Joint Probability Distributions

* → Distribution of two or more RVs.

* For two discrete RVs X & Y,

$P(X=x, Y=y) = P_{XY}(x,y)$ satisfies

① $P(x,y) \geq 0$ ② $\sum_x \sum_y P(x,y) = 1$.

* For two continuous RVs X & Y,

$P((X,Y) \in R) = \iint_R f(x,y)\,dx\,dy$ satisfies,

① $f(x,y) \geq 0$ ② $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$.

* Marginal Probability Distributions

Discrete :

X assumes the values $x_1, x_2, \ldots, x_m$ and

Y assumes the values $y_1, y_2, \ldots y_n$.

$P_X(x) = \sum_{y_k} P(x, y_k)$ | $\sum_{j=1}^{n} P_X(x_j) = 1$

$P_Y(y) = \sum_{x_k} P(x_k, y)$ | $\sum_{k=1}^{n} P_Y(y_k) = 1$.

Continuous :

$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$

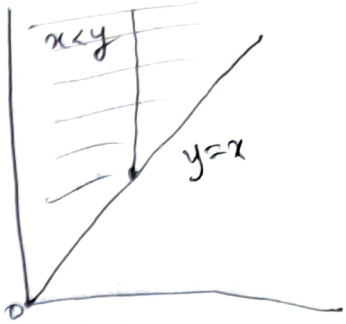
$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$.

Note : When joint distribution is given, we can calculate

$P(a<X<b) = \int_a^b f_X(x)\,dx = \int_a^b \int_{-\infty}^{\infty} f(x,y)\,dy\,dx$.

Eg:

$$f_{XY}(x,y) = \begin{cases} 6\times10^{-6} \exp\left(-0.001x - 0.002y\right), & x<y \\ 0, & \text{otherwise} \end{cases}$$



$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_{0}^{\infty}\int_{x}^{\infty} f(x,y)\,dy\,dx$$

$$= \int_{0}^{\infty} 6\times10^{-6}\left(\int_{x}^{\infty} e^{-0.002y}\,dy\right) e^{-0.001x}\,dx$$

$$= 6\times10^{-6}\int_{0}^{\infty}\left[\frac{e^{-0.002y}}{-0.002}\right]_{x}^{\infty} e^{-0.001x}\,dx$$

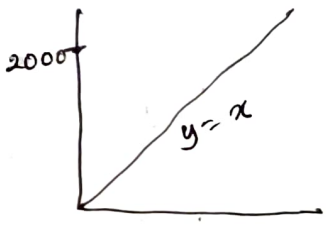
$$= \frac{6\times10^{-6}}{2\times10^{-3}}\int_{0}^{\infty} e^{-0.002x}\cdot e^{-0.001x}\,dx$$

$$= 3\times10^{-3}\int_{0}^{\infty} e^{-0.003x}\,dx$$

$$= \frac{3\times10^{-3}}{-0.003}\left[e^{-0.003x}\right]_{0}^{\infty}$$

$$= \frac{3\times10^{-3}}{-3\times10^{-3}}(-1) = 1.$$

$$P(Y>2000) = \int_{2000}^{\infty}\int_{0}^{y} 6\times10^{-6}\, e^{-0.001x-0.002y}\,dx\,dy$$



$$= 6\times10^{-6}\int_{2000}^{\infty} e^{-0.002y}\left(\int_{0}^{y} e^{-0.001x}\,dx\right)dy$$

$$= \frac{6\times10^{-6}}{10^{-3}}\int_{2000}^{\infty} e^{-0.002y}\left[e^{-0.001x}\right]_{0}^{y}\,dy.$$

$$= \frac{6\times10^{-6}}{10^{-3}}\int_{2000}^{\infty} e^{-0.002y}\left[e^{-0.001y}-1\right]dy$$

$$= 6\times10^{-3}\int_{2000}^{\infty}\left(e^{-0.003y} - e^{-0.002y}\right)dy$$

$$= \frac{6\times10^{-3}}{10^{-3}} \left[\frac{e^{-0.003y}}{-3} + \frac{e^{-0.002y}}{2} \right]_{2000}^{\infty}$$

$$= 6\left[\frac{e^{-4}}{2} - \frac{e^{-6}}{3} \right] = 0.0499$$

OR

$$f_X(y) = \int_0^y 6\times10^{-6}\, e^{-0.001x - 0.002y}\, dx$$

$$= 6\times10^{-3}\, e^{-0.002y}\left(1 - e^{-0.001y}\right)$$

$$P(Y>2000) = \int_{2000}^{\infty} f_y\, dy .$$

Problem

The joint probability fn of two discrete RVs X and Y is given by $f(x,y) = c(2x+y)$, where x and y assumes integers such that $0\le X\le 2$, $0\le y\le 3$ and $f(x,y)=0$ otherwise.

a) Find the value of the constant c.

b) Find $P(X=2, Y=1)$

c) Find $P(X\ge 1, Y\le 2)$

d) Find marginal probability functions of X and Y.

Ans

X \ Y	0	1	2	3	Total
0	0	c	2c	3c	6c
1	2c	3c	4c	5c	14c
2	4c	5c	6c	7c	22c
Total	6c	9c	12c	15c	42c ← Grand total

① We have,

$$\sum_{j=1}^{7} P_X(x_j) = 1 \quad \Rightarrow 42c = 1$$
$$\Rightarrow c = \frac{1}{42}.$$

② $P(X=2, Y=1) = 5c = \frac{5}{42}$.

③ $P(X \geq 1, Y \leq 2) = P(X=1, Y=0) + P(X=1, Y=1) + P(X=1, Y=2)$
$+ P(X=2, Y=0) + P(X=2, Y=1) + P(X=2, Y=2)$

$= 2c + 3c + 4c + 4c + 5c + 6c$

$= 24c$

$= \frac{24}{42} = \frac{4}{7}$.

④ $P(X=x) = P_X(x)$

$P_X(x)$ can be obtained from the marginal totals in the ~~last~~ right hand column of the table

$$P(X=x) = P_X(x) = \begin{cases} 6c = \frac{1}{7} & x=0 \\ 14c = \frac{1}{3} & x=1 \\ 22c = \frac{11}{21} & x=2. \end{cases}$$

$P_Y(y)$ can be obtained from the margin totals in the last row of the table

$$P(Y=y) = P_Y(y) = \begin{cases} 6c = \frac{1}{7} & y=0 \\ 9c = \frac{3}{14} & y=1 \\ 12c = \frac{2}{7} & y=2 \\ 15c = \frac{5}{14} & y=3 \end{cases}$$

Problem

The joint density function of two continuous RV X and Y is

$$f(x,y)=\begin{cases} cxy & 0<x<4,\ 1<y<5 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find c

(b) Find $P(1<X<2,\ 2<Y<3)$

(c) $P(X\geq 3,\ Y\leq 2)$

Ans

(a) We have $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

$$\therefore \int_0^4\int_1^5 cxy\,dx\,dy = 1$$

$$\Rightarrow \int_0^4 cx\left[\frac{y^2}{2}\right]_1^5 dx = 1$$

$$\Rightarrow \int_0^4 cx\ \frac{25-1}{2}\,dx = 1$$

$$\Rightarrow 12c\int_0^4 x\,dx = 1$$

$$\Rightarrow 12c\left[\frac{x^2}{2}\right]_0^4 = 1$$

$$\Rightarrow 6c\times 16 = 1$$

$$\Rightarrow c = \frac{1}{96}.$$

(b) $P(1<X<2,\ 2<Y<3) = \int_{x=1}^{2}\int_{y=2}^{3} cxy\,dx\,dy$

$$= \int_{x=1}^{2}\int_{y=2}^{3} \frac{xy}{96}\,dx\,dy$$

$$= \int_1^2 \frac{x}{96}\left[\frac{y^2}{2}\right]_2^3 dx$$

$$= \int_1^2 \frac{x}{96}\ \frac{9-4}{2}\,dx$$

$$= \frac{5}{2\times 96}\int_1^2 x\,dx$$

$$= \frac{5}{2\times 96}\left[\frac{x^2}{2}\right]_1^2$$

$$= \frac{5}{2\times 96}\left(\frac{4-1}{2}\right) = \frac{5}{128}$$

(c) $P(X\geq 3,\ Y\leq 2) = \int_{x=3}^{4}\int_{y=1}^{2}\frac{xy}{96}\,dx\,dy$

$$= \int_3^4 \frac{x}{96}\left[\frac{y^2}{2}\right]_1^2 dx$$

$$= \frac{1}{96\times 4}\int_3^4 x\,(3)\,dx$$

$$= \frac{3}{96\times 4}\left[\frac{x^2}{2}\right]_3^4$$

$$= \frac{3}{96\times 4}\ \frac{16-9}{2} = \frac{7}{128}.$$

Mean and Variance from a Joint Distribution

Continuous

$$E(X) = \int_{-\infty}^{\infty} x\,f_X(x)\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot f(x,y)\,dy\,dx$$

~~$V(X) = \int_{-\infty}^{\infty} y\,f_Y(y)$~~

$$V(X) = \int_{-\infty}^{\infty}(x-\mu_X)^2 f_X(x)\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)^2 f_{XY}(x,y)\,dy\,dx$$

Discrete :

$$E(X) = \sum_x x\,P_X(x)$$

$$V(X) = \sum_x (x-\mu_X)^2 P_X(x)$$

Problem

Find $E(X)$ and $E(Y)$

X \ Y	1	2	3	
1	0.01	0.04	0.25	$0.3 = f_X(x_1)$
2	0.02	0.03	0.2	$0.17 = f_X(x_2)$
3	0.02	0.1	0.05	$0.25 = f_X(x_3)$
4	0.15	0.1	0.05	$0.28 = f_X(x_4)$
	0.2 = $f_Y(y_1)$	0.25 = $f_Y(y_2)$	0.55 = $f_Y(y_3)$	

$$E(X) = x_1 f_X(x_1) + x_2 f_X(x_2) + x_3 f_X(x_3) + x_4 f_X(x_4)$$
$$= 1(0.3) + 2(0.17) + 3(0.2) + 4(0.28)$$
$$= 2.36.$$

$$E(Y) = y_1 f_Y(y_1) + y_2 f_Y(y_2) + y_3 f_Y(y_3)$$
$$= 1\cdot(0.2) + 2(0.25) + 3(0.55) = 2.35$$

Conditional Probability Distributions

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Conditional probability of Y given $X = x$ is

$$f_{y|x}(y) = \frac{f_{xy}(x,y)}{f_X(x)}$$

Conditional density function.

$$P_{y|x} = \frac{P_{xy}(x,y)}{P_X(x)}$$

Conditional probability functions

Properties

① $f_{y|x}(y) \geq 0$ ② $\int f_{y|x}(y)\,dy = 1$

③ $P(Y \in B \mid X = x) = \int_B f_{y|x}(y)\,dy$ for any set B in the range of Y.

Problem

The joint density function of two continuous RVs X and Y is

$$f(x,y) = \begin{cases} \frac{xy}{96} & 0<x<4,\ 1<y<5 \\ 0 & \text{otherwise} \end{cases}$$

Find the marginal distribution functions of X & Y.

Ans

The marginal distribution function for X is

$$F_X(x) = P(X \le x) = \int_{x=-\infty}^{x} \int_{y=-\infty}^{\infty} f(x,y)\, dx\, dy$$

For $x<0$,

$$F_X(x) = 0.$$

For $1<x<4$,

$$F_X(x) = P(X \le x) = \int_{0}^{x} \int_{y=1}^{5} \frac{xy}{96} dx\, dy$$

$$= \int_0^x \frac{x}{96} \left[\frac{y^2}{2}\right]_1^5 dx =$$

$$= \int_0^x \frac{x}{96} \frac{25-1}{2} dx$$

$$= \frac{12}{96} \left[\frac{x^2}{2}\right]_0^x = \frac{x^2}{16}.$$

For $x \ge 4$,

$$F_Y(x) = \int_0^x \int_{1=y}^{5} \frac{xy}{96} dx\, dy = \int_0^4 \int_1^5 \frac{xy}{96} dx\, dy + \int_4^x \int_1^5 f(x,y)\, dx\, dy$$

$$= \int_0^4 \frac{x}{96} \left[\frac{y^2}{2}\right]_1^5 dx + 0$$

$$= 1.$$

$$\therefore F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{16} & 0 \le x < 4 \\ 1 & x \ge 4 \end{cases}$$

(b) For $y < 1$, $F_Y(y) = 0$

For $1 < y < 5$,

$$F_Y(y) = P(Y \le y) = \int_{x=0}^{4} \int_{y=1}^{y} \frac{xy}{96} \, dx \, dy$$

$$= \frac{1}{96} \int_0^4 x \left[\frac{y^2}{2}\right]_1^y dx$$

$$= \frac{y^2-1}{2 \times 96} \int_0^4 x \, dx$$

$$= \frac{y^2-1}{2 \times 96} \left[\frac{x^2}{2}\right]_0^4 = \frac{y^2-1}{24}$$

For $y \ge 5$, $F_Y(y) = 1$

$$F_Y(y) = \begin{cases} 0 & y < 1 \\ \frac{y^2-1}{24} & 1 \le y < 5 \\ 1 & y \ge 5 \end{cases}$$

Problem

y \ X	1	2	3	
1	0.01	0.02	0.25	0.28
2	0.02	0.03	0.2	0.25
3	0.02	0.1	0.05	0.17
4	0.15	0.1	0.05	0.3
	0.2	0.25	0.55	1

① Find $P(y=1 \mid x=3)$ ② $P(y=2 \mid x=3)$

③ $P(y=3 \mid x=3)$ ④ $P(y=4 \mid x=3)$

Ans

$$P(x=x \mid y=y) = \frac{f_{xy}(x,y)}{f_y(y)}$$

$$P(y=y \mid x=x) = \frac{f_{xy}(x,y)}{f_x(x)}$$

① $P(y=1, x=3) = \frac{P(x=3, y=1)}{P(x=3)} = \frac{P(3,1)}{P(x=3)} = \frac{0.25}{0.55} = 0.454$

② $P(y=2 \mid x=3) = \frac{P(x=3, y=2)}{P(x=3)} = \frac{P(3,2)}{P(x=3)} = \frac{0.2}{0.55} = 0.364$

③ $P(y=3 \mid x=3) = \frac{P(x=3, y=4)}{P(x=3)} = \frac{0.05}{0.55} = 0.091$

④ $P(y=4 \mid x=3) = \frac{P(x=3, y=4)}{P(x=3)} = \frac{0.05}{0.55} = 0.091.$

Problem

Assume that joint probability density function for X & Y is $f_{xy}(x,y) = 6 \times 10^{-6} \exp[-0.001x - 0.002y]$ for $x < y$. Find

① $f_{y|x}(y)$ ② $P(y > 2000 \mid x = 1500)$

Ans

$$f_{y|x}(y) = \frac{f_{xy}(x,y)}{f_x(x)}$$

$$f_X(x) = \int_y 6\times10^{-6} \exp(-0.001x - 0.002y)\, dy$$

$$= \int_x^{\infty} 6\times10^{-6} \exp(-0.001x - 0.002y)\, dy$$

$$= 6\times10^{-6}\, e^{-0.001x} \left[\frac{e^{-0.002x}}{-0.002} \right]_x^{\infty}$$

$$= 0.003\, e^{-0.003x} \qquad , x>0.$$

$$\therefore f_{Y|x}(y) = \frac{f_{xy}(x,y)}{f_X(x)} = \frac{6\times10^{-6} \exp(-0.001x - 0.002y)}{0.003\, e^{-0.003x}}$$

$$= 2\times10^{-3}\, e^{0.002x - 0.002y} \qquad x<y$$

(3) $$P(Y>2000 \mid X=1500) = \int_{2000}^{\infty} f_{Y|x=1500}\, dy$$

$$= \int_{2000}^{\infty} 2\times10^{-3}\, e^{3}\, e^{-0.002y}\, dy$$

$$= 0.368.$$

Conditional Mean and Variance.

The conditional mean of Y given $X=x$, denoted as $E(Y|x)$ or $\mu_{Y|x}$ is

$$E(Y|x) = \int_y y\, f_{Y|x}(y) = \int_{-\infty}^{\infty} \not{f}$$

The conditional variance of Y given $X=x$ is,

$$V(Y|x) = \int_y (y - \mu_{Y|x})^2 f_{Y|x}(y)\, dy = \int_y y^2 f_{Y|x}(y) - \mu^2_{Y|x}$$

Eg:

y \ x	1	2	3
1	0.01	0.02	0.25
2	0.02	0.03	0.2
3	0.02	0.1	0.05
4	0.15	0.1	0.05

$$E(Y|X=1) = \mu_{Y|X=1} = 1(0.01) + 2(0.02) + 3(0.02) + 4(0.15)$$
$$= 0.71$$

$$V(Y|X=1) = (1-0.71)^2\,0.01 + (2-0.71)^2(0.02) + (3-0.71)^2(0.02) + (4-0.71)^2(0.15).$$

Independent R.V.

$$f_{XY}(x,y) = f_X(x)\, f_Y(y) \quad \forall x, y.$$
$$\Updownarrow$$
$$f_{Y|X}(y) = f_Y(y) \quad \forall x,y \ \& \ f_X(x) > 0$$
$$\Updownarrow$$
$$f_{X|Y}(y) = f_X(x) \quad \forall x,y \ \& \ f_Y(y) > 0.$$
$$\Updownarrow$$
$$P(X \in A, Y \in B) = P(X \in A) \times P(Y \in B)$$

Problem

The joint density fn of X on Y is given by

$$p(x,y) = \frac{1}{27}(2x+y) \quad ; \ x = 0,1,2 \quad y = 0,1,2.$$

Are X & Y independent ?

Two RVs X and Y are said to be independent if $P(x,y) = P(x) \cdot P(y) \quad \forall x, y$.

Consider $P(0,0)$

$P(0,0) = 0$

$P(x=0) = \frac{3}{27}$

$P(y=0) = \frac{6}{27}$

$\therefore P(0,0) \neq P(x=0) \times P(y=0)$

$\therefore$ X & Y are not independent.

Eg:

$$f_{XY}(x,y) = 2\times10^{-6}\, e^{-0.001x - 0.002y}, \quad x,y \geq 0$$

$$f_X(x) = \int_0^\infty f_{XY}(x,y)\,dy = 0.01\, e^{-0.001x}$$

$$f_Y(y) = \int_0^\infty f_{XY}(x,y)\,dx = 0.002\, e^{-0.002y}$$

$$f_X(x) f_Y(y) = 2\times10^{-6}\, e^{-0.001x - 0.002y}$$

$$= f_{XY}(x,y)$$

$\therefore$ X & Y are independent.

$$\therefore P(X>1000, Y<1000) = P(X>1000) \times P(Y<1000)$$

$$= \int_{1000}^{\infty} f_X\,dx \times \int_0^{1000} f_Y\,dy$$

$$= e^{-1}\left(1 - e^{-2}\right)$$

Note

$$E\big(h(x,y)\big) = \sum_x \sum_y h(x,y)\, f(x,y)$$

$$\infty \int_x \int_y h(x,y)\, f(x,y)\, dy\, dx .$$

Covariance

Given two R.Vs X & Y. The relationship between them can be established if exists, by using covariance of X & Y.

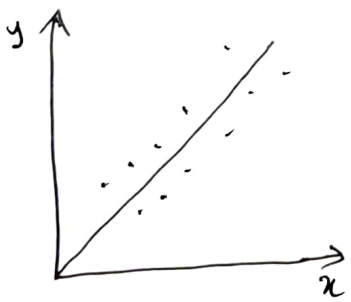
$$\sigma_{xy} = \text{cov}(X,Y) = E\big[(X-\mu_x)(Y-\mu_y)\big]$$

$$= E(XY) - \mu_x \mu_y$$

$$= E(XY) - E(X)\cdot E(Y)$$

Note

Covariance is a linear measure of relationship between X and Y



+ve covariance

-ve covariance

0 covariance

Correlation

It is a normalized measure of covariance defined by

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{V(x)\,V(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\& \quad -1 < \rho_{xy} \leq 1$$

When ρ_{xy} is closer to or equal to ± 1, we say higher or perfect correlation +vely or -vely.

Note

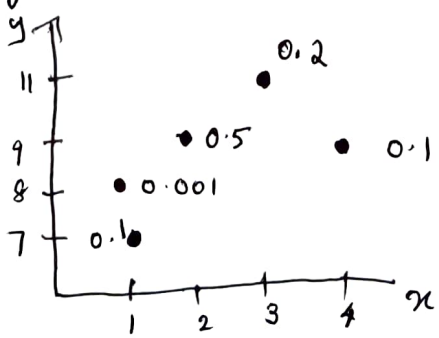
* If X & Y are independent RVs, then

$$\sigma_{xy} = \rho_{xy} = 0.$$

* $\rho_{xy} = 1$ — Perfect positive correlation
* $\rho_{xy} = -1$ — Perfect negative correlation
* $\rho_{xy} = 0$ — No correlation.

Eg:

The joint probability mass fn is given in the figure below.



Calculate ρ_{xy} & σ_{xy}.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\sigma_{xy} = E(xy) - E(x)\,E(y)$$

$$E(xy) = 7(0.1) + (18)(0.5) + 33(0.2) + 8(0.1) + (36)(0.1)$$

$$E(x) = 1(0.2) + 2(0.5) + 3(0.2) + 4(0.1)$$

$$E(y) = 7(0.1) + 8(0.1) + 9(0.6) + 11(0.2)$$

$$\sigma_{xy} = E(xy) - E(x)\,E(y)$$

Result

- $E(C_1X_1 + C_2X_2 + \cdots + C_nX_n) = C_1E(X_1) + C_2E(X_2) + \cdots + C_nE(X_n)$
- $V(C_1X_1 + C_2X_2 + \cdots + C_nX_n) = C_1^2V(X_1) + C_2^2V(X_2) + \cdots + C_n^2V(X_n)$.

Problem

Let X_1 & X_2 are normal RVs representing length and breadth of a rectangle. Given ~~$E(X_1) = 2$~~. $\mu_1 = 2$, $\mu_2 = 5$, $\sigma_1 = 0.1$, $\sigma_2 = 0.2$. X & Y are independent R.V.s. Find $P(\text{Perimeter} > 14.5)$.

Ans

Let Y = perimeter of the rectangle.

$$= 2X_1 + 2X_2.$$

$$\therefore E(Y) = E(2X_1 + 2X_2) = 2E(X_1) + 2E(X_2)$$
$$= 2\mu_1 + 2\mu_2$$
$$= 2(2) + 2(5) = 14 \rightarrow \text{average perimeter.}$$

$$V(Y) = V(2X_1 + 2X_2)$$
$$= 2^2 V(X_1) + 2^2 V(X_2)$$
$$= 2^2 \sigma_1^2 + 2^2 \sigma_2^2$$
$$= 4 \cdot (0.1)^2 + 4 (0.2)^2$$
$$= 0.2$$

$$\therefore P(Y > 14.5) = P\left(\frac{Y - \mu_Y}{\sigma_Y} > \frac{14.5 - 14}{\sqrt{0.2}} \right)$$
$$= P(Z > 1.12) = 0.13 .$$

Results

- For a discrete RV X and a RV $Y = h(X)$ which is 1-1 and onto and given p.m.f P_X of X.

Then

$$f_Y(y) = P(Y = y) = P(X = h^{-1}(y))$$
$$= f_X(h^{-1}(y))$$

$y = h(x)$
$x = h^{-1} y$.

- For a continuous RV. X with pdf f_X. Let $Y = h(X)$, which is one-one and onto.

Then $f_Y(y) = f_X[h^{-1}(y)] \, |J|$,

where J is the Jacobian of the transformation
$= [h^{-1}(y)]'$.

- Let $f_X(x) = \frac{x}{8}$, $0 \le x \le 4$. Find pdf of $Y = h(x) = 2x + 4$.

$$x = h^{-1}(y) = \frac{1}{2}(y-4)$$

$$J = (h^{-1}(y))' = \frac{1}{2}$$

$$\therefore f_X(y) = \frac{\frac{1}{2}(y-4)}{8} \times \frac{1}{2}$$

$$= \frac{y-4}{32}$$

Moments

The r^{th} moment about origin of the RV X is defined as

$$\mu_r' = E(X^r) = \begin{cases} \sum_x x^r p(x) & \rightarrow \text{discrete} \\ \int_{-\infty}^{\infty} x^r f(x)\,dx & \rightarrow \text{continuous} \end{cases}$$

$$\mu_1' = E(X) = \mu$$

$$\mu_2' = E(X^2) \Rightarrow V(X) = \mu_2' - \mu_1^2.$$

Moment Generating Function (MGF)

The MGF of the RV X is the expected value of e^{tx}.

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} p(x) & \rightarrow \text{Discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\,dx & \rightarrow \text{Continuous} \end{cases}$$

and

$$\mu_r' = \left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0} = E(X^r)$$

- Find the mean and variance of a Binomial distribution using MGF.

$$P(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, \ldots, n.$$

$$M_x(t) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} e^{tx}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x}$$

$$= \left[pe^t + (1-p)\right]^n.$$

$$M_x'(t) = n\left[pe^t + (1-p)\right]^{n-1} \times pe^t$$

$$M_x''(t) = n(n-1)\left[pe^t + (1-p)\right]^{n-2} pe^t \cdot pe^t + n\left[pe^t + (1-p)\right]^{n-1} pe^t$$

$$= npe^t\left(pe^t + (1-p)\right)^{n-2}\left((n-1)pe^t + pe^t + (1-p)\right)$$

$$= npe^t\left(pe^t + (1-p)\right)^{n-2}\left(npe^t + (1-p)\right)$$

$$M_x'(0) = E(X) = np\left[p + 1 - p\right] = np$$

$$M_x''(0) = E(X^2) = np\left(p + 1 - p\right)^{n-2}(np + 1 - p)$$

$$= np(np + 1 - p)$$

$$V(X) = E(X^2) - \left[E(X)\right]^2 = n^2p^2 - np - np^2 - n^2p^2$$

$$= np(1-p)$$

$\mu = np$

$\sigma^2 = np(1-p)$

- Find the mean and variance of a normal distribution using MGF.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

$$M_x(t) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} e^{tx}\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[tx - \frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp -\left[\frac{x^2 - 2(\mu+\sigma^2 t)x + \mu^2}{2\sigma^2}\right] dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{(x-(\mu+\sigma^2 t))^2 - 2\mu\sigma^2 t - \sigma^4 t^2}{2\sigma^2}\right)\right] dx$$

$$= \frac{e^{\mu t + \frac{\sigma^2 t^2}{2}}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}\right] dx$$

put $\frac{x-(\mu+\sigma^2 t)}{\sqrt{2}\,\sigma} = u. \Rightarrow dx = \sqrt{2}\,\sigma\, du$

$$= \frac{e^{\mu t + \frac{\sigma^2 t^2}{2}}}{\sigma\sqrt{2}\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-u^2} du\, \sqrt{2}\,\sigma$$

$$= \frac{e^{\mu t + \frac{\sigma^2 t^2}{2}}}{\sqrt{\pi}} \times \sqrt{\pi}$$

$\therefore V(x) = E(x^2) - E(x)^2 = \sigma^2$

$$\therefore M_x(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

$$M_x'(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} (\mu + \sigma^2 t)$$

$$M_x''(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} (\mu + \sigma^2 t)^2 + \left(e^{\mu t + \frac{\sigma^2 t^2}{2}}\right)\sigma^2$$

Note

For two independent RVs X & Y,

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Numerical Summaries of Data

- **Population :**
  Large group of data which needs to be studied
- **Sample :**
  A small part a representation of the population and the process of obtaining which is called sampling
- **Statistical Inference :**
  Inferring certain facts about the population from results found from a sample.
- **Random Sample :**
  If each members of the population has the same chance of being in the sample, then random sample.

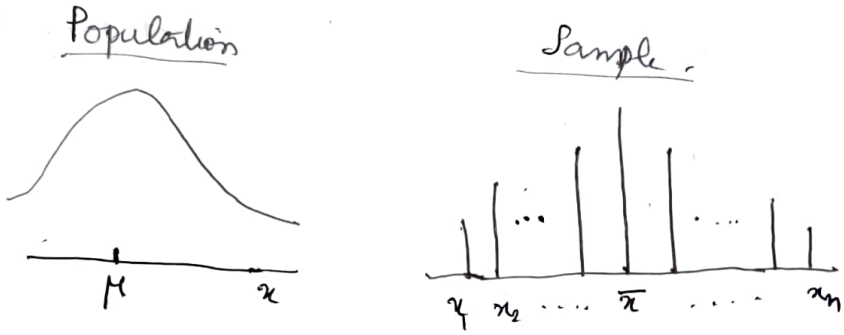
Note :

A population is considered to be known, when we know the prob. distribution of the associated RV. The associated quantities such as $\mu, \sigma^2, \sigma, \beta_2$ etc. are called population parameters, which are known, then we say the population is known.

* Sample Analysis.

- Sample mean : $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Sample variance : $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$.
- Standard deviation : $\sqrt{s^2}$
- Sample range

$$r = \max_i (x_i) - \min_i (x_i)$$



Stem & Leaf diagram.

A stem and leaf diagram is a good way to obtain an informative visual display of a data set $x_1, x_2, \ldots, x_n$, where each number x_i consists of at least two digits.

Construction

1) Divide each number x_i into two parts :
   a stem, consisting of one or more of the leading digits and a leaf consisting of the remaining digit
2) List the stem values in a vertical column

3) Record the leaf for each observation beside its stem.

4) Write the units for stems and leaves on the display.

Eg:

Let $x_1, \ldots, x_m$ are the data & let

$x_i = a_1 a_2 a_3 \ldots a_n$ in decimal form and we divide each x_i into two parts namely stem & leaf.

(Tens) Stem	(ones) Leaf	Freq / Unit
0	7, 9	2
1	3, 5, 8	3
2	5	1
3	2, 3, 2, 7	4
4	1, 5, 9, 8	4
5	4, 3, 7, 8, 5	5
6	1, 4, 5, 8, 9	5
7	1, 2, 3, 4, 5, 6, 7	7
8	9, 8, 7	3
9	2, 5, 9, 8	3.
		N = 37

Actual list : 7, 9, 13, 15, 18, 25, 32, 33, 32, 39, 41, 45, 49, 48, 54, 53, 57, 58, 55, 61, 64, 65, 68, 69, 71, 72, 73, 74, 75, 76, 77, 89, 88, 87, 92, 95, 91.

Tens (Stem)	Ones (Leaf)
0L	
0R	7, 9
1L	3
1R	5, 8

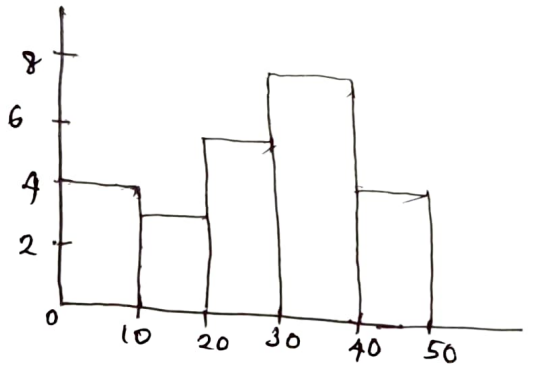
When there are too many leaves in one stem, the stems can be subdivided

(Tens) Stems	(Ones) Leaf
6L	1,4
6R	5,8,9
7L	1,2,3,4
7R	5,6,7.

Histograms

Here, the data is divided into preferably equal interval classes and the data in each class is recorded as frequency.

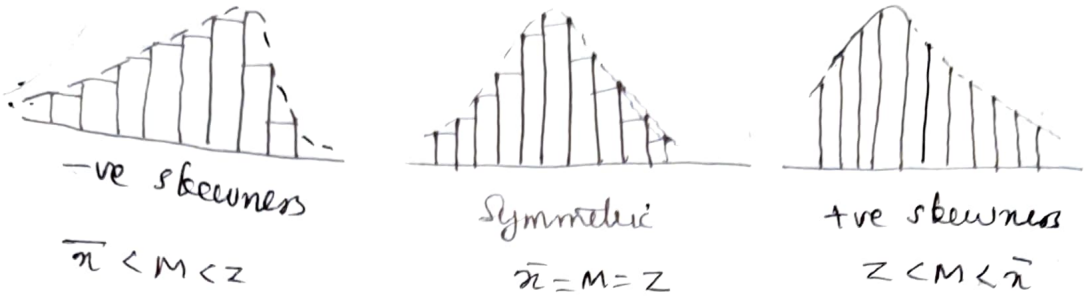
Class :	0-10	10-20	20-30	30-40	40-50
Frequency :	4	3	6	8	4



We can also find the cumulative & relatively by adding up the frequencies consecutively & taking the ratio of a particular frequency to the total, respectively.

Class	Frequency	Cumulative freq.	Relative freq.
0-10	4	4	4/25
10-20	3	7	7/25
20-30	6	13	13/25
30-40	8	21	21/25
40-50	4	25	25/25 = 1
	25		

For large enough data a histogram can also determine the shape & skewness of the data.

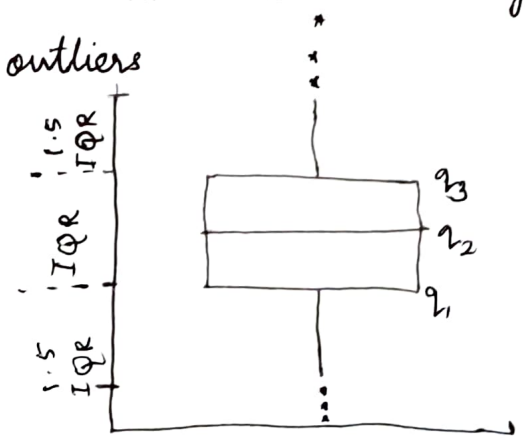


M - median, $\bar{x}$ - mean, z - mode (when it is defined)

Box plots

For a given data set, we determine all the quartiles q_1, q_2, q_3, the difference $q_3 - q_2$ (upper quartile — lower quartile) is determined and is called IQR (Inter Quartile Range)

We draw a diagram with the data on the Y-axis (in the increasing order). We draw a box whose lower side is q_1 & upper side is q_3 and a middle line q_2 ignoring the width. We draw two straight lines above & below the box for the data within the 1.5 IQR range. The data beyond this is denoted as single points and are called outliers

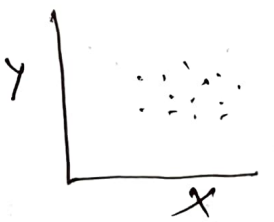


Time Series plot / Time sequence plot

Time series plot is a graph between time and data, from which we can see the evolution of data, we can also see a trend (increasing or decreasing) and cycle (if exists)

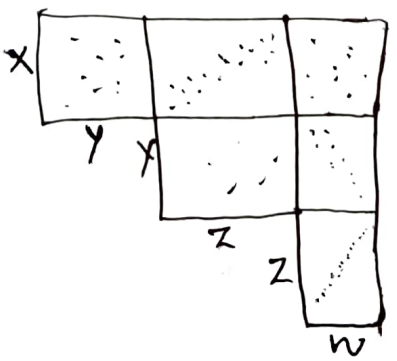
Scatter diagram.

A scatter diagram is constructed by plotting each pair of observations with one measurement in the pair on the vertical axis of the graph and the other measurement in the pair on the horizontal axis



Relation between X & Y

When two or more variables exists, the matrix of scatter diagram may be useful.



Sample correlation coefficient / Karl Pearson Correlation Coeff.

$$\rho_{xy} = \frac{\sum_i \Delta x_i \, \Delta y_i}{\sqrt{\sum_i \Delta x_i^2 \sum \Delta y_i^2}}$$

$$\Delta x_i = x_i - \bar{x}$$

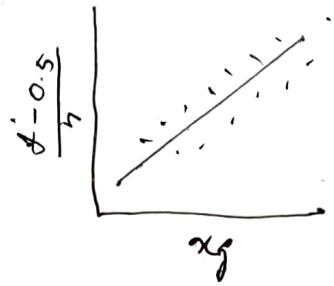
$$\Delta y_i = y_i - \bar{x}$$

$$-1 \le \rho_{xy} \le 1$$

Probability Plot

A probability plot is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of data.

- If a probability plot gives an approximate straight line, then the probability distribution assumed for the probability population is correct.



Point Estimation

A guess of parameter about a population using a sample.

If X is a RV with pdf $f(x)$ and an unknown parameter θ and of $X_1, X_2, \cdots X_n$ is a random sample of size n from X, then

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

is called a point estimator and a particular value is point estimate.

Eg: Consider a large population. Let the RV X denotes the age of the population and

θ : average age of the population $(= \mu = E(X))$

Sample-1	x_1	x_2	....	x_{100}
Sample-2	x_1'	x_2'	...	x_{100}'
⋮				
Sample-i	x_1^i	x_2^i	...	x_{100}^i
	↓	↓		↓
	X_1	X_2		X_{100}

$\theta = h(X_1, X_2, \ldots, X_n)$

↓

R.V.

Random Sample

The random variables $X_1, X_2, \ldots X_n$ are a random sample of size n if

(a) the X_i's are independent RVs and

(b) Every X_i has the same probability distribution

Statistic

A statistic is any function of the observations in a random sample.

$$h(X_1, X_2 \ldots, X_n)$$

Sampling Distribution.

The probability distribution of a statistic is called a sampling distribution.

$$\text{Eg}: \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$P(\bar{x})$ is called sampling distribution of the mean.

Central Limit Theorem.

If $X_1, \ldots, X_n$ is a random sample of size 'n' taken from a population (either finite or infinite) with mean μ and finite variance σ^2 and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$ is the standard normal distribution.

Unbiased estimator

The point estimator $\hat{\Theta}$ is an unbiased estimator for the parameter θ if

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

If the estimator is not unbiased, then the difference

$$E(\hat{\Theta}) - \theta.$$

is called the bias of the estimator $\hat{\Theta}$.

Eg:

The sample mean $\bar{X}$ and var S^2 are unbiased estimators of μ & σ^2

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$

$$= \frac{1}{n}\left(E(X_1) + \cdots + E(X_n)\right)$$

$$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{n\mu}{n} = \mu$$

$$E(S^2) = E\left(\frac{\Sigma (X_i - \bar{X})^2}{n-1}\right)$$

$$= \frac{1}{n-1} E\left(\Sigma (X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right)$$

$$= \frac{1}{n-1} E\left(\Sigma X_i^2 - 2\bar{X}\Sigma X_i + \bar{X}^2\right)$$

$$= \frac{1}{n-1} E\left(\Sigma X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right)$$

$$= \frac{1}{n-1} E\left(\Sigma X_i^2 - n\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left[\{X_i\}^2 - n\bar{X}^2\right]$$

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

$$= \sigma^2$$

Note

An unbiased estimator of a parameter may not be unique.

Eg: mean and median for some sample may estimate μ.

* Minimum variance unbiased estimator (MVUE):

The estimator which has the minimum variance is MVUE.

Result

The sample mean $\bar{x}$ is the MVUE for μ, where X is having a normal distribution.

Standard error of an estimator.

The standard error of an estimator $\hat{\theta}$ is its standard deviation given by $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$.

Eg: Let X be a normal RV.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\therefore \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

If σ is unknown, we can give estimated standard error

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

Mean squared Error of an Estimator.

The Mean Squared Error of an estimator $\hat{\theta}$ of the parameter θ is defined as

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

$$= E\left[(\hat{\theta} - E(\hat{\theta}))^2\right] + \left[\theta - E[\hat{\theta}]\right]^2$$

$$= V(\hat{\theta}) + (\text{bias})^2.$$

For an unbiased estimator, $MSE = \text{Variance}$

Methods of Point Estimation

① Method of moments

Let $X_1, X_2 \ldots, X_n$ be a random sample from the probability distribution $f(x)$ [continuous / discrete]

The k^{th} population moment (or distribution moment) is $E(X^k)$, $k = 1, 2, \ldots$ The corresponding k^{th} sample moment is

$$\frac{1}{n} \sum_{i=1}^{n} X_i^k \quad , k = 1, 2, \ldots$$

First sample moment $= \frac{\sum x_i}{n} = \bar{x}$

First population moments $= \mu$.

Moment Estimators.

Let $X_1, X_2 \cdots, X_n$ be a random sample from either a probability mass fn or a probability density fn with m unknown parameters $\theta_1, \theta_2, \cdots, \theta_m$. The moment estimators $\hat{\theta}_1, \hat{\theta}_2 \cdots, \hat{\theta}_m$ are found by equating the first m population moments to the first m sample moment and solving the resulting equations for the unknown parameters.

Q. Let $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean μ & variance σ^2. Estimate μ & σ^2.

put $\mu = \bar{X}$ and $\mu^2 + \sigma^2 = \frac{1}{n} \sum X_i^2$.

$$\mu = \bar{X} \text{ & } \sigma^2 = \frac{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2}{n}$$

$$= \frac{\sum (x_i - \bar{x})^2}{n}.$$

(2) Method of Maximum Likelyhood :-

Let X be a R.V with pdf $f(x, \theta)$ where θ is the unknown parameter. Let $x_1, x_2, \ldots, x_n$ be a sample of size n. The likelyhood fn of the sample is

$$L(\theta) = f(x_1, \theta)\, f(x_2, \theta) \cdots f(x_n, \theta).$$

The maximum likelyhood estimator (MLE) of θ is the value of θ that maximizes $L(\theta)$.

Q. Bernoulli distribution

$$P(x,p) = \begin{cases} p^x (1-p)^{1-x} & x = 0, 1. \\ 0 & , \text{otherwise} \end{cases}$$

Here p is the parameter. Then the likelyhood fn is

$$L(p) = p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

$$\ln L(p) = (\sum x_i) \ln p + (n - \sum x_i) \ln (1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0.$$

$$\Rightarrow \hat{p} = \frac{1}{n} \sum x_i$$

Q. $X \sim N(\mu, \sigma^2)$. Find MLE for μ & σ^2.

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \sigma \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\therefore L(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum (x-\mu)^2}$$

For maximum we let

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = 0$$

and $\frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = 0$

i.e. $\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$

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

gives

$\hat{\mu} = \bar{x}$ & $\sigma^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2$.