clearly the desired proportion of red balls must coincide with p. Under sampling with replacement, the various drawings yield a set of independent events and thus the probability of obtaining the given sequence of events is the product of the probabilities of the individual drawings. With two red and three black balls, the probability of the given sequence of events is $p^2(1-p)^3$. But the observed sequence of outcomes is only one way of getting two red and three black balls. The total number of ways of getting two red and three black is $\binom{5}{2} = 10$.

Hence the implied binomial probability is $B(2;5,p) = 10p^2(1-p)^3$.Since the red balls is fixed, this expression is a function of p -the likelihood function of the sample. Thus the likelihood function for the observed number of red balls is,

$$L(p;2,5) = 10p^2(1-p)^3, \;\; 0 \le p \le 1$$

Hence the maximum likelihood method has us choose the value for p, which makes the observed outcome of two red and three black balls the most probable outcome. To make our choice of p, let us perform the following experiment: we specify a whole range of possible p's and select the one that maximizes L . That is, we will chose the p, $\hat{p}$, that maximizes the probability of getting the actual sample outcome. Hence $\hat{p}$ best explains the realized sample. All this is carried out in Table.

p	$L(p;2,5)$
0.0	*0.0000*
0.1	*0.0729*
0.2	*0.2048*
0.3	*0.3087*
0.4 ($\hat{p} = 0.4$)	***0.3456 (max. value)***
0.5	*0.3125*
0.6	*0.2304*
0.7	*0.1323*
0.8	*0.0552*
0.9	*0.0081*
1	*0.0000*



Example: Given that the sample random variables $X_1, X_2, \dots, X_n$ are i.i.d taken from $Geo(p)$, $0 \leq p \leq 1$. Find MLE of p.

Solution:

For the geometric distribution, the pmf is given by

$$f(x, p) = p(1-p)^{x-1}, \quad 0 \leq p \leq 1, \quad x = 1, 2, 3, \dots.$$

Hence, the likelihood function is

$$L(p) = \prod_{i=1}^{n} \left[p\,(1-p)^{x-1} \right] = p^n\,(1-p)^{-n+\sum_{i=1}^{n} x_i}.$$

Taking the natural logarithm of $L(p)$,

$$\ln L = n \ln p + \left(-n + \sum_{i=1}^{n} x_i \right) \ln (1-p).$$

Taking the derivative with respect to p, we have

$$\frac{d \ln L}{dp} = \frac{n}{p} - \frac{\left(-n + \sum_{i=1}^{n} x_i \right)}{(1-p)}.$$

Equating $\frac{d \ln L(p)}{dp}$ to zero, we have

$$\frac{n}{p} - \frac{\left(-n + \sum_{i=1}^{n} x_i \right)}{(1-p)} = 0.$$

Solving for p,

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

Thus, we obtain a maximum likelihood estimator of p as

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}}.$$

Example: Given that the sample random variables $X_1, X_2, \dots, X_n$ are i.i.d taken from $Poisson$ distribution with parameter λ, $0 \leq p \leq 1$. Find MLE of λ.

Solution:

We have the probability mass function

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots, \quad \lambda > 0.$$

Hence, the likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}.$$

Then, taking the natural logarithm, we have

$$\ln L(\lambda) = \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \sum_{i=1}^{n} \ln (x_i!)$$

and differentiating with respect to λ results in

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

and

$$\frac{d \ln L(\lambda)}{d\lambda} = 0, \text{ implies } \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0.$$

That is,

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

Hence, the MLE of λ is

$$\hat{\lambda} = \overline{X}.$$

Example: Let $X_1, X_2, \ldots, X_n$ are i.i.d $N(\mu, \sigma^2)$ random variables.

- ***a.*** If μ is unknown and σ^2 is unknown, find MLE for μ.
- ***b.*** If μ is known and σ^2 is unknown, find MLE for σ^2 .
- ***c.*** If μ and σ^2 are both unknown, find MLE for both σ^2 *and* μ .

Solution:

a)

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

Now

$$\ln L(\mu) = -(n/2)\ln(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)^2$$

and

$$\frac{d \ln L(\mu)}{d\mu} = (\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu)$$

Equating this last result to zero and solving for μ yields

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$$

b)

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

and

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Now

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

Setting the derivative equal to zero and solving for σ^2 *we get*

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

c)

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

and

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Now

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

The solutions to the above equation yield the maximum likelihood estimators

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Example: Given that the sample random variables $X_1, X_2, \ldots, X_n$ are i.i.d taken from $G(\alpha, \beta)$. Find the MLEs for unknown parameters α *and* β.

Solution:

The pdf for the gamma distribution is given by

$$f(x) = \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, & x > 0, \quad \alpha > 0, \quad \beta > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function is given by

$$L = L(\alpha, \beta) = \frac{1}{(\Gamma(\alpha)\beta^{\alpha})^n} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\sum_{i=1}^{n} x_i/\beta}.$$

Taking the logarithms gives

$$\ln L = -n \ln \Gamma(\alpha) - n\alpha \ln \beta + (\alpha - 1) \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \frac{x}{\beta}.$$

Now taking the partial derivatives with respect to α and β and setting both equal to zero, we have

$$\frac{\partial}{\partial \alpha} \ln L = -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - n \ln \beta + \sum_{i=1}^{n} \ln x_i = 0$$

$$\frac{\partial}{\partial \beta} \ln L = -n \frac{\alpha}{\beta} + \sum_{i=1}^{n} \frac{x_i}{\beta^2} = 0.$$

Solving the second one to get β in terms of α, we have

$$\beta = \frac{\overline{x}}{\alpha}.$$

Substituting this β in the first equation, we have to solve

$$-n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - n \ln \frac{\overline{x}}{\alpha} + \sum_{i=1}^{n} \ln x_i = 0$$

for α > 0. There is no closed-form solution for α and β. In this case, one can use numerical methods such as the Newton--Raphson method to solve for α, and then use this value to find β.

1.3 Method of Least Squares

The last technique for obtaining a good point estimator of a parameter θ is the method of least squares. Least squares estimators are determined in a fashion such that desirable properties of an estimator are essentially built into them by virtue of the process by which they are constructed. In this regard, least squares estimators are *best linear unbiased estimators*. Best means that out of the class of all unbiased linear estimators of θ, the least squares estimators have minimum variance, and thus minimum mean squared error. Additionally, least squared estimators have the advantage that knowledge of the form the population probability density function is not required.

Example: Let us determine a least squares estimators for the population mean μ. Given that the sample random variables $Y_1, Y_2, \dots, Y_n$ are i.i.d, it follows that $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$ for all $i = 1,2, \dots, n$. Now, let $Y_i = \mu + \varepsilon_i$, $i = 1,2, \dots, n$, where ε_i is an observational random variable with $E(\varepsilon_i) = 0 \; and \; V(\varepsilon_i) = \sigma^2$. Under random sampling, ε_i accounts for the difference between Y_i and its mean μ.

Then principle of least squares directs us to choose μ (the $Y_i's$ are fixed, $i = 1,2, \dots, n$) so as to minimize the sum of the squared deviations between the observed Y_i values and their mean μ. That is we should choose μ so as to minimize,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \mu)^2$$

To this end we set,

$$\frac{d \sum_{i=1}^{n} \varepsilon_i^2}{d\mu} = -2 \sum_{i=1}^{n} (Y_i - \mu) = 0$$

So as to obtain $\hat{\mu} = \bar{Y}$. (Note that $\frac{d^2 \sum_{i=1}^{n} \varepsilon_i^2}{d\mu^2} = 2n > 0$ as required for a minimum.)

Hence the least squares estimator of the population mean μ is the sample mean $\bar{Y}$.

2. Properties of Point Estimators

Three different methods of finding estimators for population parameters have been introduced in the preceding section. We have seen that it is possible to have several estimators for the same parameter. For a practitioner of statistics, an important question is going to be which of many available sample statistics, such as mean, median, smallest observation, or largest observation, should be chosen to represent the entire sample? Should we use the method of moments estimator, the maximum likelihood estimator, or an estimator obtained through some other method of least squares? Now we introduce some common ways to distinguish between them by looking at some desirable properties of these estimators.

2.1 Unbiasedness:

It is desirable to have the property that the expected value of an estimator of a parameter is equal to the true value of the parameter. Such estimators are called unbiased estimators.

Definition: Let $\hat{\theta}$ be a point estimator for a parameter θ. Then $\hat{\theta}$ is an *unbiased estimator* if

$$E(\hat{\theta}) = \theta$$

If $E(\hat{\theta}) \neq \theta$, then $\hat{\theta}$ is said to be *biased estimator*. The bias of a point estimator $\hat{\theta}$ is given by

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

** *Note that bias is a constant.*

Example:

If X is a binomial random variable, show that
(a) $\hat{P} = X/n$ is an unbiased estimator of p;
(b) $P' = \frac{X+\sqrt{n}/2}{n+\sqrt{n}}$ is a biased estimator of p.

Solution:

(a) $E(X) = np$; $E(\hat{P}) = E(X/n) = E(X)/n = np/n = p$.
(b) $E(P') = \frac{E(X)+\sqrt{n}/2}{n+\sqrt{n}} = \frac{np+\sqrt{n}/2}{n+\sqrt{n}} \neq p$.

Example: Let $X_1, X_2, \dots, X_n$ depicts a set of *i.i.d* sample random variables taken from a Bernoulli population with parameter p. Show that the method of moment estimator is also an unbiased estimator.

Solution:

We can verify that the moment estimator of p is

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{Y}{n}.$$

Because for binomial random variables, $E(Y) = np$, it follows that

$$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{1}{n} \cdot np = p.$$

Hence, $\hat{p} = Y/n$ is an unbiased estimator for p.

■

Example: Let $X_1, X_2, \dots, X_n$ depicts a set of i.i.d sample random variables taken from a population with finite mean μ. Show that the sample mean $\bar{X}$ and $\frac{1}{3}\bar{X} + \frac{2}{3}X_1$ are both unbiased estimators of μ.

Solution:

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

$$E\left(\frac{1}{3}\bar{X} + \frac{2}{3}X_1\right) = E\left(\frac{1}{3}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \frac{2}{3}X_1\right) = \frac{1}{3n}\sum_{i=1}^{n} E(X_i) + \frac{2}{3}E(X_1) = \frac{1}{3n}n\mu + \frac{2}{3}\mu = \mu$$

Hence, sample mean $\bar{X}$ and $\frac{1}{3}\bar{X} + \frac{2}{3}X_1$ are both unbiased estimators of μ.

Example:

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ. Show that

$$\hat{\theta}_3 = a\hat{\theta}_1 + (1-a)\hat{\theta}_2,\ 0 \le a \le 1$$

is an unbiased estimator of θ. Note that $\hat{\theta}_3$ is a convex combination of $\hat{\theta}_1$ and $\hat{\theta}_2$. In addition, assume that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, and $Var(\hat{\theta}_1) = \sigma_1^2$ and $Var(\hat{\theta}_2) = \sigma_2^2$. How should the constant a be chosen in order to minimize the variance of $\hat{\theta}_3$?

Solution:

We are given that $E(\hat{\theta}_1) = \theta$ *and* $E(\hat{\theta}_2) = \theta$. *Therefore,*

$$E(\hat{\theta}_3) = E\left[a\hat{\theta}_1 + (1-a)\hat{\theta}_2\right] = aE\hat{\theta}_1 + (1-a)E\hat{\theta}_2$$
$$= a\theta + (1-a)\theta = \theta.$$

Hence $\hat{\theta}_3$ *is unbiased. By independence,*

$$Var(\hat{\theta}_3) = Var\left[a\hat{\theta}_1 + (1-a)\hat{\theta}_2\right]$$
$$= a^2 Var(\hat{\theta}_1) + (1-a)^2 Var(\hat{\theta}_2)$$
$$= a^2\sigma_1^2 + (1-a)^2\sigma_2^2.$$

To find the minimum,

$$\frac{d}{da}Var(\hat{\theta}_3) = 2a\sigma_1^2 - 2(1-a)\sigma_2^2 = 0,$$

gives us

$$a = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Because $\frac{d^2}{da^2}V(\hat{\theta}_3) = 2\sigma_1^2 + 2\sigma_2^2 > 0$, $V(\hat{\theta}_3)$ *has a minimum at this value of 'a'. Thus, if* $\sigma_1^2 = \sigma_2^2$, *then* $a = 1/2$.

—

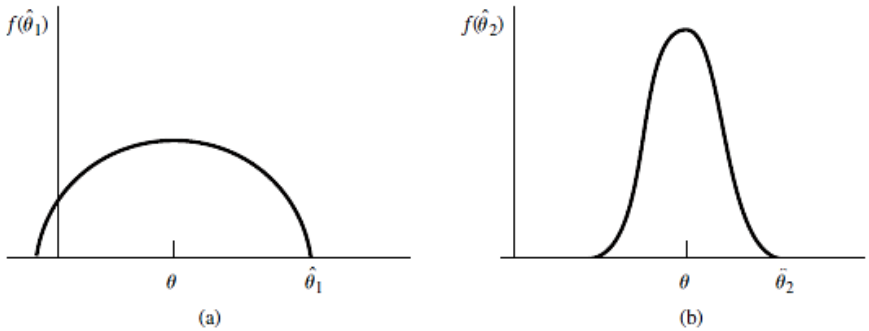
✓ For instance, if the i.i.d sample random variables, X_i , $i = 1,2,\dots,n$, each have the same mean μ and σ^2 as the population random variable X, then:

- If $\hat{\theta} = \bar{X}$, then $E(\bar{X}) = \mu$ for any population distribution. *Hence, the sample mean is an unbiased estimator of the population mean.*
- If a statistic $\hat{\theta} = \sum_{i=1}^{n} a_i X_i$ is a linear combination of the sample random variables X_i , $i = 1,2,\dots,n$, then, $E(\hat{\theta}) = E(\sum_{i=1}^{n} a_i X_i) = \mu \sum_{i=1}^{n} a_i$. If $\sum_{i=1}^{n} a_i = 1$, *then* $\hat{\theta}$ *is an unbiased estimator of* μ.
- For X a binomial random variable, if $\hat{\theta} = \hat{p} = \frac{X}{n}$, then $E(\hat{p}) = p$. *Hence the sample proportion serves as unbiased estimator of population proportion* p.
- If $\hat{\theta}$ is an unbiased estimator of a parameter θ, it does not follow that every function of $\hat{\theta}$, $h(\hat{\theta})$, is an unbiased estimator of $h(\theta)$.

✓ *As we have seen, there can be many unbiased estimators of a parameter* θ. *Which one of these estimators can we choose?*

- *If we have to choose an unbiased estimator, it will be desirable to choose the one with the least variance.* Figure 1 shows sampling distributions for unbiased point estimators for θ. We would prefer that our estimator have the type of distribution indicated in Figure 1(b) because the smaller variance guarantees that in repeated sampling a higher fraction of values of $\hat{\theta}_2$ will close to θ.

FIGURE 1
Sampling distributions for two unbiased estimators: (a) estimator with large variation; (b) estimator with small variation



- *If an estimator is biased, then we should prefer the one with low bias as well as low variance.* Generally, it is better to have an estimator that has low bias as well as low variance. Rather than using the bias and variance of a point

estimator to characterize its goodness, we might employ $E\left(\left(\hat{\theta}-\theta\right)^2\right)$, the average of the square of the distance between the estimator and its target parameter.

This leads us to the following definition,

Definiton: The mean square error of a point estimator $\hat{\theta}$ *is*,

$$MSE(\hat{\theta}) = E\left(\left(\hat{\theta}-\theta\right)^2\right)$$

The mean square error of an estimator $\hat{\theta}$, $MSE(\hat{\theta})$, is a function of both its variance and its bias. If $B(\hat{\theta})$ denotes the bias of the estimator $\hat{\theta}$, it can be shown that,

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left(B(\hat{\theta})\right)^2$$

Through the following calculations, we will now show that the MSE is a measure that combines both bias and variance.

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

because letting $B = E(\hat{\theta}) - \theta$, we get

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

B is called the *bias* of the estimator. Also, $E(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) = 0$.

Because the bias is zero for unbiased estimators, it is clear that $MSE(\hat{\theta}) = Var(\hat{\theta})$. Mean square error measures, on average, how close an estimator comes to the true value of the parameter. Hence, this could be used as a criterion for determining when one estimator is "better" than another. However, in general, it is difficult to find $\hat{\theta}$ to minimize $MSE(\hat{\theta})$. For this reason, most of the time, we look only at unbiased estimators in order to minimize $Var(\hat{\theta})$. This leads to the following definition.

Definiton: *The unbiased estimator* $\hat{\theta}$ *that minimizes the mean square error is called the* ***minimum variance unbiased estimator (MVUE)*** of θ.

Consider two unbiased estimator, $\hat{\theta}_1$ and $\hat{\theta}_2$ of population parameter θ. Then, $MSE(\hat{\theta}_1) = Var(\hat{\theta}_1)$ and $MSE(\hat{\theta}_2) = Var(\hat{\theta}_2)$. If $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$, then we can state that $\hat{\theta}_1$ is a minimum variance unbiased estimator of θ.

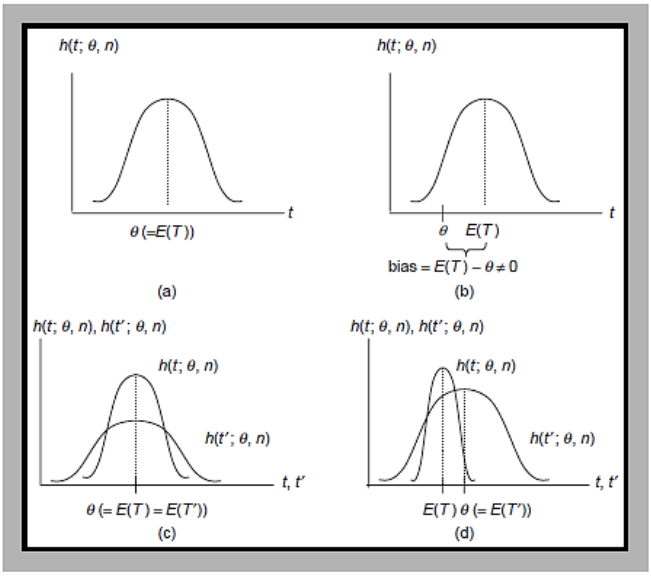


Figure (a) T is an unbiased estimator of θ; (b) T is a biased estimator of θ; (c) Both T and T' are an unbiased estimator of θ but $V(T) < V(T')$; (d) A biased estimator T may be preferable to an unbiased estimator T' of θ.

Figure 1(a): $E(T) = \theta : T$ is an unbiased estimator of θ, its sampling distribution is centered exactly on θ.

Figure 1(b): $E(T) \neq \theta : T$ is a biased estimator of θ, its sampling distribution is not centered on θ.

Figure 1(c): $E(T) = E(T') = \theta$, $Var(T) < Var(T') : T \; and \; T'$ are unbiased estimators of θ, its sampling distribution is centered on θ. It seems preferable to choose the one with the smallest variance since using T yields a higher probability of obtaining an estimate that is closer to θ.

Figure 1(d): $E(T) \neq \theta$, $E(T') = \theta$, $Var(T) < Var(T') : T$ is a biased estimator of θ and the bias is small, then T may be preferable to an unbiased estimator T' whose variance is large.

Example:

Let X_1, X_2, X_3 be a sample of size $n = 3$ from a distribution with unknown mean μ, $-\infty < \mu < \infty$, where the variance σ^2 is a known positive number. Show that both $\hat{\theta}_1 = \overline{X}$ and $\hat{\theta}_2 = [(2X_1 + X_2 + 5X_3)/8]$ are unbiased estimators for μ. Compare the variances of $\hat{\theta}_1$ and $\hat{\theta}_2$.

Solution

We have

$$E(\hat{\theta}_1) = E(\overline{X}) = \frac{1}{3} \cdot 3\mu = \mu,$$

and

$$\begin{aligned} E(\hat{\theta}_2) &= \frac{1}{8}[2EX_1 + EX_2 + 5EX_3] \\ &= \frac{1}{8}[2\mu + \mu + 5\mu] = \mu. \end{aligned}$$

Hence, both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimators.
However,

$$Var(\hat{\theta}_1) = \frac{\sigma^2}{3},$$

whereas

$$\begin{aligned} Var(\hat{\theta}_2) &= Var\left(\frac{2X_1 + X_2 + 5X_3}{8}\right) \\ &= \frac{4}{64}\sigma^2 + \frac{1}{64}\sigma^2 + \frac{25}{64}\sigma^2 = \frac{30}{64}\sigma^2. \end{aligned}$$

Because $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$, we see that $\overline{X}$ is a better unbiased estimator in the sense that the variance of $\overline{X}$ is smaller.

—

Example : Let $Y_1, Y_2, \ldots, Y_n$ depicts a set of i.i.d sample random variables taken from a population with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$. Show that,

$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is a biased estimator for σ^2 and that,

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is a unbiased estimator for σ^2 .

Solution :

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2.$$

Hence,

$$E\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] = E\left(\sum_{i=1}^{n} Y_i^2\right) - nE(\overline{Y}^2) = \sum_{i=1}^{n} E(Y_i^2) - nE(\overline{Y}^2).$$

Notice that $E(Y_i^2)$ is the same for $i = 1, 2, \ldots, n$. We use this and the fact that the variance of a random variable is given by $V(Y) = E(Y^2) - [E(Y)]^2$ to conclude that $E(Y_i^2) = V(Y_i) + [E(Y_i)]^2 = \sigma^2 + \mu^2$, $E(\overline{Y}^2) = V(\overline{Y}) + [E(\overline{Y})]^2 = \sigma^2/n + \mu^2$, and that

$$\begin{aligned} E\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] &= \sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= n\sigma^2 - \sigma^2 = (n-1)\sigma^2. \end{aligned}$$

It follows that

$$E(S'^2) = \frac{1}{n}E\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] = \frac{1}{n}(n-1)\sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2$$

and that S'^2 is biased because $E(S'^2) \neq \sigma^2$. However,

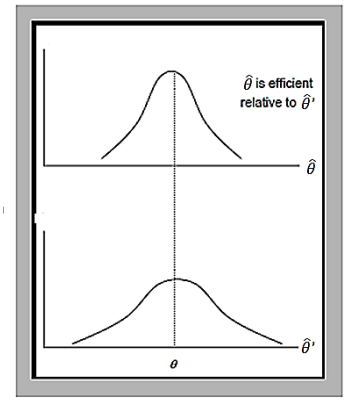
$$E(S^2) = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2,$$

so we see that S^2 is an unbiased estimator for σ^2.

$$Bias(S'^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

2.2. Efficiency:

We have seen that there can be more than one unbiased estimator for a parameter θ. We have also mentioned that the one with the least variance is desirable. Here, we introduce the concept of efficiency. If there are two unbiased estimators, it is desirable to have the one with a smaller variance. Specifically, $\hat{\theta}$ is termed an *efficient* (alternatively *minimum variance unbiased* or *best unbiased*) *estimator* of θ if $\hat{\theta}$ is unbiased and $V(\hat{\theta}) \leq V(\hat{\theta}')$ for all possible values of θ, where $\hat{\theta}'$ is any other unbiased estimator of θ.



$\hat{\theta}$ relative to $\hat{\theta}'$ is an efficient or best unbiased estimator of θ.

Hence $\hat{\theta}$ is more efficient than $\hat{\theta}'$ since the sampling distribution of $\hat{\theta}$ is more closely concentrated about θ than the sampling distribution of $\hat{\theta}'$.

How do we determine whether or not a given estimator is efficient?

Definition: If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of population parameter θ, the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is the ratio,

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two biased estimators of population parameter θ, the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is the ratio,

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}$$

If $Var(\hat{\theta}_2) > Var(\hat{\theta}_1)$ $\left(MSE(\hat{\theta}_2) > MSE(\hat{\theta}_1)\right)$, or equivalently, $e(\hat{\theta}_1, \hat{\theta}_2) > 1$, then $\hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$.

Example:

Suppose that $\hat{\Theta}_1$, $\hat{\Theta}_2$, and $\hat{\Theta}_3$ are estimators of θ. We know that

$E(\hat{\Theta}_1) = E(\hat{\Theta}_2) = \theta, E(\hat{\Theta}_3) \neq \theta, V(\hat{\Theta}_1) = 12, V(\hat{\Theta}_2) = 10$, and $E(\hat{\Theta}_3 - \theta)^2 = 6$.

Compare these three estimators. Which do you prefer? Why?

Solution:

$E(\hat{\Theta}_1)=\theta$	No bias	$V(\hat{\Theta}_1)=12=MSE(\hat{\Theta}_1)$
$E(\hat{\Theta}_2)=\theta$	No bias	$V(\hat{\Theta}_2)=10=MSE(\hat{\Theta}_2)$
$E(\hat{\Theta}_3)\neq\theta$	Bias	$MSE(\hat{\Theta}_3)=6$ [note that this includes (bias^2)]

To compare the three estimators, calculate the relative efficiencies:

$\frac{MSE(\hat{\Theta}_1)}{MSE(\hat{\Theta}_2)}=\frac{12}{10}=1.2$, since rel. eff. > 1 use $\hat{\Theta}_2$ as the estimator for θ

$\frac{MSE(\hat{\Theta}_1)}{MSE(\hat{\Theta}_3)}=\frac{12}{6}=2$, since rel. eff. > 1 use $\hat{\Theta}_3$ as the estimator for θ

$\frac{MSE(\hat{\Theta}_2)}{MSE(\hat{\Theta}_3)}=\frac{10}{6}=1.8$, since rel. eff. > 1 use $\hat{\Theta}_3$ as the estimator for θ

Conclusion:

$\hat{\Theta}_3$ is the most efficient estimator with bias, but it is biased. $\hat{\Theta}_2$ is the best "unbiased" estimator.

Example:

Let $X_1, \ldots, X_n, n > 3$, be a random sample from a population with a true mean μ and variance σ^2. Consider the following three estimators of μ:

$$\hat{\theta}_1 = \frac{1}{3}(X_1 + X_2 + X_3),$$

$$\hat{\theta}_2 = \frac{1}{8}X_1 + \frac{3}{4(n-2)}(X_2 + \cdots + X_{n-1}) + \frac{1}{8}X_n,$$

and

$$\hat{\theta}_3 = \overline{X}.$$

(a) Show that each of the three estimators is unbiased.

(b) Find $e(\hat{\theta}_2, \hat{\theta}_1)$, $e(\hat{\theta}_3, \hat{\theta}_1)$, and $e(\hat{\theta}_3, \hat{\theta}_2)$. Which of the three estimators is more efficient?

Solution:

(a) *Given* $E(X_i) = \mu, i = 1, 2, \ldots, n$. *Then,*

$$E(\hat{\theta}_1) = \frac{1}{3}[E(X_1) + E(X_2) + E(X_3)] = \frac{3\mu}{3} = \mu$$

$$E(\hat{\theta}_2) = \frac{1}{8}E(X_1) + \frac{3}{4(n-2)}(E(X_2) + \cdots + E(X_{n-1})) + \frac{1}{8}E(X_n)$$

$$= \frac{1}{8}\mu + \frac{3}{4(n-2)}(n-2)\mu + \frac{1}{8}\mu = \mu$$

$$E(\hat{\theta}_3) = E(\overline{X}) = \mu.$$

Hence, $\hat{\theta}_1, \hat{\theta}_2$, *and* $\hat{\theta}_3$ *are unbiased estimators of* μ.

(b) *Computing the variances, we have*

$$Var(\hat{\theta}_1) = \frac{1}{9}(Var(X_1) + Var(X_2) + Var(X_3))$$

$$= \frac{1}{9}3\sigma^2 = \frac{\sigma^2}{3}.$$

$$Var(\hat{\theta}_2) = \frac{\sigma^2}{64} + \frac{9(n-2)\sigma^2}{16(n-2)^2} + \frac{\sigma^2}{64}$$

$$= \frac{2\sigma^2}{64} + \frac{9\sigma^2}{16(n-2)} = \frac{n+16}{32(n-2)}\sigma^2.$$

$$Var(\hat{\theta}_3) = \frac{\sigma^2}{n}.$$

The relative efficiencies are

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} = \frac{\sigma^2(n+16)/32(n-2)}{\sigma^2/3}$$

$$= \frac{3(n+16)}{32(n-2)} < 1 \text{ for } n > 3.$$

Thus, for $n \geq 4$, $\hat{\theta}_2$ *is more efficient than* $\hat{\theta}_1$.

$$e(\hat{\theta}_3, \hat{\theta}_1) = \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_3)} = \frac{\sigma^2/3}{\sigma^2/n} = \frac{n}{3} > 1 \text{ for } n \geq 4.$$

Hence, for $n > 3$, $\hat{\theta}_3$ *is more efficient than* $\hat{\theta}_1$.

$$e(\hat{\theta}_3, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_3)} = \frac{\frac{n+16}{32(n-2)}\sigma^2}{\sigma^2/n}$$

$$= \frac{n^2+16n}{32(n-2)} > 1 \text{ for } n \geq 4.$$

Therefore, even though both $\hat{\theta}_3\hat{\theta}_2$ *are based on all the* n *observations, for* $n > 3$, *the sample mean* $\hat{\theta}_3$ *is more efficient than* $\hat{\theta}_2$.

Example :

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the uniform distribution on the interval $(0, \theta)$. Two unbiased estimators for θ are

$$\hat{\theta}_1 = 2\overline{Y} \quad \text{and} \quad \hat{\theta}_2 = \left(\frac{n+1}{n}\right) Y_{(n)},$$

where $Y_{(n)} = \max(Y_1, Y_2, \ldots, Y_n)$. Find the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$.

Hint:

$$E(Y_{(n)}) = \left(\frac{n}{n+1}\right)\theta, \qquad V(Y_{(n)}) = \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right]\theta^2$$

Solution :

Because each Y_i has a uniform distribution on the interval $(0, \theta)$, $\mu = E(Y_i) = \theta/2$ and $\sigma^2 = V(Y_i) = \theta^2/12$. Therefore,

$$E(\hat{\theta}_1) = E(2\overline{Y}) = 2E(\overline{Y}) = 2(\mu) = 2\left(\frac{\theta}{2}\right) = \theta,$$

and $\hat{\theta}_1$ is unbiased, as claimed. Further,

$$V(\hat{\theta}_1) = V(2\overline{Y}) = 4V(\overline{Y}) = 4\left[\frac{V(Y_i)}{n}\right] = \left(\frac{4}{n}\right)\left(\frac{\theta^2}{12}\right) = \frac{\theta^2}{3n}.$$

$$E(\hat{\theta}_2) = \left(\frac{n+1}{n}\right)E(Y_{(n)}) = \left(\frac{n+1}{n}\right)\left(\frac{n}{n+1}\right) = \theta$$

that is, $\hat{\theta}_2$ is an unbiased estimator for θ.

$$V(\hat{\theta}_2) = V\left[\left(\frac{n+1}{n}\right)Y_{(n)}\right] = \left(\frac{n+1}{n}\right)^2 V(Y_{(n)})$$
$$= \left[\frac{(n+1)^2}{n(n+2)} - 1\right]\theta^2 = \frac{\theta^2}{n(n+2)}.$$

Therefore, the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is given by

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)} = \frac{\theta^2/[n(n+2)]}{\theta^2/3n} = \frac{3}{n+2}.$$

This efficiency is less than 1 if $n > 1$. That is, if $n > 1$, $\hat{\theta}_2$ has a smaller variance than $\hat{\theta}_1$, and therefore $\hat{\theta}_2$ is generally preferable to $\hat{\theta}_1$ as an estimator of θ. ■

Example:

Let $X_1, \ldots, X_n, n \geq 2$ be a random sample from a normal population with a true mean μ and variance σ^2. Consider the following two estimators of σ^2: $\hat{\theta}_1 = S^2$, and $\hat{\theta}_2 = S'^2$. Find $e(\hat{\theta}_1, \hat{\theta}_2)$.

$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Solution:

$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$. So,

$$E\left(\frac{(n-1)S^2}{\sigma^2}\right) = (n-1) \Rightarrow \frac{(n-1)E(S^2)}{\sigma^2} = (n-1) \Rightarrow E(S^2) = \sigma^2$$

$$Var\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1) \Rightarrow \frac{(n-1)^2 Var(S^2)}{\sigma^4} = 2(n-1) \Rightarrow Var(S^2) = \frac{2\sigma^4}{(n-1)}$$

Because $E(S^2) = \sigma^2$, S^2 *is an unbiased estimator of* σ^2. Then, $MSE(S^2) = \frac{2\sigma^4}{(n-1)}$

$$nS'^2 = S^2(n-1) \Rightarrow S'^2 = \frac{S^2(n-1)}{n}$$

$$E(S'^2) = \frac{E(S^2)(n-1)}{n} = \frac{(n-1)\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n}$$

Because $E(S'^2) \neq \sigma^2$, S'^2 *is a biased estimator of* σ^2 *and* $Bias\ is\ -\frac{\sigma^2}{n}$.

$$Var(S'^2) = \frac{Var(S^2)(n-1)^2}{n^2} = \frac{2\sigma^4}{(n-1)}\frac{(n-1)^2}{n^2} = \frac{(n-1)2\sigma^4}{n^2}$$

$$MSE(S'^2) = \frac{(n-1)2\sigma^4}{n^2} + \left(-\frac{\sigma^2}{n}\right)^2 = \frac{(2n-1)\sigma^4}{n^2}$$

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)} = \frac{MSE(S'^2)}{MSE(S^2)}$$

$$= \frac{\frac{(2n-1)\sigma^4}{n^2}}{\frac{2\sigma^4}{(n-1)}} = \frac{(2n-1)(n-1)}{2n^2}$$

For $n \geq 2$, it can be seen that $e(\hat{\theta}_1, \hat{\theta}_2) < 1$. Hence, S'^2 is relatively more efficient than S^2.

We now turn to a more comprehensive approach to finding an efficient or minimum variance unbiased estimator.

Definition: An unbiased estimator $\hat{\theta}_0$, is said to be a *uniformly minimum variance unbiased estimator* (UMVUE) for the parameter θ if, for any other unbiased estimator $\hat{\theta}$,

$$Var(\hat{\theta}_0) \leq Var(\hat{\theta})$$

for all possible values of θ. It is not always easy to find an UMVUE for a parameter.

Our search for any such estimator is facilitated by the notion of finding a lower limit *CR*(θ, *n*) (hereafter called the *Cramér-Rao* (1945, 1946)) on the variance of any unbiased estimator of a parameter θ. If $\hat{\theta}$ is any unbiased estimator of θ, then $V(\hat{\theta}) \geq CR(\theta, n)$. This lower bound enables us to determine if a given unbiased estimator has the (theoretically) smallest possible variance in the sense that, if $V(\hat{\theta}) = CR(\theta, n)$, then $\hat{\theta}$ represents the *most efficient* estimator of θ. In this regard, if we can find an unbiased estimator $\hat{\theta}$ for which $V(\hat{\theta}) = CR(\theta, n)$, then the *most efficient* estimator is actually a *minimum variance bound estimator*.

Theorem: *Cramér-Rao Inequality :*

Let $X_1, X_2, \dots, X_n$ ve iid random sample variables from a population with $pdf\ (or\ pf\ )\ f_\theta(x)$ that depends on a parameter θ. If $\hat{\theta}$ is an unbiased estimator of θ, and if

$$Var(\hat{\theta}) \geq \frac{1}{-E\left(\frac{d^2 \ln L}{d\theta^2}\right)} = CR(\theta, n)$$

Then, $\hat{\theta}$ is a uniformly *minimum variance unbiased estimator* (UMVUE) of θ. Hence the variance of $\hat{\theta}$ is never smaller than *CR*(θ, *n*), which is constant for a fixed *n*. If $\hat{\theta}$ is an unbiased estimator of θ and strict equality holds in Theorem, then $\hat{\theta}$ is the *most efficient* or *minimum variance bound estimator* of θ. Hence the realizations of the *most efficient* estimator are those that are most concentrated about θ and thus have the highest probability of being close to θ.

Example: Let $X_1, X_2, \dots, X_n$ be a set of independent and identically distributed sample random variables taken from a normally distributed population. Assuming that σ^2 is known, what is the Cramér-Rao lower bound on the variance of any unbiased estimator of μ?

Solution:

$$\mathcal{L}(\mu,\sigma^2;x,n)=\frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}},$$

$$\log\mathcal{L}=-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}$$

and thus

$$\frac{\partial\log\mathcal{L}}{\partial\mu}=\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2},\quad\frac{\partial^2\log\mathcal{L}}{\partial\mu^2}=-\frac{n}{\sigma^2}.$$

Then

$$-E\left(\frac{\partial^2\log\mathcal{L}}{\partial\mu^2}\right)=-E\left(-\frac{n}{\sigma^2}\right)=\frac{n}{\sigma^2}$$

so that

$$Var(\hat{\theta})\geq\frac{1}{-E\left(\frac{\partial^2\log\mathcal{L}}{\partial\mu^2}\right)}=\frac{\sigma^2}{n}=CR(\mu,n).$$

Since $\bar{X}$ is an unbiased estimator of μ and $Var(\bar{X})=\frac{\sigma^2}{n}$, it follows that $\hat{\theta}=\hat{\mu}=\bar{X}$ is a minimum variance bound estimator of μ.

Example: Let $X_1, X_2, \ldots, X_n$ be a set of independent and identically distributed sample random variables taken from a normally distributed population. Assuming that μ is known, what is the Cramér-Rao lower bound on the variance of any unbiased estimator of σ^2?

Solution:

$$\frac{\partial \log \mathcal{L}}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\left(\frac{1}{\sigma^4}\right),$$

$$\frac{\partial^2 \log \mathcal{L}}{\partial (\sigma^2)^2} = \frac{n}{2}\frac{1}{\sigma^4} - \sum_{i=1}^{n}(x_i - \mu)^2\left(\frac{1}{\sigma^6}\right).$$

Then

$$-E\left(\frac{\partial^2 \log \mathcal{L}}{\partial (\sigma^2)^2}\right) = \frac{-n}{2\sigma^4} + \frac{1}{\sigma^6}E\left(\sum_{i=1}^{n}(x_i - \mu)^2\right) = \frac{-n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} = \frac{n}{2\sigma^4}$$

and thus,

$$Var(\hat{\theta}) \geq \frac{1}{-E\left(\frac{\partial^2 \log \mathcal{L}}{\partial(\sigma^2)^2}\right)} = \frac{2\sigma^4}{n} = CR(\sigma^2, n).$$

Since $MLE\ of\ \sigma^2, S'^2 = \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{n}$ is an unbiased estimator of σ^2 and

$Var(S'^2) = \frac{2\sigma^4}{n}$, it follows that $\hat{\theta} = \hat{\sigma}^2 = S'^2$ is a minimum variance bound estimator of σ^2.

Example: Let $X_1, X_2, \ldots, X_n$ be a set of independent and identically distributed sample random variables taken from a Poisson distributed population with unknown parameter λ. What is the Cramér-Rao lower bound on the variance of any unbiased estimator of λ?

Solution:

$$\mathcal{L}(\lambda; x, n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

with

$$\log \mathcal{L} = -n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log x_i!,$$

we have

$$\frac{\partial \log \mathcal{L}}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}, \quad \frac{\partial^2 \log \mathcal{L}}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}.$$

Then

$$-E\left(\frac{\partial^2 \log \mathcal{L}}{\partial \lambda^2}\right) = \frac{1}{\lambda^2} \sum_{i=1}^{n} E(x_i) = \frac{1}{\lambda^2} n\lambda = \frac{n}{\lambda}$$

and thus, from (10.10),

$$Var\left(\hat{\theta}\right) \geq \frac{1}{-E\left(\frac{\partial^2 \log \mathcal{L}}{\partial \lambda^2}\right)} = \frac{n}{\lambda} = \frac{\sigma^2}{n} = CR(\lambda, n).$$

Since the sample mean $\bar{X}$ is an unbiased estimator of λ and the variance of the sample mean is $\frac{\sigma^2}{n}$, it follows that $Var\left(\hat{\theta}\right) = Var(\bar{X}) = \frac{\sigma^2}{n}$, and thus $\hat{\theta} = \bar{X}$ is a minimum variance bound estimator of λ.

2.3 Sufficiency:

In the statistical inference problems on a parameter, one of the major questions is: Can a specific statistic replace the entire data without losing pertinent information?

Suppose $X_1, X_2, \dots, X_n$ is a iid random sample variables from a probability distribution with unknown parameter θ. In general, statisticians look for ways of reducing a set of data so that these data can be more easily understood without losing the meaning associated with the entire collection of observations. Intuitively, a statistic T is a sufficient statistic for a parameter θ if U contains all the information available in the data about the value of θ.

For example, the sample mean may contain all the relevant information about the parameter μ, and in that case $T = \bar{X}$ is called a sufficient statistic for μ. An estimator that is a function of a sufficient statistic can be deemed to be a "good" estimator, because it depends on fewer data values. When we have a sufficient statistic T for θ, we need to concentrate only on T because it exhausts all the information that the sample has about θ. That is, knowledge of the actual n observations does not contribute anything more to the inference about θ.

Sufficient statistics often can be used to develop estimators that are efficient (have minimum variance among all unbiased estimators). In fact, as we shall now see, if a minimum variance bound or most efficient estimator exists, it will be found to be a sufficient statistic. Moreover, if an efficient estimator of θ exists, it is expressible as a function of a sufficient statistic.

Definiton:

Let $X_1, X_2, \dots, X_n$ be a random sample from a probability distribution with unknown parameter θ. Then, the statistic $T = g(X_1, X_2, \dots, X_n)$ is said to be sufficient for θ if the conditional pdf or pf of $X_1, X_2, \dots, X_n$ given $T = t$ does not depend on θ for any value of t. An estimator of θ, that is a function of a sufficient statistic for θ is said to be a sufficient estimator of θ.

Example: Let $X_1, X_2, \dots, X_n$ be iid random variables with parameter p. Show that $T = \sum_{i=1}^{n} X_i$ is sufficient for p.

Solution:

$T = \sum_{i=1}^{n} X_i$ be the total number of successes in the n independent trials. If the value of T is known, can we gain any additional information about p by examining the value of any alternative estimator that also depends on $X_1, X_2, \dots, X_n$? According to our definition of sufficiency, the conditional distribution of $X_1, X_2, \dots, X_n$ given $T = t$ must be independent of p. Is it?

The joint probability function,

$$f(X_1, X_2, \dots, X_n ; p) = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}, \qquad 0 \le \theta \le 1$$

Because, $T = \sum_{i=1}^{n} X_i$ we have,

$$f(X_1, X_2, \dots, X_n ; p) = p^{T}(1-p)^{n-T}, \qquad 0 \le \theta \le 1$$

Also, because, $T \sim B(n, p)$, we have,

$$f(t ; p) = \binom{n}{t} p^t (1-p)^{n-t}$$

Also,

$$f(X_1, X_2, \dots, X_n | \, T = t) = \frac{f(X_1, X_2, \dots, X_n ; p)}{f(t ; p)} = \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}$$

which is independent of p. Therefore, $T = \sum_{i=1}^{n} X_i$ is sufficient estimator for p.

The definition of a sufficient statistic can tell us how to check to see if a particular estimator is sufficient for a parameter θ, but it does not tell us how to actually go about finding a sufficient statistic (if one exists). To address the issue of operationally determining a

sufficient statistic, we turn to following Theorem, the *Fisher-Neyman Factorization Theorem* (1922, 1924, 1935):

Theorem: *Fisher-Neyman Factorization Theorem:*

Let $X_1, X_2, \ldots, X_n$ be a random sample taken from a population with probability density function f (x; θ). The estimator $T = g(X_1, X_2, \ldots, X_n, n)$ is a sufficient statistic for the parameter θ if and only if the likelihood function of the sample factors as the product of two nonnegative functions $h(t; \theta, n)$ and $j(x_1, x_2, \ldots, x_n, n)$ or

$$\mathcal{L}(\theta; x_1, \ldots, x_n, n) = h(t; \theta, n) \cdot j(x_1, \ldots, x_n, n)$$

for every realization $t = g(x_1, x_2, \ldots, x_n, n)$ of T and all admissible values of θ. Although the function j is independent of θ (it may possibly be a constant), the function h depends on the sample realizations via the estimator T and thus this estimator constitutes a sufficient statistic for θ.

Example: Let the sample random variables $X_1, X_2, \ldots, X_n$ be drawn from a Poisson population with parameter $\lambda > 0$. Show that the minimum variance bound estimator $\bar{X}$ is sufficient for λ.

Solution:

$p(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$,

$$\mathcal{L}(\lambda; x, n) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} = \frac{e^{-n\lambda} \lambda^{n\bar{x}}}{\prod_{i=1}^{n} x_i!} = h(\bar{x}; \lambda, n) \cdot j(x_1, \ldots, x_n, n),$$

where $h(\bar{x}; \lambda, n) = e^{-n\lambda} \lambda^{n\bar{x}}$.

Note that h depends upon the X_i, $i = 1, \ldots, n$, only through the function or realization $t = \bar{x}$. So, $\bar{X}$ is a sufficient statistic for λ.

Example:

If the sample random variables $X_1,\ldots,X_n$ are taken from a $N(\mu,\sigma)$ population with probability density function $f(x;\mu,\sigma) = (2\pi\sigma^2)^{-\frac{1}{2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x,\mu < +\infty, \sigma > 0$, then, with σ^2 known, the likelihood function factors as

$$\begin{aligned}\mathcal{L}(\mu,\sigma^2;x,n) &= (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\sum_{i=1}^{n}\frac{(x_i-\bar{x}+\bar{x}-\mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2}{2\sigma^2}} \\ &= \left[e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}}\right]\left[(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2\sigma^2}}\right] \\ &= h(\bar{x};\mu,n)\ .\ j(x_1,\ldots,x_n,n)\end{aligned}$$

where $h(\bar{x};\mu,n) = e^{-n(\bar{x}-\mu)^2/2\sigma^2}$. Since h is a function of the sample random variables only through $t = \bar{x}$, it follows from the factorization criterion that, for all x and μ, $\overline{X}$ is a sufficient statistic for μ as well as a minimum variance bound estimator for the same.

Example:

If the sample random variables are extracted from a $N(\mu,\sigma)$ distribution with $\mu = 0$, then the population probability density function appears as $f(x;\mu,\sigma) = (2\pi\sigma^2)^{-1/2}e^{-x^2/2\sigma^2}, -\infty < x,\mu < +\infty, \sigma > 0$, and the likelihood function assumes the form

$$\begin{aligned}\mathcal{L}(\mu,\sigma^2;x,n) &= (2\pi\sigma^2)^{-\frac{n}{2}}e^{-\sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}} \\ &= \left[(\sigma^2)^{-\frac{n}{2}}e^{-\frac{ns_0^2}{2\sigma^2}}\right]\left[(2\pi)^{-\frac{n}{2}}\right] = h(s_0^2;\sigma^2,n)\cdot j(x_1,\ldots,x_n,n),\end{aligned}$$

where $h(s_0^2;\sigma^2,n) = (\sigma^2)^{-n/2}e^{-ns_0^2/2\sigma^2}$ and the function j is a constant. (Remember that with μ known, $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ serves as a minimum variance bound estimator of σ^2.) Since h depends on the $X_i, i = 1,\ldots,n$, only through the function or realization $s_0^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2$ of $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2$, the factorization criterion enables us to conclude that, for all x and σ^2, $\sum_{i=1}^{n}X_i^2$ as well as S_0^2 is sufficient for σ^2. ■

Definition: *Minimal Sufficient Statistics*

It may be the case that there is more than one sufficient statistic associated with the parameter θ of some population probability density (or mass) function. Hence it is only natural to ask whether one sufficient statistic is better than any other such statistic in the sense that it achieves the highest possible degree of data reduction without loss of information about θ. A sufficient statistic that satisfies this requirement will be termed a minimal sufficient statistic. More formally, a sufficient statistic T for a parameter θ is termed a *minimal sufficient statistic* if, for any other sufficient statistic T' for θ, T is a function of T'. A general procedure for finding a minimal sufficient statistic is provided by the following Theorem, the *Lehmann–Scheffé Theorem* (1950, 1955, 1956):

Theorem: *Lehmann–Scheffé Theorem:*

Let $L(\theta;\ x_1, x_2, \dots, x_n, n)$ denote the likelihood function of a random sample taken from the population probability density function $f\ (x;\ \theta)$. Suppose there exists a function $T = g(X_1, X_2, \dots, X_n, n)$ such that, for the two sets of sample realizations $\{x_1, x_2, \dots, x_n\}$ and $\{y_1, y_2, \dots, y_n\}$, the likelihood ratio

$$\frac{L(\theta;\ x_1, x_2, \dots, x_n, n)}{L(\theta;\ y_1, y_2, \dots, y_n, n)}$$

is independent of θ if and only if $g(x_1, x_2, \dots, x_n, n) = g(y_1, y_2, \dots, y_n, n)$. Then, T is a minimal sufficient statistics for θ

Example:

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be two sets of sample random variables taken from a Poisson population with probability mass function $p(x;\lambda) = e^{-\lambda}\lambda^x/x!, x = 0, 1, \dots; \lambda > 0$.

$$\frac{\mathcal{L}(\lambda;\ x,\ n)}{\mathcal{L}(\lambda;\ y,\ n)} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i} \Big/ \prod_{i=1}^{n} x_i!}{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i} \Big/ \prod_{i=1}^{n} y_i!} = \lambda^{\sum_{i=1}^{n}(x_i - y_i)} \left(\frac{\prod_{i=1}^{n} y_i!}{\prod_{i=1}^{n} x_i!} \right) = \lambda^{n(\bar{x}-\bar{y})} \left(\frac{\prod_{i=1}^{n} y_i!}{\prod_{i=1}^{n} x_i!} \right).$$

Clearly this ratio is free of λ if and only if $\bar{x} = \bar{y}$. Hence $\overline{X}$ is a minimal sufficient statistic for λ. ■

Let us now consider the importance or usefulness of a sufficient statistic (provided, of course, that one exists). As will be seen next, sufficient statistics often can be used to construct estimators that have minimum variance among all unbiased estimators or are efficient. In fact, sufficiency is a necessary condition for efficiency in that an estimator cannot be efficient unless it utilizes all the sample information.

We note first that if an unbiased estimator T of a parameter θ is a function of a sufficient statistic S, then T has a variance that is smaller than that of any other unbiased estimator of θ that is not dependent on S. Second, if T is a minimum variance bound or *most efficient* estimator of θ, then T is also a sufficient statistic (although the converse does not necessarily hold).

This first point may be legitimized by the *Rao-Blackwell Theorem* (presented next), which establishes a connection between unbiased estimators and sufficient statistics. In fact, this theorem informs us that we may improve upon an unbiased estimator by conditioning it on a sufficient statistic. That is, an unbiased estimator and a sufficient statistic for a parameter θ may be combined to yield a single estimator that is both unbiased and sufficient for θ and has a variance no larger (and usually smaller) than that of the original unbiased estimator.

To see this let T be any unbiased estimator of a parameter θ and let the statistic S be sufficient for θ. Then the Rao-Blackwell Theorem indicates that another estimator T' can be derived from S (i.e., expressed as a function of S) such that T' is unbiased and sufficient for θ with the variance of T' uniformly less than or equal to the variance of T.

Theorem: *Rao-Blackwell Theorem :*

Let $X_1, X_2, \dots, X_n$ be a random sample taken from the population probability density function $f(x; \theta)$ and let $T = g(X_1, X_2, \dots, X_n, n)$ be any unbiased estimator of the parameter θ. For $S = u(X_1, X_2, \dots, X_n, n)$ a sufficient statistic for θ, define $T' = E(T/S)$, where T is not a function of S alone. Then $E(T') = \theta$ and $V(T') \leq V(T)$ for all θ (with $V(T') < V(T)$ for some θ unless $T = T'$ with probability 1).

What this theorem tells us is that if we condition any unbiased estimator T on a sufficient statistic S to obtain a new unbiased estimator T', then T' is also sufficient for θ and is a uniformly *better* unbiased estimator of θ. Moreover, the resulting efficient estimator is unique.

It was mentioned earlier that minimum variance bound estimators and sufficient statistics are related. In fact, as we shall now see, under certain conditions they are one and the same. To see this let us return to the Factorization Theorem and rewrite the factorization criterion or sufficiency condition,

$$\log \mathcal{L}(\theta; x_1, \ldots, x_n, n) = \log h(t; \theta, n) + \log j(x_1, \ldots, x_n, n).$$

Then the factorization criterion may be respecified as

$$\frac{\partial \log \mathcal{L}}{\partial \theta} = \frac{\partial \log h(t; \theta, n)}{\partial \theta} = \alpha(\theta, n)\left(t(x_1, \ldots, x_n, n) - \theta\right),$$

then the sufficient statistic T is also a minimum variance bound estimator of the parameter θ.

Example:

We noted earlier that if our random sample is taken from a Poisson population, then the likelihood function of the sample can be factored as

$$\mathcal{L}(\lambda; x, n) = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{\prod_{i=1}^{n} x_i!} = h(\bar{x}; \lambda, n) \cdot j(x_1, \ldots, x_n, n),$$

with $h(\bar{x}; \lambda, n) = e^{-n\lambda}\lambda^{n\bar{x}}$. Then $\log h = -n\lambda + n\bar{x} \log \lambda$ and thus

$$\frac{\partial \log h}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda} = \frac{n}{\lambda}(\bar{x} - \lambda),$$

where, $\alpha(\lambda, n) = \frac{n}{\lambda}$ and $t(x_1, \ldots, x_n, n) - \lambda = \bar{x} - \lambda$.

Hence the sufficient statistic $\overline{X}$ is also a minimum variance bound estimator of λ.

***Definition**: Jointly sufficient statistics*

Specifically, for $X_1, X_2, \dots, X_n$ a set of sample random variables taken from the population probability density function $f\ (x;\ \theta)$, the statistics $S_k = u^k(X_1, X_2, \dots, X_n, n)$, $k = 1, \dots, r$, are said to be *jointly sufficient* if and only if the joint probability density function of $X_1, X_2, \dots, X_n$ given $S_1, S_2, \dots, S_n$ is independent of θ for any set of realizations $s_k = u^k(x_1, x_2, \dots, x_n, n)$, $k = 1, \dots, r$; that is, the conditional distribution of $X_1, X_2, \dots, X_n$ given $S_1, S_2, \dots, S_n$ or $f\,(X_1, X_2, \dots, X_n | S_1, S_2, \dots, S_n)$ does not depend on θ.

We next examine a set of statistics for their joint sufficiency by considering the following Theorem, the *Generalized Fisher-Neyman Factorization Theorem:*

***Theorem :** Generalized Fisher-Neyman Factorization Theorem*

Let $X_1, \dots, X_n$ be a random sample taken from a population with probability density function $f(x;\theta)$. The set of estimators $S_k = u^k(X_1, \dots, X_n, n), k = 1, \dots, r$, is jointly sufficient for the parameter θ if and only if the likelihood function of the sample factors as the product of the two nonnegative functions $h(s_1, \dots, s_r; \theta, n)$ and $j(x_1, \dots, x_r, n)$ or

$$\mathcal{L}(\theta; x_1, \dots, x_n, n) = h(s_1, \dots, s_r; \theta, n) \cdot j(x_1, \dots, x_n, n)$$

for every set of realizations $s_k = u^k(x_1, \dots, x_n, n), k = 1, \dots, r$, of the S_k's, $k = 1, \dots, r$, and all admissible values of θ.

Here the function j is independent of θ (it may be a constant) and h depends on the sample realizations via the estimators $S_k, k = 1, \dots, r$, and thus the S_k's, $k = 1, \dots, r$, constitute a set of jointly sufficient statistics for θ. Equation is termed the *factorization criterion for jointly sufficient statistics.*

Example:

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$.

If μ and σ^2 are both unknown, show that $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$ are jointly sufficient for μ and σ^2.

Solution

The likelihood function of the sample is

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

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

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

When both μ and σ^2 are unknown, use

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

and

$$j(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}}.$$

Hence, $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$ are jointly sufficient for μ and σ^2.

We noted earlier that a minimal sufficient statistic is one that achieves the highest possible degree of data reduction without losing any information about a parameter θ. Moreover, a general procedure for finding a minimal sufficient statistic was provided by the Lehman-Scheffé Theorem. Let us now extend this theorem to the determination of a set of jointly minimal sufficient statistics. We thus have the *generalized Lehman-Scheffé Theorem*:

Theorem: *Generalized Lehman-Scheffé Theorem:*

Let $\mathcal{L}(\theta; x_1, \ldots, x_n, n)$ denote the likelihood function of a random sample taken from the probability density function $f(x;\theta)$. Suppose there exists a set of functions $S_k = u^k(X_1, \ldots, X_n, n), k = 1, \ldots, r$, such that, for two sets of sample realizations $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ the likelihood ratio

$$\frac{\mathcal{L}(\theta; x_1, \ldots, x_n, n)}{\mathcal{L}(\theta; y_1, \ldots, y_n, n)}$$

is independent of θ if and only if $u^k(x_1, \ldots, x_n, n) = u^k(y_1, \ldots, y_n, n)$, $k = 1, \ldots, r$. Then the $S_k, k = 1, \ldots, r$, represent, jointly, a set of minimal sufficient statistics for θ.

Example:

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be two sets of sample random variables taken from a normal probability density function with unknown mean μ and variance σ^2. From (10.20),

$$\begin{aligned}
\frac{\mathcal{L}(\mu, \sigma^2; x, n)}{\mathcal{L}(\mu, \sigma^2; y, n)} &= \frac{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}}{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}} \\
&= e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2 - \sum_{i=1}^{n}(y_i-\mu)^2\right)} \\
&= e^{-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}y_i^2\right) - 2\mu\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}y_i\right)\right]} \\
&= e^{-\frac{1}{2\sigma^2}\left[(s_2^x - s_2^y) - 2\mu(s_1^x - s_1^y)\right]},
\end{aligned}$$

where $s_2^x = \sum_{i=1}^{n} x_i^2, s_2^y = \sum_{i=1}^{n} y_i^2, s_1^x = \sum_{i=1}^{n} x_i$, and $s_1^y = \sum_{i=1}^{n} y_i$. Clearly this ratio is free of μ and σ^2 if and only if $s_2^x = s_2^y$ and $s_1^x = s_1^y$. Thus $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ are jointly minimal sufficient statistics for μ and σ^2, respectively. And by virtue of the argument used in the preceding example problem, it follows that $\bar{X}$ and S^2 are also jointly minimal sufficient statistics for μ and σ^2.

2.4 Consistency

Definition:

The estimator $\hat{\theta}_n$ is said to be a *consistent estimator* of θ if, for any positive number ε,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \le \varepsilon) = 1$$

or, equivalently,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

The notation $\hat{\theta}_n$ expresses that the estimator for θ is calculated by using a sample of size n. For example, $\overline{Y}_2$ is the average of two observations

e.g :

Suppose that a coin, which has probability p of resulting in heads, is tossed n times. If the tosses are independent, then Y, the number of heads among the n tosses, has a binomial distribution. If the true value of p is unknown, the sample proportion Y/n is an estimator of p.

What happens to this sample proportion as the number of tosses n increases?

Our intuition leads us to believe that as n gets larger, Y/n should get closer to the true value of p. That is, as the amount of information in the sample increases, our estimator should get closer to the quantity being estimated.

FIGURE 2 Values of $\hat{p} = Y/n$ for a single sequence of 1000 Bernoulli trials, $p = 0.5$

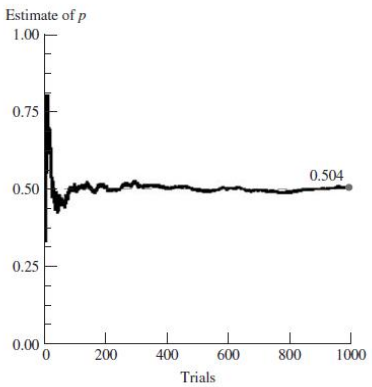


Figure 2 illustrates the values of $\hat{p} = Y/n$ for a single sequence of 1000 Bernoulli trials when the true value of p is 0.5. Notice that the values of $\hat{p}$ bounce around 0.5 when the number of trials is small but approach and stay very close to $p = 0.5$ as the number of trials increases.

The single sequence of 1000 trials illustrated in Figure 9.1 resulted (for larger n) in values for the estimate that were very close to the true value, $p = 0.5$.

Theorem: If $\hat{\theta}$ is an unbiased estimator of θ, is a consistent estimator for θ if,

$$\lim_{n\to\infty} Var(\hat{\theta}) = 0$$

And if $\hat{\theta}$ is a biased estimator of θ, is a consistent estimator for θ if,

$$\lim_{n\to\infty} Var(\hat{\theta}) = 0 \ and \ \lim_{n\to\infty} Bias(\hat{\theta}) = 0 \ ; \ or \ \lim_{n\to\infty} MSE(\hat{\theta}) = 0$$

PROCEDURE TO TEST FOR CONSISTENCY

1. Check whether the estimator $\hat{\theta}_n$ is unbiased or not.
2. Calculate $Var(\hat{\theta}_n)$ and $B(\hat{\theta}_n)$, the bias of $\hat{\theta}_n$.
3. An unbiased estimator is consistent if $Var(\hat{\theta}_n) \to 0$ as $n \to \infty$.
4. A biased estimator is consistent if both

$$Var(\hat{\theta}_n) \to 0 \text{ and } B(\hat{\theta}_n) \to 0 \text{ as } n \to \infty.$$

Example:

Let $X_1, \ldots, X_n$ be a random sample with true mean μ and finite variance. Then, the sample mean $\overline{X}$ is a consistent estimator of the population mean μ.

Solution:

We show this result in two ways.

(1) *Using Chebyshev's inequality,* $P\{|X - \mu| \geq \varepsilon\} \leq \frac{Var(x)}{\varepsilon^2}$, *we obtain*

$$P\left[\left|\overline{X} - \mu\right| \leq k\right] \geq 1 - \frac{\sigma_{\overline{X}}^2}{k^2}$$

$$= 1 - \frac{\sigma^2}{k^2 n} \to 1 \text{as } n \to \infty.$$

Hence, $\overline{X}$ *is a consistent estimator of* μ.

(ii) *First note that* $\overline{X}$ *is an unbiased estimator of* μ. *Because* $Var(\overline{X}) = (\sigma^2/n)$, *we have*

$$\lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

Thus, from the previous theorem, $\overline{X}$ *is a consistent estimator of* μ.

Example:

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ population.

(a) Show that the sample variance S^2 is a consistent estimator for σ^2.

(b) Show that the maximum likelihood estimators for μ and σ^2 are consistent estimators for μ and σ^2.

Solution:

(a) *We have already seen that* $ES^2 = \sigma^2$, *and hence,* S^2 *is an unbiased estimator of* σ^2. *Because the sample is drawn from a normal distribution, we know that* $[(n-1)S^2/\sigma^2]$ *has a chi-square distribution with* $(n-1)$ *d.f. and*

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

Thus,

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

This implies that

$$Var(S^2) = \frac{2\sigma^4}{n-1} \to 0 \text{ as } n \to \infty.$$

Hence, S^2 *is a consistent estimator of the variance of a normal population.*

(b) *We have seen that the MLE of* μ *is* $\hat{\mu} = \overline{X}$, *and that of* σ^2 *is* $\hat{\sigma}_n^2 = (1/n)\sum_{i=1}^n (X_i - \overline{X})^2$. *Now* $\hat{\mu}$ *is an unbiased estimator of* μ, *and* $Var(\overline{X}) = (\sigma^2/n) \to 0$ *as* $n \to \infty$.
$\overline{X}$ *is a consistent estimator for* μ.

Now we will use the identity

$$E\left[(\hat{\theta}_n - \theta)^2\right] = Var(\hat{\theta}_n) + \left[B(\hat{\theta}_n)\right]^2$$

to show that the MLE for σ^2 is biased with

$$E(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2 \quad \text{and} \quad B(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

Thus, $\hat{\sigma}_n^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2 = ((n-1)/n)S^2$. Using part (a), we get

$$Var(\hat{\sigma}_n^2) = \frac{(n-1)^2}{n^2}Var(S^2) \quad = \frac{(n-1)^2 2\sigma^4}{n^2(n-1)} = \frac{2(n-1)(\sigma^2)^2}{n^2}.$$

Therefore,

$$\lim_{n\to\infty} B(\hat{\sigma}_n^2) = \lim_{n\to\infty} \frac{-\sigma^2}{n} = 0, \text{ and } \lim_{n\to\infty} Var(\hat{\sigma}_n^2)$$
$$= \lim_{n\to\infty} \frac{2(n-1)(\sigma^2)^2}{n^2} = 0.$$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is a consistent estimator of σ^2.