

# Mathematical Statistics

---

## Miscellaneous

### Definition

$\Gamma(\alpha)$  denotes the gamma function which is defined as  $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt$ , where  $\alpha > 0$ . It is a commonly used extension of the factorial function.

### Proposition

The Gamma function has the following properties:

1. The improper integral converges for all  $\alpha > 0$ .
2.  $\Gamma(\alpha) > 0$  for all  $\alpha > 0$ ,  $\Gamma(\alpha) \rightarrow \infty$  as  $\alpha \rightarrow 0$ .
3.  $\Gamma(1) = 1$
4.  $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
5. For  $n \in \mathbb{N}$ ,  $\Gamma(n) = (n - 1)!$
6.  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

### Definition

For  $\alpha > 0, \beta > 0$ , the beta function is defined as  $B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1 - x)^{\beta-1} dx$ .

Alternatively, the beta function can be written as  $B(\alpha, \beta) = \int_0^{\infty} \frac{t^{\alpha-1}}{(1 + t)^{\alpha+\beta}} dt$

**Proposition** (Relation between beta function and gamma function)

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

**Proposition** (Limit Comparison Test for Improper Integrals)

Let  $a \in \mathbb{R}$  and  $f, g : [a, \infty) \rightarrow \mathbb{R}$  be such that both  $f$  and  $g$  are integrable on  $[a, x]$  for every  $x \geq a$  with  $f(t) > 0$  and  $g(t) > 0$  for all large  $t$ .

Assume that  $\lim_{t \rightarrow \infty} \frac{f(t)}{g(t)} = l$  where  $l \in [0, \infty]$ . Then:

- If  $l \in (0, \infty)$ , then  $\int_a^\infty f(x)dx$  converges  $\iff \int_a^\infty g(x)dx$  converges
- If  $l = 0$ , and  $\int_a^\infty g(x)dx$  converges then  $\int_a^\infty f(x)dx$  converges
- If  $l = \infty$  and  $\int_a^\infty f(x)dx$  converges then  $\int_a^\infty g(x)dx$  converges absolutely

**Proposition**

Let  $(a_n)$  be a sequence of real numbers such that  $a_n \rightarrow a$ . Then  $\lim_{n \rightarrow \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a$

**Definition**

The **indicator function** of the set  $A$  is defined as  $I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$

**Definition**

If  $f(t)$  is a function defined for all  $t \geq 0$ , its Laplace transform is defined as

$$F(s) = \int_0^\infty e^{-st} f(t) dt \text{ where } s \text{ is a real number}$$

**Note**

Laplace transform  $F(s)$  may not exist for all real number  $s$ .

**Proposition**

If  $f(t)$  is defined and piecewise continuous on every finite subinterval of  $(0, \infty)$  and satisfies the following growth restriction  $|f(t)| \leq M e^{ct}$  for all  $t \geq 0$  and some constants  $M > 0$  and  $c \in \mathbb{R}$ , the Laplace transform  $F(s)$  exists for all  $s \geq c$ .

**Proposition**

If the Laplace transform of a given function exists, it is uniquely determined.

Conversely, if two functions have the same transform, these functions cannot differ over an interval of positive length.

If two continuous functions have the same transform, they are completely identical.

**Proposition**

Suppose that  $f(x, y)$  and its first and second partial derivatives are continuous throughout a disk centered at  $(a, b)$  and that  $f_x(a, b) = f_y(a, b) = 0$ . Then:

1.  $f$  has a local maximum at  $(a, b)$  if  $f_{xx} < 0$  and  $f_{xx}f_{yy} - f_{xy}^2 > 0$  at  $(a, b)$
2.  $f$  has a local minimum at  $(a, b)$  if  $f_{xx} > 0$  and  $f_{xx}f_{yy} - f_{xy}^2 > 0$  at  $(a, b)$
3.  $f$  has a saddle point at  $(a, b)$  if  $f_{xx}f_{yy} - f_{xy}^2 < 0$  at  $(a, b)$

The test is inconclusive if  $f_{xx}f_{yy} - f_{xy}^2 = 0$  at  $(a, b)$ . In this case, we must find some other way to determine the behaviour of  $f$  at  $(a, b)$ .

---

## Fundamentals

### Definition

Two random vectors  $(X_1, X_2, \dots, X_m)$  and  $(Y_1, Y_2, \dots, Y_n)$  are said to be **independent** if  $F(x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n) = F_1(x_1, x_2, \dots, x_m)F_2(y_1, y_2, \dots, y_n)$  for all  $(x_1, x_2, \dots, x_m, y_1, y_2, \dots, y_n) \in \mathbb{R}^{m+n}$  where  $F, F_1, F_2$  are the joint CDF's of  $(X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n)$ ,  $(X_1, X_2, \dots, X_m)$  and  $(Y_1, Y_2, \dots, Y_n)$  respectively.

### Theorem

Let  $X = (X_1, X_2, \dots, X_m)$  and  $Y = (Y_1, Y_2, \dots, Y_n)$  be independent random vectors. Then the component  $X_j$  of  $X$  ( $j = 1, 2, \dots, m$ ) and the component  $Y_k$  of  $Y$  ( $k = 1, 2, \dots, n$ ) are independent random variables. If  $h$  and  $g$  are Borel-measurable functions,  $h(X_1, X_2, \dots, X_m)$  and  $g(Y_1, Y_2, \dots, Y_n)$  are independent.

### Theorem

Suppose  $(X, Y)$  have joint pdf  $f$ . Then  $X$  and  $Y$  are independent iff for some constant  $k > 0$  and non-negative functions  $f_1$  and  $f_2$ ,  $f(x, y) = kf_1(x)f_2(y)$  for all  $(x, y) \in \mathbb{R}^2$

### Theorem

Let  $X$  be a random variable with pdf  $f(x)$ . Then the pdf of  $aX + b$  where  $a \neq 0, b \in \mathbb{R}$  is given by  $\frac{1}{|a|}f\left(\frac{x-b}{a}\right)$

### Theorem

If  $X$  and  $Y$  are independent continuous random variables with pdfs  $f_X(x)$  and  $f_Y(y)$ , then the pdf of  $Z = X + Y$  is  $f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z-w) dw$

**Proposition**

Let  $(X, Y)$  be a random vector with joint density  $f(x, y)$  and  $g, h$  be continuous and differentiable real valued functions of two variables. Then to obtain the joint pdf of  $(g(X, Y), h(X, Y))$  we consider the equations  $g(x, y) = z$  and  $h(x, y) = w$ . There may be many such points  $(x, y)$  which map to  $z$  and  $w$  under  $g$  and  $h$  respectively.

Let the points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  represent the points which satisfy  $g(x_i, y_i) = z$  and  $h(x_i, y_i) = w$ .

We can find these points as  $\{x_i\} = g^{-1}(z, w)$  and  $\{y_i\} = h^{-1}(z, w)$ .

$$\text{Compute } J(x_i, y_i) = \begin{bmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{bmatrix}_{(x=x_i, y=y_i)}$$

$$\text{Then the joint pdf is } k(z, w) = \sum_i \frac{1}{|J(x_i, y_i)|} f(x_i, y_i)$$


---

**Expectation, Variance and Moments****Definition**

The **expectation** of a discrete random variable  $X$  having values  $x_1, x_2, \dots, x_n$  and

probability function  $f(x)$  is defined as  $E(X) = \sum_{i=1}^n x_i f(x_i)$ .

If  $X$  is a discrete random variable taking on infinite set of values  $x_1, x_2, \dots$ , then

$$E(X) = \sum_{i=1}^{\infty} x_i f(x_i) \text{ provided the infinite series converges absolutely.}$$

For a continuous random variable  $X$  with distribution function  $f(x)$ , the expectation of  $X$  is defined as  $E(X) = \int_{-\infty}^{\infty} x f(x) dx$  provided the integral converges absolutely.

**Theorem**

The expectation has the following properties:

1.  $E(cX) = cE(X)$  where  $c$  is any constant
2. If  $X$  and  $Y$  are any random variables then  $E(X + Y) = E(X) + E(Y)$
3. If  $X$  and  $Y$  are independent random variables, then  $E(XY) = E(X)E(Y)$

**Definition**

The **variance** is defined as  $\text{Var}(X) = \sigma^2 = E[(X - \mu)^2]$ . The **standard deviation** is defined as  $\sigma = \sqrt{\text{Var}(X)}$ .

**Theorem**

The variance has the following properties:

1.  $\sigma^2 = E[(X - \mu)^2] = E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$  where  $\mu = E(X)$
2. If  $c$  is any constant,  $\text{Var}(cX) = c^2\text{Var}(X)$
3. The quantity  $E[(X - a)^2]$  is a minimum where  $a = \mu = E(X)$
4. If  $X$  and  $Y$  are independent random variables, then  
 $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$  and  $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$

**Definition**

The  $r$ -th **moment** of a random variable  $X$  about the mean  $\mu$  is defined as  $\mu_r = E[(X - \mu)^r]$  where  $r = 0, 1, 2, \dots$

The  $r$ -th moment of  $X$  about the origin is defined as  $\mu'_r = E(X^r)$ .

It follows that  $\mu_0 = 0$ ,  $\mu_1 = 1$ ,  $\mu_2 = \sigma^2$

**Theorem** (Law of the unconscious statistician - LOTUS)

Let  $X$  be a discrete random variable with probability function  $f(x)$ . Then  $Y = g(x)$  is also a discrete random variable.

The probability function of  $Y$  is

$$h(y) = P(Y = y) = \sum_{\{x|g(x)=y\}} P(X = x) = \sum_{\{x|g(x)=y\}} f(x).$$

If  $X$  takes on values  $x_1, x_2, \dots, x_n$  and  $Y$  takes on values  $y_1, y_2, \dots, y_m$ , then  $m \leq n$  and  $y_1 h(y_1) + y_2 h(y_2) + \dots + y_m h(y_m) = g(x_1)f(x_1) + g(x_2)f(x_2) + \dots + g(x_n)f(x_n)$  which lets us write the expectation of  $Y$  as

$$E(Y) = g(x_1)f(x_1) + g(x_2)f(x_2) + \dots + g(x_n)f(x_n) = \sum_{i=1}^n g(x_i)f(x_i).$$

Similarly when  $X$  is a continuous random variable and  $Y = g(X)$ , then

$$E(Y) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

**Definition**

Let  $X$  be a random variable defined on  $(\Omega, \mathcal{F}, P)$ . The function  $M(t) = E[e^{tX}]$  is called the **moment generating function** (MGF) of the random variable  $X$  if the expectation on the right side exists in some neighbourhood of the origin. If the expectation on the right side does not exist in any neighbourhood of the origin, then we say the MGF does not exist.

The  $r$ -th derivative of the moment generating function is the  $r$ -th moment about the origin  $\mu'_r$ .

### Theorem

If the MGF  $M(s)$  of a random variable  $X$  exists, then the MGF  $M(s)$  has derivatives of all orders at  $s = 0$  and

$$M^{(k)}(s)|_{s=0} = EX^k \text{ for positive integer } k$$

### Theorem

The moment generating function has the following properties:

1. For any constants  $a$  and  $b$ , the mgf of the random variable  $aX + b$  is given by  

$$M_{aX+b} = e^{bt} M_X(at)$$
2. If  $X$  and  $Y$  are independent random variables having moment generating functions  $M_X(t)$  and  $M_Y(t)$  respectively, then  $M_{X+Y}(t) = M_X(t)M_Y(t)$
3. **Uniqueness Theorem** Suppose that  $X$  and  $Y$  are random variables having moment generating functions  $M_X(t)$  and  $M_Y(t)$  respectively. Then  $X$  and  $Y$  have the same probability distribution if and only if  $M_X(t) = M_Y(t)$  identically.

### Definition

Let  $X_1, X_2, \dots, X_n$  be a jointly distributed or  $(X_1, X_2, \dots, X_n)$  be a random vector. If

$E[\exp(\sum_{j=1}^n t_j X_j)]$  exists for  $|t_j| \leq h_j, j = 1, 2, \dots, n$ , we write

$M(t_1, t_2, \dots, t_n) = E[\exp(t_1 X_1 + t_2 X_2 + \dots + t_n X_n)]$  and call it the MGF of  $X_1, X_2, \dots, X_n$  or simply, the **joint moment generating function** (joint MGF) of the random vector  $(X_1, X_2, \dots, X_n)$

### Theorem

The joint MGF  $M(t_1, t_2)$  uniquely determines the joint distribution of  $(X, Y)$ . Conversely, if the joint MGF exists it is unique.

### Theorem

The joint MGF  $M(t_1, t_2)$  completely determines the marginal distributions of  $X$  and  $Y$ .  
 $M(t_1, 0) = E[\exp(t_1 X)] = M_X(t_1)$  and  $M(0, t_2) = E[\exp(t_2 Y)] = M_Y(t_2)$

### Theorem

$X$  and  $Y$  are independent random variables if and only if  $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$  for all  $t_1 \in [-h_1, h_1], t_2 \in [-h_2, h_2]$

## Distributions

### Definition

Let  $p$  be the probability that an event will happen in any single Bernoulli trial (trial with outcomes either success or failure). The probability that an event will happen exactly  $x$  times in  $n$  trials is given by the **Binomial Random Variable** with pmf distribution

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \text{ where } x = 0, 1, \dots, n.$$

### Proposition

The binomial random variable has mean  $\mu = np$  and variance  $\sigma^2 = npq$ .

### Definition

The **Poisson Random Variable** has pmf distribution  $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ ,  $x = 0, 1, 2, \dots$  and  $\lambda > 0$ .

### Proposition

The Poisson Random Variable has mean  $\mu = \lambda$  and variance  $\sigma^2 = \lambda$ .

### Proposition

The Poisson Random Variable with  $\lambda = np$  is the limiting case of the Binomial Distribution. It approximates the Binomial Random variable  $\text{Binomial}(n, p)$  when  $n$  is large and probability of occurrence of an event  $p$  is close to 0.

### Definition

The **uniform random variable** has the pdf distribution  $f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$ .

### Proposition

The uniform random variable has mean  $\mu = \frac{a+b}{2}$  and variance  $\sigma^2 = \frac{1}{12}(b-a)^2$ .

### Definition

The **exponential random variable** has the pdf distribution

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

### Proposition

The exponential random variable has mean  $\frac{1}{\lambda}$  and variance  $\frac{1}{\lambda^2}$ .

### Definition

The **normal random variable**, also known as the gaussian random variable, has pdf distribution  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$  where  $-\infty < x < \infty$  where  $\mu$  and  $\sigma$  are the mean and standard deviation respectively.

**Definition**

When  $\mu = 0$  and  $\sigma = 1$ , we get the **standard normal random variable** with distribution  $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$ .

We can write any normal random variable  $Y \sim N(\mu, \sigma)$  in terms of the standard normal random variable  $X \sim N(0, 1)$  as  $Y = \sigma X + \mu$ .

**Proposition**

When  $n$  is large and neither  $p$  or  $q$  is too close to 0, the binomial random variable  $X$  can be approximated by a normal distribution with mean  $np$  and standard deviation  $\sqrt{npq}$ . The approximation is very good when  $np, nq > 5$ .

**Proposition**

The Poisson Distribution approaches the normal distribution  $N(\lambda, \sqrt{\lambda})$  as  $\lambda \rightarrow \infty$ , i.e. the Poisson distribution is asymptotically normal.

**Definition**

A variable the pdf  $f(x) = \frac{1}{\sigma\pi(1 + \left(\frac{x-\mu}{\sigma}\right)^2)}$ ,  $x \in \mathbb{R}$  where  $\sigma > 0$ ,  $\mu \in \mathbb{R}$  is called a

**Cauchy random variable** with parameter  $\mu$  and  $\sigma^2$ . We write  $X \sim C(\mu, \sigma^2)$  for Cauchy random  $X$  with pdf.

**Definition**

When  $\mu = 0$  and  $\sigma^2 = 1$ , we get the **standard Cauchy random variable**  $C(0, 1)$  with distribution  $f(x) = \frac{1}{\pi(1 + x^2)}$ .

We can write any Cauchy random variable  $Y = C(\mu, \sigma^2)$  in terms of the standard Cauchy random variable  $X = C(0, 1)$  as  $Y = \sigma X + \mu$ .

**Proposition**

The mean, variance, higher moments, moment generating function of a Cauchy random variable do not exist.

**Definition**

A random variable with the pdf  $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$  where

$0 < x < \infty$ ,  $\alpha > 0$ ,  $\beta > 0$  is called a **gamma random variable**  $G(\alpha, \beta)$  with parameters  $\alpha$  and  $\beta$ .

**Proposition**

When  $\alpha = 1$ , we see that the gamma distribution is a generalization of the exponential distribution as

$G(1, \beta) = \exp(\beta)$  with pdf  $f(x) = \frac{1}{\beta} e^{-x/\beta}$ ,  $x > 0$ .



**Proposition**

The gamma distribution has mean  $\mu = \alpha\beta$  and variance  $\sigma^2 = \alpha\beta^2$ .

**Definition**

The **chi-square distribution** is

$$f(x) = \frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}} x^{\frac{p}{2}-1} e^{-x/2} \text{ where } 0 < x < \infty.$$

$\chi_p^2$  denotes a chi-square random variable with  $p$  degrees of freedom.

The chi-square distribution is a special case of the gamma distribution with  $\alpha = \frac{p}{2}$  and  $\beta = 2$ .

**Proposition**

The mean of the chi-square distribution is given by  $\mu = p$  and the variance is given by  $\sigma^2 = 2p$ .

**Proposition**

Let  $X \sim N(0, 1)$ . Then  $X^2 \sim \chi_1^2$

**Proposition**

Let  $X_1, X_2, \dots, X_n$  be independent normal random variables with mean 0 and variance 1. Then  $\chi^2 = X_1^2 + X_2^2 + \dots + X_p^2$  is chi-square distributed with  $p$  degrees of freedom.

**Definition**

A random variable  $T$  has the **Students t-distribution** with  $p$  degrees of freedom, and

we write  $T \sim t_p$  if it has pdf  $f_p(t) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})} \frac{1}{(p\pi)^{\frac{1}{2}}} \frac{1}{(1 + \frac{t^2}{p})^{\frac{(p+1)}{2}}}$  for  $-\infty < t < \infty$

**Proposition**

If  $p = 1$ ,  $T$  is a Cauchy(0, 1) distribution with distribution  $f_p(t) = \frac{\Gamma(1)}{\Gamma(\frac{1}{2})} \frac{1}{\pi^{\frac{1}{2}}} \frac{1}{1 + t^2}$ . So we will assume that  $p > 1$ .

**Proposition**

Let  $T \sim t_p$ . Then  $E[T^r]$  exists for  $r < p$  and

$$E[T^r] = \begin{cases} 0 & \text{if } r \text{ is odd} \\ p^{\frac{r}{2}} \frac{\Gamma(\frac{r+1}{2})\Gamma(\frac{p-r}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{1}{2})} & \text{if } r \text{ is even} \end{cases}$$

**Proposition**

The Students  $t$ -distribution has no MGF because it does not have moments of all orders.

**Proposition**

Let  $T \sim t_p$ ,  $p > 2$  be a random variable with Student's  $t$ -distribution. Then  $T$  has mean  $\mu = 0$  and variance  $\sigma^2 = \frac{p}{p-2}$ .

**Definition**

A random variable  $X \sim F(m, n)$  has the  **$F$ -distribution** with  $m$  and  $n$  degrees of freedom if it has pdf

$$f_F(t) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{t^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n}t\right)^{\frac{m+n}{2}}} \text{ where } t > 0$$

**Proposition**

If  $X \sim t_n$ , then  $X^2 \sim F(1, n)$ . In particular if  $X \sim C(0, 1)$ , i.e.  $X \sim t_1$ , then  $X^2 \sim F(1, 1)$ .

**Proposition**

Let  $X \sim F(m, n)$ . Then for  $k \in \mathbb{N}$ ,  $E[X^k] = \left(\frac{n}{m}\right)^k \frac{\Gamma(k + \frac{m}{2})\Gamma(\frac{n}{2} - k)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}$  for  $n > 2k$ .

**Proposition**

Let  $X \sim F(m, n)$ . Then  $X$  has mean  $\mu = \frac{n}{n-2}$  and variance  $\sigma^2 = \frac{n^2(2m+2n-4)}{m(n-2)^2(n-4)}$  for  $n > 4$ .

**Definition**

A random variable  $X \sim \text{beta}(\alpha, \beta)$  has the **beta distribution** if it has pdf

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \text{ for } 0 < x < 1 \text{ and } \alpha, \beta > 0.$$

**Definition**

A random variable  $X$  has the **Pareto distribution** with parameters  $\alpha > 0$  and  $\beta > 0$  if it has pdf

$$f(x) = \begin{cases} \frac{\beta \alpha^\beta}{x^{\beta+1}} & x \geq \alpha \\ 0 & x \leq \alpha \end{cases}$$

**Proposition**

For Pareto's distribution with parameter  $\alpha$  and  $\beta$ , the moment of order  $n$  exists if and only if  $n < \beta$ .

**Definition**

The **Weibull Distribution** (Weibull( $\gamma, \beta$ )) has pdf with parameters  $\gamma > 0, \beta > 0$  defined as  $f(x) = \frac{\gamma}{\beta} x^{\gamma-1} \exp(-\frac{x^\gamma}{\beta})$  where  $x \geq 0$ .

---

## Sampling Theory

We can either sample *with replacement* or *without replacement*. A finite population sampled with replacement can be considered infinite. Sampling from a very large finite population can similarly be considered as sampling from an infinite population.

To properly choose the sample, we can make sure that every member of the population has an equal chance of being in the sample. Normally, since the sample size is much smaller than the population size, sampling without replacement will give practically the same results as sampling with replacement.

For a sample of size  $n$  from a population which we assume has distribution  $f(x)$ , we can choose members of the population at random, each selection corresponding to a random variable  $X_1, X_2, \dots, X_n$  with corresponding values  $x_1, x_2, \dots, x_n$ . In case we are assuming sampling without replacement,  $X_1, X_2, \dots, X_n$  will be independent and identically distributed random variables with probability distribution  $f(x)$ .

**Definition**

Let  $X$  be a random variable with a distribution  $f$ , and let  $X_1, X_2, \dots, X_n$  be iid random variables with the common distribution  $f$ .

Then the collection  $X_1, X_2, \dots, X_n$  is called a **random sample** of size  $n$  from the population  $f$ .

Since  $X_1, X_2, \dots, X_n$  are iid, the joint distribution of the random sample is

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2)\dots f(x_n).$$

Any quantity obtained from a sample for the purpose of estimating a population parameter is called a sample statistic, or briefly statistic. Mathematically, a sample statistic for a sample of size  $n$  can be defined as a function of the random variables  $X_1, X_2, \dots, X_n$  as  $T(X_1, X_2, \dots, X_n)$ . This itself is a random variable whose values can be represented as  $T(x_1, x_2, \dots, x_n)$ .

### Definition

Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from the population whose distribution is  $f(x|\theta)$  (the distribution  $f$  with unknown parameter  $\theta$ ). Let  $T(x_1, x_2, \dots, x_n)$  be a real-valued or vector-valued function whose domain includes the range of  $(X_1, X_2, \dots, X_n)$ . Then the random variable or random vector  $Y = T(X_1, \dots, X_n)$  is called a **statistic** provided that  $T$  is not a function of any unknown parameter  $\theta$ .

For example consider  $X \approx N(\mu, \sigma^2)$  where  $\mu$  is known but  $\sigma$  is unknown. Then  $\frac{\sum_{i=1}^n X_i}{\sigma^2}$  is not a statistic but  $\frac{\sum_{i=1}^n X_i}{\mu^2}$  is a statistic.

Two common statistics are the sample mean and sample variance.

### Definition

The **sample mean** is the arithmetic average of the values in the random sample. It is denoted by  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ .

The **sample variance** is the statistic defined by  $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$

The **sample standard deviation** is the statistic defined by  $S = \sqrt{S^2}$ .

### Definition

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population  $f(x|\theta)$ . We say that a statistic  $T(X_1, X_2, \dots, X_n)$  is an **unbiased estimator** of the parameter  $\theta$  if  $E(T) = \theta$  for all possible values of  $\theta$ .

### Theorem

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with mean  $\mu$  and variance  $\sigma^2 < \infty$ . Then:

1.  $E(\bar{X}) = \mu$
2.  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
3.  $E(S^2) = \sigma^2$

From the above theorem we see that the sample mean  $\bar{X}$  is an unbiased estimator of the population mean  $\mu$  and the sample variance  $S^2$  is an unbiased estimator of the population variance  $\sigma^2$ . (The reason we included  $\frac{1}{n-1}$  in the definition of the sample variance was to make it an unbiased estimator)

**Definition**

Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a population  $f(x|\theta)$ . The probability distribution of a statistic  $T(X_1, X_2, \dots, X_n)$  is called the sampling distribution of  $T$ .

**Theorem** (MGF of the sample mean)

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with MGF  $M_X(t)$ . Then the MGF of the sample mean is  $M_{\bar{X}}(t) = (M_X(t/n))^n$ .

**Theorem**

Let  $X_1, X_2, \dots, X_n$  be a random sample from a  $N(\mu, \sigma^2)$  distribution, and let  $\bar{X}$  denote the sample mean.

Then  $\bar{X}$  and the random vector  $(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})$  are independent.

**Theorem**

Let  $X_1, X_2, \dots, X_n$  be a random sample from a  $N(\mu, \sigma^2)$  distribution, and let  $\bar{X}$  denote the sample mean and  $S^2$  denote the sample variance. Then  $\bar{X}$  and  $S^2$  are independent random variables.

The converse of this theorem is also true: if the sample mean and sample variance of a random sample are independent random variables then population distribution is normal.

**Theorem**

Let  $X_1, X_2, \dots, X_n$  be a random sample from a  $N(\mu, \sigma^2)$  distribution and let  $\bar{X}$  denote the sample mean and  $S^2$  denote the sample variance. Then  $(n-1)\frac{S^2}{\sigma^2}$  has a chi-square distribution with  $(n-1)$  degrees of freedom.

**Definition**

Let  $X_1, X_2, \dots, X_n$  be a random sample and  $x_1, x_2, \dots, x_n$  be values taken by these random variables. Arrange  $(x_1, x_2, \dots, x_n)$  in increasing order of magnitude so,  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  where  $x_{(1)} = \min(x_1, x_2, \dots, x_n)$ ,  $x_{(2)}$  is the second smallest value and so on and  $x_{(n)} = \max(x_1, x_2, \dots, x_n)$ . If any two  $x_i, x_j$  are equal, their order does not matter.

The function  $X_{(k)}$  of  $(X_1, X_2, \dots, X_n)$  that takes on the value  $x_{(k)}$  in each possible sequence  $(x_1, x_2, \dots, x_n)$  of values assumed by  $(X_1, X_2, \dots, X_n)$  is known as the  **$k$ -th order statistic**.

$X_{(1)}, X_{(2)}, \dots, X_{(n)}$  is called the **set of order statistics** for  $(X_1, X_2, \dots, X_n)$

**Definition**

The **sample range**  $R = X_{(n)} - X_{(1)}$  is the distance between the smallest and largest observations. It is the measure of the dispersion in the sample and should reflect the dispersion of the population.

**Definition**

In terms of order statistics, the **sample median**  $M$  is defined by

$$M = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \frac{1}{2} [X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}] & \text{if } n \text{ is even} \end{cases}.$$

**Theorem**

Let  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  denote the order statistics of the random sample  $X_1, X_2, \dots, X_n$  from a continuous population with cdf  $F_X(x)$  and the pdf  $f_X(x)$ . Then the pdf of  $X_{(j)}$  is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j} \text{ for } x \in \mathbb{R}$$

**Theorem**

Let  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  denote the order statistics of the random sample  $X_1, X_2, \dots, X_n$  from a continuous population with cdf  $F_X(x)$  and the pdf  $f_X(x)$ . Then the joint pdf of all the order statistics is given by

$$f_{X_{(1)}, X_{(2)}, \dots, X_{(n)}} = \begin{cases} n! f_X(x_1) \dots f_X(x_n) & -\infty < x_1 < x_2 < \dots < x_n < \infty \\ 0 & \text{otherwise} \end{cases}$$

**Theorem**

Let  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  denote the order statistics of the random sample  $X_1, X_2, \dots, X_n$  from a continuous population with cdf  $F_X(x)$  and the pdf  $f_X(x)$ .

Then the joint pdf of  $X_{(i)}$  and  $X_{(j)}$  where  $1 \leq i \leq j \leq n$  is given by

$$f_{X_{(i)}, X_{(j)}}(u, v) = \begin{cases} \frac{n! f_X(u) f_X(v) [F_X(u)]^{i-1} [1 - F_X(v)]^{n-j} [F_X(v) - F_X(u)]^{j-1-i}}{(i-1)!(j-1-i)!(n-j)!} & -\infty < u < v < \infty \\ 0 & \text{otherwise} \end{cases}$$

**Theorem**

Suppose  $X_1, X_2, \dots, X_n$  are iid random variables with common pdf  $f$  and CDF  $F$ . Let  $g$  be a real valued function such that  $E|g(X)| < \infty$  where  $X \sim F$ . Then for  $1 \leq j \leq n$ ,  $E|g(X_{(j)})|$  exists. Converse holds as well.

**Note**

If  $E|g(X_{(j)})| = \infty$  for some  $j$ , then  $E|g(X)| = \infty$  and conversely, if  $E|g(X)| = \infty$ , then  $E|g(X_{(j)})| = \infty$  for some  $j$ .

**Definition**

Suppose a sequence of random variables  $(X_n)_{n \geq 1}$  and a random variable  $X$  are defined on a probability space  $(\Omega, \mathcal{F}, P)$ . We say that the sequence of random variables

$X_1, X_2, \dots$  **converges in probability** to the random variable  $X$  (written  $X_n \xrightarrow{P} X$ ) if for every  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0$  or equivalently  $\lim_{n \rightarrow \infty} P(|X_n - X| \leq \epsilon) = 1$

**Note**

The random variables in the sequence  $X_1, X_2, \dots$  are typically not iid random variables

**Theorem**

Suppose  $X_n \xrightarrow{p} X$  and  $Y_n \xrightarrow{p} Y$ . Then  $X_n \pm Y_n \xrightarrow{p} X \pm Y$  and  $X_n Y_n \xrightarrow{p} XY$

**Theorem**

Suppose  $X_n \xrightarrow{p} a$ , where  $a$  is a non-zero constant. Then  $\frac{1}{X_n} \xrightarrow{p} \frac{1}{a}$

**Theorem**

Let  $X_n \xrightarrow{p} X$  and  $h$  be a real valued continuous function of a real variable. Then  $h(X_n) \xrightarrow{p} h(X)$

**Definition**

We say that a sequence of estimators  $W_n = W_n(X_1, \dots, X_n)$  is a **consistent sequence of estimators** of the parameter  $\theta$  if  $W_n \xrightarrow{p} \theta$  as  $n \rightarrow \infty$  for each fixed  $\theta \in \Theta$ .

**Note**

This basically means the estimator converges to the proper value as the sample size becomes infinite, i.e. approaches the size of the population itself.

**Theorem** (Weak Law of Large Numbers)

Let  $X_1, X_2, \dots$  be iid random variables with  $EX_i = \theta$ . Define  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Then  $\bar{X}_n \xrightarrow{p} \theta$ .

The weak law of large numbers states that for any population with a finite mean  $\theta$ , the sample mean  $\bar{X}_n$  is a consistent estimator for the population mean  $\theta$ .

**Theorem** (Markov's Inequality)

Let  $X$  be a random variable with finite  $r$ -moment where  $r > 0$ . Then for every  $\epsilon > 0$ ,

$$P(|X| \geq \epsilon) \leq \frac{E|X|^r}{\epsilon^r}$$

**Theorem** (Chebyshev's Inequality)

Let  $X$  be a random variable with finite mean  $\mu$  and finite variance  $\sigma^2$ . Then for every

$$\epsilon > 0, P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

**Theorem**

Suppose we have a sequence  $X_1, X_2, \dots$  of iid random variables with  $EX_i = \mu$  and

$Var(X_i) = \sigma^2 < \infty$ . Define the sample variance  $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ . Then a

sufficient condition that  $S_n^2$  converges in probability to  $\sigma^2$  is that  $Var(S_n^2) \rightarrow 0$  as  $n \rightarrow \infty$ .

**Theorem**

Consider a sequence of estimators  $W_n$  each having finite mean and variance. If  $W_n$  is a sequence of estimators such that  $EW_n \rightarrow \theta$  and  $Var(W_n) \rightarrow 0$  as  $n \rightarrow \infty$ , then  $W_n$  is consistent for  $\theta$ .

**Theorem**

Let  $W_n$  be a consistent sequence of estimators for a parameter  $\theta$ . Let  $a_1, a_2, \dots$  and  $b_1, b_2, \dots$  be sequences of real numbers such that  $a_n \rightarrow 1$  and  $b_n \rightarrow 0$ . Then the sequence  $U_n = a_n W_n + b_n$  is a consistent sequence of estimators of  $\theta$ .

**Theorem**

If  $S_n^2$  is a consistent estimator of  $\sigma^2$ , then the sample standard deviation  $S_n = \sqrt{S_n^2}$  is a consistent estimator of  $\sigma$

**Definition**

Let  $X_1, X_2, \dots, X_n$  be a random sample. The **sample moment of order  $k$**  (where  $k$  is a positive integer) is defined as  $m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$

**Note**

Even if the population does not have any moment, sample moments of all orders exists

**Definition**

We say that a sequence of random variables  $X_1, X_2, \dots$  **converges in distribution** to a random variable  $X$  (written  $X_n \xrightarrow{d} X$ ) if  $\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$  at all points  $x$  where  $F_X(x)$  is continuous.

**Note**

The convergence of distribution functions does not imply the convergence of corresponding PMFs or PDFs.

**Theorem**

Assume that the random variables  $X$  and  $X_n$  (for each  $n$ ) are non-negative and integer valued. Then  $X_n \xrightarrow{d} X \iff \lim_{n \rightarrow \infty} f_{X_n}(k) = f_X(k)$  for all  $k = 0, 1, 2, \dots$

**Theorem**

Let  $(X_n)$  and  $X$  be random variables with pdf such that  $f_{X_n}(x) \rightarrow f_X(x)$  for almost all  $x \in \mathbb{R}$ . Then  $X_n \xrightarrow{d} X$ .

**Note**

Convergence in distribution does not have the usual properties associated with convergence. For example, unless  $X_n$  and  $Y_n$  are independent, then in general  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{d} Y$  does not imply that  $X_n + Y_n \xrightarrow{d} X + Y$



**Theorem**

If  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{d} a$  where  $a$  is a constant, then  $X_n + Y_n \xrightarrow{d} X + a$  and  $X_n Y_n \xrightarrow{d} aX$

**Theorem**

If the sequence of random variables  $X_1, X_2, \dots$  converges in probability to a random variable  $X$ , then the sequence also converges in distribution to  $X$ .

**Note**

The converse of the above theorem is not true.

**Theorem**

Suppose  $X_n \xrightarrow{d} a$  where  $a$  is a constant. Then  $X_n \xrightarrow{p} a$

**Theorem** (Central Limit Theorem)

Let  $X_1, X_2, \dots$  be a sequence of independent and identically distributed random variables, each having a finite mean  $\mu$  and non-zero variance  $\sigma^2$ .

Then  $\frac{Z_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1)$  where  $Z_n = X_1 + X_2 + \dots + X_n$

**Theorem** (Continuity Theorem)

Let  $(X_n)$  be a sequence of random variables with corresponding MGFs  $(M_n)$  and suppose that  $M_n(t)$  exists for  $|t| \leq t_0$  for every  $n$ . If there exists a random variable  $X$  with corresponding MGF  $M$  which exists for  $|t| \leq t_1 \leq t_0$  such that  $M_n(t) \rightarrow M(t)$  as  $n \rightarrow \infty$  for every  $t \in [-t_1, t_1]$  then  $X_n \xrightarrow{d} X$ .

**Definition**

Let  $(X_n)$  be a sequence of random variables. We say that  $X_n$  is **asymptotically normal** with mean  $\mu$  and variance  $\sigma_n^2$  and we write  $X_n$  is  $AN(\mu_n, \sigma_n^2)$  if  $\sigma_n > 0$  and

$$\frac{X_n - \mu}{\sigma_n} \xrightarrow{d} N(0, 1) \text{ as } n \rightarrow \infty.$$

**Note**

Here  $\mu_n$  is not necessarily the mean of  $X_n$  and  $\sigma_n^2$  is not necessarily its variance.

**Theorem** (Delta Method)

Suppose  $Y_n$  is  $AN(\mu, \sigma^2)$  with  $\sigma_n \rightarrow 0$  and  $\mu$  a fixed real number. Let  $g$  be a real-valued function which is differentiable at  $x = \mu$ , with  $g'(\mu) \neq 0$ . Then  $g(Y_n)$  is  $AN(g(\mu), [g'(\mu)]^2 \sigma_n^2)$

**Theorem** ( $k$ -th order Delta Method)

Suppose  $Y_n$  is  $AN(\mu, \sigma^2)$  with  $\sigma_n \rightarrow 0$  and  $\mu$  a fixed real number. Let  $g$  be a real valued function which is differentiable  $k$  times,  $k \geq 1$  at  $x = \mu$  with  $g^{(i)}(\mu) = 0$  for

$1 \leq i \leq k-1$ ,  $g^{(k)}(\mu) \neq 0$ . Then  $\frac{g(Y_n) - g(\mu)}{\frac{1}{k!} g^{(k)}(\mu) \sigma_n^k} \xrightarrow{d} Z^k$  where  $Z \sim N(0, 1)$

---

# Principles of Data Reduction

## Definition

Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a random sample from a population with unknown parameter  $\theta$ . A statistic  $T = T(\mathbf{X})$  is a **sufficient statistic** for  $\theta$  if the conditional distribution of the sample  $\mathbf{X}$  given  $T = t$  does not depend on  $\theta$ , i.e.

$P(\mathbf{X} = (x_1, x_2, \dots, x_n) | T = t)$  does not depend on  $\theta$

## Remark

Sufficient if probability of sample vector given the statistic does not depend on the parameter

## Theorem

Let  $f(\mathbf{x}|\theta)$  denote the joint pdf or pmf of a sample  $\mathbf{X}$  and  $q(t|\theta)$  is the pdf or pmf of a statistic  $T(\mathbf{X})$ . Then  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if for every  $\mathbf{x}$  in the sample

space, the ratio  $\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$  does not depend on  $\theta$ .

## Remark

Sufficient if joint pmf/pdf of sample vector divided by pmf/pdf of statistic does not depend on parameter.

## Theorem (Factorization Theorem)

Let  $f(\mathbf{x}|\theta)$  denote the joint pdf or pmf of a sample  $\mathbf{X}$ . A statistic  $T(\mathbf{X})$  is a sufficient statistic for  $\theta$  if and only if there exist functions  $g(t|\theta)$  and  $h(\mathbf{X})$  such that for all sample point  $\mathbf{x}$  and all parameter points  $\theta$ ,  $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$

## Remark

Sufficient iff the joint pdf of the sample vector can be factorized into a function of the statistic (which depends on the parameter) and a function independent of the parameter. The function of the statistic does not need to be the pdf/pmf of the statistic

## Note

The above theorem helps us construct sufficient statistics instead of guessing at them.

## Note

It is always true that the entire sample is a sufficient statistics since  $T(\mathbf{x}) = \mathbf{x}$  and  $h(\mathbf{x}) = 1$  satisfies the above theorem. But this does not help with data reduction.

### Note

For samples from a normal distribution with parameters  $\mu$  and  $\sigma^2$ ,  $T(\mathbf{X}) = (\bar{X}, S^2)$  is a sufficient statistic for  $(\mu, \sigma^2)$ . Note for other distributions, the sample mean and variance may not be sufficient.

### Note

If  $T$  is sufficient for  $\theta$ , then any one-one function of  $T$  is also sufficient.

### Definition

Let  $\{f(t|\theta), \theta \in \Theta\}$  ( $\theta$  may be a vector) be a family of pdfs or pmfs for a statistic  $T = T(\mathbf{X})$  (here  $T$  may be multidimensional). We say that this family is **complete** if given any real-valued function  $g$  with  $E_\theta g(T) = 0$  for all  $\theta \in \Theta$  then  $P_\theta(g(T) = 0) = 1$  for all  $\theta \in \Theta$ .

### Note

If a statistic  $T = T(\mathbf{X})$  is sufficient for the family of pdfs or pmfs  $\{f(x|\theta)|\theta \in \Theta\}$  then  $T$  is sufficient for any subclass of  $\{f(x|\theta)|\theta \in \Theta\}$ .

This does not hold for completeness, if even one member is removed from the family, it destroys completeness.

### Theorem

If  $T$  is complete, then any one-to-one mapping of  $T$  is also complete.

### Definition

We say that  $\{f(x|\theta)|\theta \in \Theta\}$  is a **one-parameter exponential family** if  $f(x|\theta)$  can be expressed as  $f(x|\theta) = c(\theta)h(x)e^{w(\theta)t(x)}$  where  $h(x) \geq 0$ ,  $t(x)$  are real valued functions of the observation  $x$  which cannot depend on  $\theta$  and  $c(\theta) \geq 0$ ,  $w(\theta)$  are real valued functions of the unknown parameter  $\theta$  which cannot depend on  $x$ .

### Definition

The form in the above definition is not unique.

### Definition

We say that  $\{f(x|\theta)|\theta \in \Theta\}$  is a  **$k$ -parameter exponential family** if  $f(x|\theta)$  can be expressed as  $f(x|\theta) = c(\theta)h(x)e^{\sum_{i=1}^k w_i(\theta)t_i(x)}$  where  $h(x) \geq 0$ ,  $t_i(x)$  are real-valued functions of the observation  $x$  which cannot depend on  $\theta = (\theta_1, \dots, \theta_k)$  and  $c(\theta) \geq 0$ ,  $w_i(\theta)$  are real-valued functions of the unknown  $k$ -dimensional parameter  $\theta$  which cannot depend on  $x$ .

### Definition

We say that  $\{f(x|\theta)|\theta \in \Theta\}$  is a **curved exponential family** if  $f(x|\theta)$  can be expressed in the form  $f(x|\theta) = c(\theta)h(x)e^{\sum_{i=1}^k w_i(\theta)t_i(x)}$  for which the dimension of the vector  $\theta$  is equal to  $d < k$ .

**Theorem**

Let  $X_1, X_2, \dots, X_n$  be iid observations from a pdf or pmf  $f(x|\theta)$  that belongs to an exponential family given by  $f(x|\theta) = c(\theta)h(x)e^{\sum_{i=1}^k w_i(\theta)t_i(x)}$  where  $\theta = (\theta_1, \theta_2, \dots, \theta_d)$  where  $d \leq k$ . Then  $T(X) = \left( \sum_{j=1}^n t_1(X_j), \dots, \sum_{j=1}^n t_k(X_j) \right)$  is a sufficient statistic for  $\theta$

**Theorem**

Let  $X_1, X_2, \dots, X_n$  be iid observations from a pdf or pmf  $f(x|\theta)$  that belongs to a  $k$ -parameter exponential family given by  $f(x|\theta) = c(\theta)h(x)e^{\sum_{i=1}^k w_i(\theta)t_i(x)}$  where  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ . Then the statistic  $T(X) = \left( \sum_{j=1}^n t_1(X_j), \dots, \sum_{j=1}^n t_k(X_j) \right)$  is complete for  $\theta$

**Note**

In the above theorems, curved exponential families are allowed in the sufficient statistic theorem, but only fully exponential families are allowed in the complete statistic theorem.

---

## Point Estimation

**Definition**

A **point estimator** is any function  $W(X_1, \dots, X_n)$  of a sample, i.e. a statistic is a point estimator.

**Note**

An estimator is a function of the sample, while an estimate is the realised value of an estimator that is obtained when the sample is taken.

**Remark**

A statistic is not an estimator. An estimator is a statistic with the added intention of measuring some property of the underlying distribution. Any real valued function of an observable random variable in a sample is called a statistic. Some statistics work well to estimate a property of a distribution.

The goal of a statistic is to summarize the information in a sample, by using sufficient statistics, and the goal of estimators is to estimate the parameters of the population under consideration.

### Result

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with probability function  $f(x|\theta_1, \dots, \theta_k)$ . Method of moments estimators are found by equating first  $k$  sample moments to the corresponding  $k$  population moments, and solving the resulting system of simultaneous equations.

### Note

The method of moments is the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. It is simple to use and almost always yields some sort of estimate. This method usually yields estimators that may be improved upon.

### Note

The method of moments estimators need not be unbiased.

### Definition

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with probability function  $f(x|\theta)$ . The **likelihood function** is defined by

$L(\theta|x_1, x_2, \dots, x_n) = f(x_1|\theta)f(x_2|\theta)\dots f(x_n|\theta)$  where  $x_1, x_2, \dots, x_n$  is the realized value of the random sample  $X_1, X_2, \dots, X_n$ .

### Note

The domain of the likelihood function is the set of all permissible values of the population parameter.

### Definition

For each sample point  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , let  $\hat{\theta}(\mathbf{x})$  be a parameter value at which  $L(\theta|\mathbf{x})$  attains its maximum as a function of  $\theta$ , with  $\mathbf{x}$  held fixed. Then  $\hat{\theta}(\mathbf{X})$  is a **maximum likelihood estimator** of the parameter  $\theta$  based on a random sample  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ .

### Note

The MLE is the parameter point for which the observed sample is most likely.

### Theorem (Invariance Property of MLEs)

If  $\hat{\theta}$  is the MLE of  $\theta$ , then for any function  $\tau(\theta)$ , the MLE of  $\tau(\theta)$  is  $\tau(\hat{\theta})$ .

### Note

The above property also holds in the multivariate case.

### Theorem (Bayes' Theorem)

Let  $\{A_1, A_2, \dots, A_N\}$  be a partition of the sample space and assume that  $P(A_i) > 0$  for all  $i = 1, 2, \dots, N$ .

For any event  $B$  such that  $P(B) > 0$  we have  $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^N P(A_k)P(B|A_k)}$  for each  $i = 1, 2, \dots, N$

Alternatively,  $f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|t)f_X(t)dt}$

### Definition

In Bayesian analysis, before the data is observed, the unknown parameter is modelled as a random variable  $\Theta$  having probability distribution  $\pi(\theta)$  called the **prior distribution**. This distribution represents our prior belief about the value of this parameter. We update the prior distribution after observing the sample to get the **posterior distribution**  $\pi(\theta|x)$ .

### Definition

Let  $\mathcal{F}$  denote the class of probability functions  $f(x|\theta)$  (indexed by  $\theta$ ). A class  $\Pi$  of prior distributions is a conjugate family for  $\mathcal{F}$  if the posterior distribution is in the class  $\Pi$  for all  $f \in \mathcal{F}$ .