MATH 154

STATISTICAL METHODS II

Semester 2, 2017

Lecture Notes

Compiled by

Sampson Twumasi-Ankrah (Ph.D)

Department of Mathematics

KNUST, Ghana

Learning Outcomes

At the end of this course students should be able to:

1. Demonstrate an understanding of the basic concepts of probability and random variables
2. Understand the concept of the sampling theory in statistic, and in particular the ability to select appropriate sampling method for data collection
3. Apply and interpret basic summary and modelling techniques for bivariate data and use inferential methods in the context of simple linear models
4. Interpret and analyse data that may be displayed in a two—way table.

Course Content

Topics include Random variables: discrete and continuous random variable and probability distributions, Approximation of Normal distribution to Poisson and Binomial distributions, introduction to Sampling theory, introduction to Correlation and Regression analysis and analysis of categorical data.

Table of Contents

UNIT 1

RANDOM VARIABLES

Let S be the sample space associated with a given random experiment. By a random variable we mean a real number X connected with the outcome of a random experiment, i.e. a random variable is variable which assigns a real value to each outcome of a random experiment.

For example: let E be the random experiment consisting of two tosses of a coin

$S = \{HH, HT, TH, TT\}$

We may define the random variable X which denotes the number of heads (0, 1 or 2)

$X = \{2, 1, 1, 0\}$

DISTRIBUTION FUNCTION

Let X be a random variable, then the function

$F_x(x) = F(x) = P(X \leq x) = P\{w : x(w) \leq x\}, -\infty < x < \infty$

is called the distribution of X.

SESSION 1.1: DISCRETE RANDOM VARIABLE

If a random variable X takes at most a countable number of values or countably infinite number of values, it is called *discrete random variable*. In other words, a real valued function defined on a discrete sample space is called a *discrete random variable.*

Example

Otitis media, a disease of the middle ear, is one of the most frequent reasons for visiting a doctor in the first two years of life other than a routine well-baby visit. Let X be the random variable that represents the number of episodes of otitis media in the first 2 years of life. Then X is a discrete random variable, which takes on the values 0, 1, 2, and so on.

PROBABILITY MASS FUNCTION (PMF)

Suppose X is an one-dimensional random variable taking at most a countably infinite number of values $x_1, x_2, ...$ With each possible outcome x_i we associate a number $p_i, P(X = x_i) = p(x_i) = p_i$, called the probability of x_i

The function $p(x_i), i = 1,2....$ satisfying the conditions

i. $p(x_i) \geq 0 \forall i$

ii. $\sum_{i=1}^{\infty} p(x_i) = 1$

is called the probability mass function or probability function of the random variable X. The collection of pairs $\{x_i, p_i\} i = 1,2,3..$ is called the probability distribution of the random variable X.

DISCRETE DISTRIBUTION FUNCTION

The distribution function of the random variable X with PMF $p(x_i), i = 1,2,3...$ is defined as

$$F(x_i) = \sum_{i: x_i \leq x} p(x_i)$$

Note:

i. $p(x_i) = P(X = x_i) = F(x_i) - F(x_{i-1})$, where F is the distribution function of the random variable X

ii. Mean of the random variable $X = E(X) = \sum_x xP(x)$

iii. Variance of the random variable X

$$Var(x) = \sum x^2 p(x) - \left(\sum xp[x]\right)^2$$

Example

The number of patients seen in the ENT department at Komfo Anokye Teaching Hospital in any given hour is a random variable represented by *x*. The probability distribution for *x* is:

x	10	11	12	13	14
$P(x)$	0.4	0.2	0.2	0.1	0.1

Find the probability that in a given hour:

a. exactly 14 patients arrive

b. At least 12 patients arrive

c. At most 11 patients arrive

solution

a. $P(x = 14) = 0.1$

b. $P(x \geq 12) = (0.2 + 0.1 + 0.1) = 0.4$

c. $P(x \leq 11) = (0.4 + 0.2)$

Example

Suppose from previous experience with hypertension drug, the drug company expects that for any clinical practice the probability that 0 patients out of 4 will be brought under control is 0.008, 1 patient out of 4 is 0.076, 2 patients out of 4 is 0.265, 3 patients out of 4 is 0.411, and all 4 patients is 0.240. This probability mass function, or probability distribution is displayed below

$\Pr(X = r)$	0.008	0.076	0.265	0.411	0.240
r	0	1	2	3	4

Notice that for any probability mass function, the probability of any particular value must be between 0 and 1 and the sum of the probabilities of all values must exactly equal 1. Thus $0 < \Pr(X = r) \le 1$, $\sum \Pr(\mathrm{X} = \mathrm{r}) = 1$, where the summation is taken over all possible values that have positive probability.

a. For any clinical practice, what is the probability that between 0 and 4 hypertensives are brought under control?
b. Find the expected value for the random variable

Solution

a. 0.076+0.265+0.411=0.752
b. The expected value of discrete random variable

$$E(X) \equiv \mu = \sum_{i=1}^{n} x_i \Pr(X = x_i)$$

$$= 0(0.008) + 1(0.076) + 2(0.265) + 3(0.411) + 4(0.240) = 2.80$$

Thus on the average about 2.8 hypertensives would be expected to be brought under control for every 4 who are treated.

Example

Consider the random variable representing the number of episodes of diarrhoea in the first 2 years of life. Suppose this random variable has a probability mass function as below

r	0	1	2	3	4	5	6
$\Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

a. What is the expected number of episodes of diarrhoea in the first 2 years of life?

b. Compute the variance and SD for the random variable representing number of episodes of diarrhoea in the first 2 years of life

Solution

a. E(X)=0(.129)+1(.264)+2(.271)+3(.185)+4(.095)+5(.039)+6(.017)=2.038

Thus, on the average a child would be expected to have 2 episodes of diarrhoea in the first 2 years of life.

$$E(X^2)=\sum_{i=1}^{R} r^2 p(r)=(0^2\times .129)+(1^2\times .264)+(2^2\times .271)+(3^3\times .185)+(4^2\times .095)$$

b. $+(5^2\times .039)+(6^2\times .017)$

$=6.12$

$Var(X)=E(X^2)-[E(X)]^2=6.12-(2.038)^2=1.967$

The standard deviation of X is, $\sigma=\sqrt{1.967}=1.402$

Example

If X is a discrete random variable having the probability distribution

$X=x$	1	2	3
$P(X=x)$	k	$2k$	k

Find $P(X\le 2)$

Solution

We know that

$$\sum P(X=x)=1 \Rightarrow 4k=1$$

$$k=\frac{1}{4}$$

$$P(X\le 2)=P(X=1)+P(X=2)\Rightarrow P(X\le 2)=\frac{3}{4}$$

Example

If X is a discrete random variable having the PMF

x	-1	0	1
$P(x)$	k	$2k$	$3k$

Find $P(X\ge 0)$

Solution

We know that

$$\sum P(X=x)=P(X=-1)+P(X=0)+P(X=1)$$

$$k+2k+3k=1\Rightarrow 6k=1$$

$$k=\frac{1}{6}$$

$$P(X\ge 0)=P(X=0)+P(X=1)=2k+3k=5k$$

$$\Rightarrow P(X\ge 0)=\frac{5}{6}$$

Example

If X is a discrete random variable with the following probability distribution

x	1	2	3	4
$P(x)$	a	$2a$	$3a$	$4a$

Find $P(2 < X < 4)$

Solution

We know that

$$\sum P(X = x) = 1$$

$$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1$$

$$10a = 1$$

$$a = \frac{1}{10}$$

$$P(2 < X < 4) = P(X = 3) = 3a = \frac{3}{10}$$

Example

If the probability distribution of X is given as:

x	1	2	3	4
$p(x)$	0.4	0.3	0.2	0.1

Find $P\left(\frac{1}{2} < X < \frac{7}{2} \middle| X > 1\right)$

Solution

By definition

$$P\left(\frac{1}{2} < X < \frac{7}{2}\middle| X > 1\right) = \frac{P\left(\frac{1}{2} < X < \frac{7}{2} \cap X > 1\right)}{P(X > 1)}$$

$$= \frac{P\left(1 < X < \frac{7}{2}\right)}{P(X > 1)}$$

$$= \frac{P(X = 2) + P(X = 3)}{1 - P(X \leq 1)}$$

$$= \frac{P(X = 2) + P(X = 3)}{1 - P(X = 1)}$$

$$\frac{0.5}{0.6} = \frac{5}{6}$$

Example

A random variable X has the probability function

x	-2	-1	0	1
$p(x)$	0.4	k	0.2	0.3

Find k and the mean of X

Solution

$$\sum p(x) = 1 \Rightarrow 0.9 + k = 1 \Rightarrow k = 0.1$$

$$mean = \sum xp(x) = -0.8 - 0.1 + 0 + 0.3 = -0.6$$

Example

If $P(X = x) = \begin{cases} kx, & x = 1, 2, 3, 4, 5 \\ 0, & otherwise \end{cases}$ **represents a probability function, Find**

i. k

ii. $P(X$ being a prime number)

iii. $P\left(\frac{1}{2} < X < \frac{5}{2} \middle| X > 1\right)$

iv. The distribution function.

Solution

i. Given

x	1	2	3	4	5
$P(X = x)$	k	$2k$	$3k$	$4k$	$5k$

Since $p(x)$ is a probability function, $\sum p(x) = 1$

$k + 2k + 3k + 4k + 5k = 1$

$$15k = 1$$

$$k = \frac{1}{15}$$

ii. P(X being a prime number)= $P(X = 1,2,3)$

$$= P(X = 2) + P(X = 3) + P(X = 5)$$

$$= 2k + 3k + 5k = 10k$$

Since $k = \frac{1}{15}, \Rightarrow P(X$ being a prime number$) = 10\left(\frac{1}{15}\right) = \frac{10}{15} = \frac{2}{3}$

iii. $P\left(\frac{1}{2} < X < \frac{5}{2} \middle| X > 1\right) = \frac{P\left(\frac{1}{2}<X<\frac{5}{2} \cap X>1\right)}{P(X>1)}$

$$= \frac{P\left(1 < X < \frac{5}{2}\right)}{P(X > 1)}$$

$$= \frac{P(X = 2)}{P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)}$$

$$= \frac{2k}{14k} = \frac{2}{14} = \frac{1}{7}$$

iv. The distribution function is

$f(x) = 0, x < 1$

$= \frac{1}{15}, 1 \le x < 2$

$= \frac{3}{15}, 2 \le x < 3$

$= \frac{6}{15}, 3 \le x < 4$

$= \frac{10}{15}, 4 \le x < 5$

$= 1, x \ge 5$

Example

If the CDF of a random variable is given by

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

Find $P(X > 1|X < 3)$

Solution

We know that

$$P(X > 1|X < 3) = \frac{P(X > 1 \cap X < 3)}{P(X < 3)}$$

$$= \frac{P(1 < X < 3)}{P(X < 3)}$$

$$= \frac{F(3) - F(1)}{F(3)}$$

$$= \frac{\frac{9}{16} - \frac{1}{16}}{\frac{9}{16}} = \frac{8}{9}$$

Example

Suppose that the random variable assumes three values 0, 1 and 2 with probabilities 1/3, 1/6, and ½ respectively. Obtain the distribution function of X

Solution

x	0	1	2
$P(X = x)$	1/3	1/6	1/2

The distribution $F(x) = P(X \leq x)$ is

$$F(x) = 0, x < 0$$

$$= \frac{1}{3}, 0 \leq x < 1$$

$$= \frac{3}{6}, 1 \leq x < 2$$

$$= 1, x \geq 2$$

Example

If the probability mass function of a random variable is given by

$P(X = r) = kr^3, r = 1,2,3,4$. Find

i. The value of k

ii. $P[(1/2 < X < 5/2)/X > 1]$,

iii. The mean and variance of X

iv. The distribution of the function X

Solution

Given $P(X = r) = kr^3, r = 1,2,3,4$

r	1	2	3	4
$p(r)$	k	$8k$	$27k$	$64k$

i. To find the value of k, we know that

$$\sum_{r=1}^{4} p(r) = 1$$

$$= k + 8k + 27k + 64k$$

$$100k = 1 \Rightarrow k = \frac{1}{100}$$

ii. $P[(1/2 < X < 5/2)/X > 1]$

$$= \frac{P(X = 1, X = 2) \cap P(X = 2, X = 3, X = 4)}{P(X = 2) + P(X = 3) + P(X = 4)}$$

$$= \frac{P(X = 2)}{P(X = 2) + P(X = 3) + P(X = 4)}$$

$$= \frac{\frac{8}{100}}{\frac{8}{100} + \frac{27}{100} + \frac{64}{100}} = \frac{8}{99}$$

iii. $$mean = E(X) = \sum_{r=1}^{4} rp(r) = 1 \times k + 2 \times 8k + 3 \times 27k + 4 \times 64k$$

$$= 354k = \frac{354}{100} = 3.54$$

iv. $$E(X^2) = \sum_{r=1}^{4} r^2 p(r) = 1 \times k + 4 \times 8k + 9 \times 27k + 16 \times 64k$$

$$= 1300k$$

$$= \frac{1300}{100} = 13$$

$$Var(X) = E(X^2) - [E(X)]^2 = 13 - (3.54)^2 = 0.4684$$

v. The distribution function of X is

$$F(x) = 0, x < 1$$

$$= \frac{1}{100}, 1 \le x \le 2$$

$$= \frac{9}{100}, 2 \le x < 3$$

$$= \frac{36}{100}, 3 \le x < 4$$

$$= 1, x \ge 4$$

SESSION 1.2: CONTINUOUS RANDOM VARIABLE

A random variable X is said to be continuous if it can take all possible values between certain limits. In other words, a random variable is said to be continuous when its different values cannot be put in one to one correspondence with a set of positive integers. Examples of continuous random variable are height, weight, age, etc.

Note: The sample space of the continuous random variable must be continuous and cannot be discrete. In most of the practical problems, continuous random variable represent measured data, such as all possible heights, weights, temperature, etc. whereas discrete random variables represent count data such as the number of defectives in a sample and so on.

PROBABILITY DENSITY FUNCTION (PDF)

Consider the small interval $\left(x-\frac{\Delta x}{2}, x+\frac{\Delta x}{2} \right)$ of length Δx round the point x. Let $f(x)$ be any continuous function of x so that $f(x)dx$ represents the probability that x falls in the infinitesimal interval $\left(x-\frac{\Delta x}{2}, x+\frac{\Delta x}{2} \right)$, which is denoted by

$$P\left(x-\frac{\Delta x}{2} \le x \le +\frac{\Delta x}{2} \right) = f(x)dx$$

Let $f(x)dx$ represent the area bounded by the curve $y = f(x), x$ axis and the ordinates at the points $x-\frac{\Delta x}{2}$ and $x+\frac{\Delta x}{2}$. The function $f(x)$ so defined is known as probability density function or density function of the random variable X.

The probability density function of a random variable X denoted by $f(x)$ has the following properties

i. $f(x) \ge 0, \forall x \in R$

ii. $\int_{-\infty}^{\infty} f(x)dx = 1$

iii. $P(a < X < b) = \int_a^b f(x)dx$

Note: In case of continuous random variables, the probability at a point is always zero, i.e. $P(X = a) = 0$ for all possible values of a.

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

The cumulative distribution $F(x)$ of a continuous random variable X with PDF $f(x)$ is given by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx, -\infty < x < \infty$$

Note: $P(a < x < b) = F(b) - F(a)$

The relation between the CDF and PDF is

$$f(x) = \frac{d}{dx}F(x)$$

If X is continuous random variable with PDF $f(x)$, then

$$mean = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$$

$$Var(X) = E(X^2) - [E(X)]^2$$

Example

Verify whether $f(x) = \begin{cases} |x|, & -1 \le x \le 1 \\ 0, & elsewhere \end{cases}$ can be PDF of a continuous random variable

Solution

For $f(x)$ to be a PDF, it should satisfy

i. $f(x) = |x| \ge 0, \forall x$ ii. $\int_{-\infty}^{\infty} f(x)dx = 1$

Given

i. $f(x)|x \geq 0, \forall x|$

ii. $\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{1} |x|dx = 2\int_{0}^{1} xdx = 2\left[\frac{x^2}{2}\right]_0^1 = 1$

Therefore, $f(x)$ can be the PDF of X

Example

A random variable X has the PDF $f(x)$ given by

$$f(x) = \begin{cases} cxe^{-2}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Find the value of c and CDF of x

Solution

If $f(x)$ is a PDF, then

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{0}^{\infty} cxe^{-x}dx = 1$$

$$c\left[x\left(\frac{e^{-x}}{-1}\right) - 1\left(\frac{e^{-x}}{1}\right)\right]_0^{\infty} = 1$$

$$c(0+1) = 1$$

$$c = 1$$

$$f(x) = xe^{-x}, x > 0$$

$$F(x) = P(X \leq x) = \int_{0}^{x} xe^{-x}dx$$

The CDF of X= $= \left[x\left(\frac{ex^{-x}}{-1}\right) - 1\left(\frac{e^{-x}}{1}\right)\right]_0^{x}$

$$= (-xe^{-x} - e^{-x}) - (0-1)$$

$$= 0, otherwise$$

Example

A continuous random variable X follows the probability law $f(x) = ax^2, 0 \le x \le 1$ Determine a and find the probability that x lies between $\frac{1}{4}$ and $\frac{1}{2}$

Solution

If $f(x)$ is a PDF, then

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{-\infty}^{0} f(x)dx + \int_{0}^{1} f(x)dx + \int_{1}^{\infty} f(x)dx = 1 \Rightarrow \int_{0}^{1} f(x)dx = 1$$

$$\int_{0}^{1} ax^2dx = 1 \Rightarrow a\left[\frac{x^3}{3}\right]_0^1 = 1$$

$$P\left(\frac{1}{4} \le x \le \frac{1}{2}\right) = \int_{\frac{1}{4}}^{\frac{1}{2}} f(x)dx$$

$$= \int_{\frac{1}{4}}^{\frac{1}{2}} 3x^2dx$$

$$= 3\left[\frac{x^3}{3}\right]_{\frac{1}{4}}^{\frac{1}{2}}$$

$$= 3\left[\frac{\left(\frac{1}{2}\right)^3}{3} - \frac{\left(\frac{1}{4}\right)^3}{3}\right] = \frac{7}{64}$$

Example

If the PDF of a random variable X is $f(x) = \frac{x}{2}$ in $0 \le x \le 2$. Find $P(X > 1.5 / X > 1)$

$$\mathrm{P}\left(X > \frac{1}{5}\middle|X > 1\right) = \frac{P(X > 1.5 \ \cap \ X > 1)}{P(X > 1)}$$

$$= \frac{P(X > 1.5)}{P(X > 1)}$$

$$= \frac{\int_{1.5}^{2} \frac{x}{2} dx}{\int_{1}^{2} \frac{x}{2} dx}$$

$$= \frac{4 - 2.25}{4 - 1} = 0.5833$$

Example

If $f(x) = kx^2$, $0 < x < 3$is to be the density function, find the value k

Solution

If $f(x)$is a PDF, then

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{0}^{3} kx^2 dx = 1 \Rightarrow 9k = 1$$

$$k = \frac{1}{9}$$

Example

If the CDF of a random variable X is given by $f(x) = 0$ for $x < 0; = \frac{x^2}{16}$ for $0 \le x < 4,$ and $= 1$ for $x \ge 4,$ find $P(X > 1 / X < 3)$

Solution

Given

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

Using $\mathrm{P}(A|B) = \frac{P(A \cap B)}{P(B)}$ we get

$$\mathrm{P}(X > 1|X < 3) = \frac{P(X > 1 \cap X < 3)}{P(X < 3)}$$

$$= \frac{P(1 < X < 3)}{P(0 < X < 3)}$$

$$= \frac{F(3) - F(1)}{F(3) - F(0)}$$

$$P(X > 1/X < 3) = \frac{\frac{8}{16}}{\frac{9}{16}} = \frac{8}{9}$$

Example

The cumulative distribution of X is $F(x) = \frac{x^3 + 1}{9}, -1 < X < 2 \text{ and} = 0, \textit{otherwise}.$ Find $P(0 < X < 1)$

Solution

Using $P(a < X < b) = F(b) - F(a),$ we get

$$P(0 < X < 1) = F(1) - F(0) = \frac{2}{9} - \frac{1}{9} = \frac{1}{9}$$

Example

The CDF of X is given by $F(x) = \begin{cases} 0, x > 0 \\ x^2, 0 \le x \le 1 \\ 1, x > 1 \end{cases}$

Find the PDF of X and obtain $P(X > 0.75)$

Solution

Given $F(x) = \begin{cases} 0, x > 0 \\ x^2, 0 \le x \le 1 \\ 1, x > 1 \end{cases}$

$$f(x) = \frac{d}{dx} F(x) = \begin{cases} 2x, 0 \le x \le 1 \\ 0, otherwise \end{cases}$$

$$P(X > 0.75) = 1 - P(X \le 0.75)$$

$$= 1 - F(0.75) = 1 - (0.75)^2 = 0.4375$$

Example

Verify whether $f(x) = 1 - |1 - x|$, for $0 < x < 2$ is a PDF of a random variable X

Solution

If $f(x)$ is a PDF, then

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\int_{-\infty}^{\infty} f(x) = \int_{0}^{1} [1-(1-x)]dx + \int_{1}^{2} [1+(1-x)]dx$$

$$= \int_{0}^{1} xdx + \int_{1}^{2} (2-x)dx$$

$$= \left[\frac{x^2}{2}\right]_0^1 + \left[2x - \frac{x^2}{2}\right]_1^2 = \frac{1}{2} + \left(2 - \frac{3}{2}\right) = \frac{1}{2} + \frac{1}{2} = 1$$

$\therefore f(x)$ is a PDF

Example

If $f(x) = kx^2,\ 0 < x < 3$ is to be the density function, find the value of k

Solution

If $f(x)$ is a PDF, then

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_{0}^{3} kx^2 dx = 1 \Rightarrow 9k = 1$$

$$k = \frac{1}{9}$$

SESSION 1.3: MATHEMATICAL EXPECTATION

Mathematical expectation of a random variable is obtained by multiplying each probable value of the variable by its corresponding probability and then adding these products.

Let a random variable X assumes the values $x_1, x_2, ...x_n$ with probabilities $p_1, p_2, ...p_n$ respectively. The mathematical expectation of the variable X is defined as:

$E(X) = x_1 p_1 + x_2 p_2 + ... + x_n p_n = \sum x_i p_i$

Note *sd*

1. Sometimes E(X) is also known as Expected value of X
2. Expected value of X is a population mean. If population mean μ is then $E(X) = \mu$

THEOREMS ON MATHEMATICAL EXPECTATION

- ***Theorem 1***: Expected value of constant term is constant, that is, if C is constant, then

E(C)=C

- ***Theorem 2***: If C is constant, then

$E(CX) = CE(X)$

- ***Theorem 3***: If a and b are constants, then

$E(aX \pm b) = aE(X) \pm b$

- ***Theorem 4***: If a, b and c are constants, then

$$E\left(\frac{aX + b}{c}\right) = \frac{1}{c}[aE(X) + b]$$

- ***Theorem 5***: If X and Yare any two random variables then,

$E(X + Y) = E(X) + E(Y)$

- ***Theorem 6***: If X and Y are two independent random variables, then

$$E(X.Y) = E(X).E(Y)$$

Remarks

1. If g(x) is any function of random variable X and f(x) is probability density function then:

$$E[g(x)] = \sum \{g(x).f(x)\}$$

2. $E[X - E(x)] = 0$ that is if $E(X)_{\{} = \mu$ then: $E(X - \mu) = 0$
3. $E\left(\frac{1}{x}\right)$ and $\left(\frac{1}{E(x)}\right)$ are not same

VARIANCE

Variance of the probability distribution of a random variable X is the mathematical expectation of $[X - E(X)]^2$ .Then

$$Var(X) = E[X - E(X)]^2$$

$$= [x_1 - E(X)^2].p(x_1) + [x_2 - E(X)]^2.p(x_2) + ...[x_n - E(X)]^2.p(x_n)$$

$$= \sum_{i=1}^{n} \{[x_i - E(X)]^{2} \times p(x_i)\}$$

Hence $Var(X) = E[X - E(X)]^2$

Another form of variance:

If X be random variable with first two moments $E(X) = \mu$ and $E(X^2) = \mu_2$ then the mathematical expectation of $(X - \mu)^2$ is defined to be the variance of the random variable X. Then

$$Var(X) = E[X - \mu]^2 = E(X^2 - 2X\mu + \mu^2) = E(X^2) - 2E(X)\mu + \mu^2$$

$$= \mu_2 - 2\mu\mu + \mu^2 = \mu_2 - 2\mu^2 + \mu^2 = \mu_2 - \mu^2 = E(X^2) - [E(X)]^2$$

$$Var(X) = E(X^2) - [E(X)]^2$$

$$= E(X^2) - \mu^2$$

Standard deviation: the standard deviation of the probability distribution of a random variable X is the positive square-root of the variance of that random variable .

Standard deviation: $\sigma = \sqrt{E(X^2) - [E(X)]^2}$ or $\sigma = \sqrt{E(X^2) - \mu^2}$

Or Standard deviation: $\sigma = \sqrt{E[X - E(X)]^2}$

Note: the variance of the random variable X is also denoted by V(X)

THEOREMS ON VARIANCE OF A RANDOM VARIABLE

- ***Theorem 7***: If C is a constant, then,

$$Var(CX) = C^2 Var(X)$$

- ***Theorem 8***: Variance of constant is zero i.e.,

$$Var(C) = 0$$

- ***Theorem 9***: If X is a random variable and C is a constant, then

$$Var(X + C) = Var(X)$$

- **Theorem 10**: If a and b are constants, then:

$$Var(aX + b) = a^2 Var(X)$$

- ***Theorem 11***: If X and Y are two independent random variables, then;

  i. $Var(X + Y) = Var(X) + Var(Y)$

  ii. $Var(X - Y) = Var(X) - Var(Y)$

MEAN AND VARIANCE OF A LINEAR COMBINATION

If $Z = aX + bY$ be a linear combination of two random of two random variables X and Y, then

Mean: $\mu_z = aE(X) + bE(Y)$

Example

For a random variable X, $p(x) = \frac{x}{x+1}$ where $x = 1,2,3$. Is $p(x)$ a probability density function**?**

Solution

Here $p(x) = \frac{x}{x+1}$

$\therefore x = 1,2,3,$ p(x) will take values $\frac{1}{2}, \frac{2}{3}$ *and* $\frac{3}{4}$

$\sum p(x) = \frac{1}{2} + \frac{2}{3} + \frac{3}{4} = \frac{23}{12} > 1.$ Now $\sum$ p(x) > 1. Hence p(x) is not a probability function.

Example

The probability distribution of a random variable x is given below. Find

$i. E(x)$ $ii. Var(X)$ $iii. E(2x-3)$ $iv. Var(2x-3)$

x	-2	-1	0	1	2
$P(x)$	0.2	0.1	0.3	0.3	0.1

Solution

Table: Computation of E(x) and V(x)

x	$p(x)$	$xp(x)$	x^2	$x^2 p(x)$
-2 -1 0 1 2	0.2 0.1 0.3 0.3 0.1	-0.4 -0.1 0.0 0.3 0.2	4 1 0 1 4	0.8 0.1 0.0 0.3 0.4
		$\sum xp(x)$		$\sum x^2 p(x) = 1.6$

i. $E(x) = \sum xp(x) = 0$

ii. $E(x)^2 = \sum x^2 p(x) = 1.6$

$Var(X) = E(x)^2 - [E(x)]^2 = 1.6 - 0 = 1.6$

iii. $E(2x-3) = 2E(x) - 3 = 2(0) - 3 = -3$

iv. $Var(2x-3) = 2^2 V(x) + V(-3) = 4(1.6) - 0 = 6.4$

UNIT 2

APPROXIMATION OF DISTRIBUTIONS

SESSION 2.1: THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

A normal distribution is often used to solve problems that involve the binomial distribution since when n is large (say, 100), the calculations are too difficult to do by hand using the binomial distribution. Recall that a binomial distribution has the following characteristics:

1) There must be a fixed number of trials.
2) The outcome of each trial must be independent.
3) Each experiment can have only two outcomes or outcomes that can be reduced to two outcomes.
4) The probability of a success must remain the same for each trial.

Also, recall that a binomial distribution is determined by n (the number of trials) and p (the probability of a success). When p is approximately 0.5, and as n increases, the shape of the binomial distribution becomes similar to that of a normal distribution. The larger n is and the closer p is to 0.5, the more similar the shape of the binomial distribution is to that of a normal distribution.

But when p is close to 0 or 1 and n is relatively small, a normal approximation is inaccurate.

As a rule of thumb, statisticians generally agree that a normal approximation should be used only when $n\,p$ and $n\,q$ are both greater than or equal to 5. (*Note:*q = 1- p.)

For example, if p is 0.3 and n is 10, then np (10)(0.3) = 3, and a normal distribution should not be used as an approximation. On the other hand, if p =0.5 and n =10, then np (10)(0.5) =5 and nq (10)(0.5) =5, and a normal distribution can be used as an approximation

In addition to the previous condition of $np \geq 5$ and $nq \geq 5$, a correction for continuity may be used in the normal approximation.

A ***CORRECTION FOR CONTINUITY*** is a correction employed when a continuous distribution is used to approximate a discrete distribution. The continuity correction means that for any specific value of X, say 8, the boundaries of X in the binomial distribution. Hence, when you employ a normal distribution to approximate the binomial, you must use the boundaries of any specific value X as they are shown in the binomial distribution.

For example, for $P(X= 8)$, the correction is $P(7.5 < X < 8.5)$. For $P(X \leq 7)$, the correction is $P(X < 7.5)$. For $P(X \geq 3)$, the correction is $P(X > 2.5)$.

The formulas for the mean and standard deviation for the binomial distribution are necessary for calculations. They are $\mu = np \quad \sigma = \sqrt{npq}$

The steps for using the normal distribution to approximate the binomial distribution are shown:

1. Check to see whether the normal approximation can be used.
2. Find the mean μ and the standard deviation σ.
3. Write the problem in probability notation, using X.
4. Rewrite the problem by using the continuity correction factor, and show the corresponding area under the normal distribution.
5. Find the corresponding z values.

6. Find the solution

Example

A magazine reported that 6% of cancer patients smoked while in school. If 300 cancer patients are selected at random, find the probability that exactly 25 smoked while in school.

Solution

Here $p = 0.06, \ \ \mathrm{q} = 0.94, \ \ \mathrm{n} = 300$

1. Check to see whether a normal approximation can be used

   $\mathrm{np} = (300)(0.06) = 18, \qquad \mathrm{nq} = (300)(0.94) = 282$

   $\text{since np} \geq 5 \text{ and nq} \geq 5, \text{the normal distribution can be used}$

2. Find the mean and standard deviation

   $\mu = np = (300)(0.06) = 18 \qquad \sigma = \sqrt{npq} = \sqrt{(300)(0.06)(0.94)} = 4.11$

3. Write the problem in probability notation: $P(X = 25)$.

4. Rewrite the problem by using the continuity correction factor.

   $P(25 - 0.5 < X < 25 + 0.5)$ $P(24.5 < X < 25.5)$

5. Find the corresponding *z* values. Since 25 represents any value between 24.5 and 25.5, find both *z* values.

   $\mathrm{z}_1 = \frac{25.5 - 18}{4.11} = 1.82 \qquad \mathrm{z}_2 = \frac{24.5 - 18}{4.11} = 1.58$

6. The area to the left of *z*=1.82 is 0.9656, and the area to the left of *z*=1.58 is 0.9429. The area between the two *z* values is 0.9656 0.9429 0.0227, or 2.27%. Hence, the probability that exactly 25 patients smoked while in school is 2.27%

Example

Of the members of a Medical Association, 10% are widowed. If 200 Medical Association members are selected at random, find the probability that 10 or more will be widowed.

Solution

1. Check to see whether a normal approximation can be used

   $\mathrm{n} = 200, \mathrm{p} = 0.10, \quad \mathrm{q} = 0.90$
   $\mathrm{np} = (200)(0.10) = 20, \qquad \mathrm{nq} = (200)(0.90) = 180$
   since $\mathrm{np} \geq 5$ and $\mathrm{nq} \geq 5$, the normal distribution can be used

2. $\mu = np = (200)(0.10) = 20 \qquad \sigma = \sqrt{npq} = \sqrt{(200)(0.10)(0.90)} = 4.24$
3. P(X $\geq$ 10)
4. $P(X > 10 - 0.5) = P(X > 9.5)$
5. $z = \dfrac{9.5 - 20}{4.24} = -2.48$
6. The area to the left of z =-2.48 is 1-0.0066= 99.34

   It can be concluded, then, that the probability of 10 or more widowed people in a random sample of 200 Medical Association members is 99.34%

Trial

If a baseball player's batting average is 0.320 (32%), find the probability that the player will get at most 26 hits in 100 times at bat.

SESSION 2.2: Normal Approximation to Poisson Distribution

The normal distribution can also be used to approximate the Poisson distribution whenever the parameter λ, the expected number of successes, equals or exceeds 5. Since the value of the mean and the variance of a Poisson distribution are the same,

$$\mu = \sigma^2 = \lambda$$

Then the standard deviation is

$$\sigma = \sqrt{\lambda}$$

Substituting into the transformation Equation

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

so that, for large enough λ, the random variable Z is approximately normally distributed. Hence, to find approximate probabilities corresponding to the values of the Poisson random variable X the Equation below is used

$$Z \equiv \frac{X_a - \lambda}{\sqrt{\lambda}}$$

where λ = expected number of successes or mean of the Poisson distribution σ = $\sqrt{\lambda}$, the standard deviation of the Poisson distribution; X_a = adjusted number of successes, x, for the discrete random variable X, such that X_a = X- 0.5 or X_a = X + 0.5 as appropriate (apply continuity correction)

Example

In a lab there are 45 accidents per year and the number of accidents per year follows a Poisson distribution. Use the normal approximation to find the probability that there are more than 50 accidents in a year.

Solution

Because $\lambda > 20$ a normal approximation can be used.

Let X be the random variable of the number of accidents per year.

To find $P(X > 50)$ apply a continuity correction and find $P(X > 50.5)$

For the normal approximation $\mu = \lambda = 45$ and $\sigma = \sqrt{\lambda} = 6.71$ (to 3 s. f.)

$$P(X > 50.5) = P(Z \geq \frac{50.5 - 45}{6.71})$$

$$= P(Z \geq 0.820)$$

$$=0.2061$$

The probability that there are more than 50 accidents in a year is 0.2061

UNIT 3

SAMPLING METHODS

SESSION 3.1: Introduction and Definition of Terms

Sampling involves the selection of a number of a study units from a defined population. The population is too large for us to consider collecting information from all its members. If the whole population is taken there is no need of statistical inference. Usually, a representative subgroup of the population (sample) is included in the investigation. A representative sample has all the important characteristics of the population from which it is drawn

Advantages of samples

1. cost - sampling saves time, labour and money
2. quality of data - more time and effort can be spent on getting reliable data on each individual included in the sample.- Due to the use of better trained personnel, more careful supervision and processing a sample can actually produce precise results

If we have to draw a sample, we will be confronted with the following questions:

- ❖ What is the group of people (population) from which we want to draw a sample?
- ❖ How many people do we need in our sample?
- ❖ How will these people be selected?

Apart from persons, a population may consist of mosquitoes, villages, institutions, etc.

Common terms used in sampling

1. ***Reference population (also called source population or target population) -*** the population of interest, to which the investigators would like to generalize the results of the study, and from which a representative sample is to be drawn.

2. **Study or sample population** - the population included in the sample.
3. **Sampling unit** - the unit of selection in the sampling process
4. **Study unit** - the unit on which information is collected.
   (i) the sampling unit is not necessarily the same as the study unit.
   (ii) if the objective is to determine the availability of latrine, then the study unit would be the household; if the objective is to determine the prevalence of trachoma, then the study unit would be the individual.
5. **Sampling frame** - the list of all the units in the reference population, from which a sample is to be picked.
6. **Sampling fraction (Sampling interval)** - the ratio of the number of units in the sample to the number of units in the reference population **(n/N)**

SESSION 3.2: Sampling Methods (Two broad divisions)

A. Non-probability Sampling Methods

1. Used when a sampling frame does not exist
2. No random selection (unrepresentative of the given population)
3. Inappropriate if the aim is to measure variables and generalize findings obtained from a sample to the population.

Two commonly used non-probability sampling methods are:

1. **Convenience sampling**: is a method in which for convenience sake the study units that happen to be available at the time of data collection are selected.
2. **Quota sampling**: is a method that ensures that a certain number of sample units from different categories with specific characteristics are represented. In this method the investigator interviews as many people in each category of study unit as he can find until he has filled his quota.

Both the above methods do not claim to be representative of the entire population.

B. Probability Sampling Methods

- A sampling frame exists or can be compiled.
- Involve random selection procedures. All units of the population should have an equal or at least a known chance of being included in the sample.
- Generalization is possible (from sample to population)

Types of Probability Sampling methods

1. Simple random sampling (SRS)

- This is the most basic scheme of random sampling.
- Each unit in the sampling frame has an equal chance of being selected
- representativeness of the sample is ensured.

However, it is costly to conduct SRS. Moreover, minority subgroups of interest in the population my not be present in the sample in sufficient numbers for study.

To select a simple random sample you need to:

- Make a numbered list of all the units in the population from which you want to draw a sample.
- Each unit on the list should be numbered in sequence from 1 to N (where N is the size of the population)

- Decide on the size of the sample Select the required number of study units, using a “lottery” method or a table of random numbers.

Methods of SRS

"**Lottery” method**: for a small population it may be possible to use the “lottery” method: each unit in the population is represented by a slip of paper, these are put in a box and mixed, and a sample of the required

size is drawn from the box.

Table of random numbers: if there are many units, however, the above technique soon becomes laborious. Selection of the units is greatly facilitated and made more accurate by using a set of random numbers in which a large number of digits is set out in random order. The property of a table of random numbers is that, whichever way it is read, vertically in columns or horizontally in rows, the order of the digits is random. Nowadays, any scientific calculator has the same facilities.

2. Systematic Sampling

Individuals are chosen at regular intervals (for example, every kth) from the sampling frame. The first unit to be selected is taken at random from among the first k units. For example, a systematic sample is to be selected from 1200 students of a school. The sample size is decided to be 100. The sampling fraction is: 100 /1200 = 1/12. Hence, the sample interval is 12. The number of the first student to be included in the sample is chosen randomly, for example by blindly picking one out of twelve pieces of paper, numbered 1 to 12. If number 6 is picked, every twelfth student will be included in the sample, starting with student number 6, until 100 students are selected. The numbers selected would be 6,18,30,42,etc.

Merits

• Systematic sampling is usually less time consuming and easier to perform than simple random sampling. It provides a good approximation to SRS.

• Unlike SRS, systematic sampling can be conducted without a sampling frame (useful in some situations where a sampling frame is not readily available). Eg., In patients attending a health center, where it is not possible to predict in advance who will be attending.

Demerits

• If there is any sort of cyclic pattern in the ordering of the subjects which coincides with the sampling interval, the sample will not be representative of the population.

Examples

- ❖ List of married couples arranged with men's names alternatively with the women's names (every 2nd, 4th , etc.) will result in a sample of all men or women).
- ❖ If we want to select a random sample of a certain day (sampling fraction on which to count clinic attendance, this day may fall on the same day of the week, which might, for example be a market day.

3. Stratified Sampling

It is appropriate when the distribution of the characteristic to be studied is strongly affected by certain variable (heterogeneous population). The population is first divided into groups (strata) according to a characteristic of interest (eg., sex, geographic area, prevalence of disease, etc.). A separate sample is then taken independently from each stratum, by simple random or systematic sampling.

• **proportional allocation** - if the same sampling fraction is used for each stratum.

• **non- proportional allocation** - if a different sampling fraction is used for each stratum or if the strata are unequal in size and a fixed number of units is selected from each stratum.

Merit

- The representativeness of the sample is improved. That is, adequate representation of minority subgroups of interest can be ensured by stratification and by varying the sampling fraction between strata as required.

Demerit

- Sampling frame for the entire population has to be prepared separately for each stratum.

4. Cluster sampling

In this sampling scheme, selection of the required sample is done on groups of study units (clusters) instead of each study unit individually. The sampling unit is a cluster, and the sampling frame is a list of these clusters.

procedure

- The reference population (homogeneous) is divided into clusters. These clusters are often geographic units (eg districts, villages, etc.)

- A sample of such clusters is selected

- All the units in the selected clusters are studied

It is preferable to select a large number of small clusters rather than a small number of large clusters.

Merit

A list of all the individual study units in the reference population is not required. It is sufficient to have a list of clusters.

Demerit

It is based on the assumption that the characteristic to be studied is uniformly distributed throughout the reference population, which may not always be the case. Hence, sampling error is usually higher than for a simple random sample of the same size.

5. Multi-stage sampling

This method is appropriate when the reference population is large and widely scattered . Selection is done in stages until the final sampling unit (eg., households or persons) are arrived at. The primary sampling unit (PSU) is the sampling unit (usually large size) in the first sampling stage. The secondary sampling unit (SSU) is the sampling unit in the second sampling stage, etc.

Merit - Cuts the cost of preparing sampling frame

Demerit - Sampling error is increased compared with a simple random sample. Multistage sampling gives less precise estimates than sample random sampling for the same sample size, but the reduction in cost usually far outweighs this, and allows for a larger sample size.

SESSION 3.3: ERRORS IN SAMPLING

When we take a sample, our results will not exactly equal the correct results for the whole population. That is, our results will be subject to errors.

1. Sampling error (random error)

A sample is a subset of a population. Because of this property of samples, results obtained from them cannot reflect the full range of variation found in the larger group (population). This type of error, arising from the sampling process itself, is called sampling error, which is a form of random error. Sampling error can be minimized by increasing the size of the sample. When $n = N \Rightarrow$ sampling error = 0

2. Non-sampling error (bias)

It is a type of systematic error in the design or conduct of a sampling procedure which results in distortion of the sample, so that it is no longer representative of the reference population. We can eliminate or reduce the non-sampling error (bias) by careful design of the sampling procedure and not by increasing the sample size.

Example: If you take male students only from a student dormitory in Ethiopia in order to determine the proportion of smokers, you would result in an overestimate, since females are less likely to smoke. Increasing the number of male students would not remove the bias.

- There are several possible sources of bias in sampling (eg., accessibility bias, volunteer bias, etc.)

- The best known source of bias is non response. It is the failure to obtain information on some of the subjects included in the sample to be studied.

- Non response results in significant bias when the following two conditions are both fulfilled.

  - ❖ When non-respondents constitute a significant proportion of the sample (about 15% or more)
  - ❖ When non-respondents differ significantly from respondents.

- There are several ways to deal with this problem and reduce the possibility of bias:

a) Data collection tools (questionnaire) have to be pre-tested.

b) If non response is due to absence of the subjects, repeated attempts should be considered to contact study subjects who were absent at the time of the initial visit.

c) To include additional people in the sample, so that non respondents who were absent during data

collection can be replaced (make sure that their absence is not related to the topic being studied).

NB: The number of non-responses should be documented according to type, so as to facilitate an assessment of the extent of bias introduced by non-response.

UNIT 4

REGRESSION AND CORRELATION

HOW CAN WE EXPLORE THE ASSOCIATION BETWEEN TWO QUANTITATIVE VARIABLES?

An association exists between two variables if a particular value of one variable is more likely to occur with certain values of the other variable.

For higher levels of energy use, does the CO_2 level in the atmosphere tend to be higher? If so, then there is an association between energy use and CO_2 level.

Positive Association: As x goes up, y tends to go up.

Negative Association: As x goes up, y tends to go down.

SESSION 4.1: CORRELATION

The Pearson correlation coefficient, r, measures the strength and the direction of a straight-line relationship.

•The **strength** of the relationship is determined by the closeness of the points to a straight line.

•The ***direction*** is determined by whether one variable generally increases or generally decreases when the other variable increases.

•***r*** is always between –1 and +1

•**magnitude** indicates the strength

•***r* = –1 or +1** indicates a perfect linear relationship

•**sign** indicates the direction

- ***r* = 0** indicates no linear relationship

The following data were collected to study the relationship between the body weight, y and height, x, of new born babies at the University Hospital.

Babies	X	y	x^2	y^2	xy
1	2	2	4	4	4
2	3	5	9	25	15
3	4	7	16	49	28
4	5	10	25	100	50
5	6	11	36	121	66
	20	35	90	299	163
	$\nwarrow$	$\nwarrow$	$\nwarrow$	$\nwarrow$	$\nwarrow$
	$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum xy$

Pearson correlation coefficient, r.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{5(163) - (20)(35)}{\sqrt{5(90) - (20)^2} \times \sqrt{5(299) - (35)^2}} = 0.9897$$

Example

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table. It is required to find the correlation between age and weight

Age	Weight
7	12
6	8
8	12
5	10
6	11
9	13

Solution

Age(x)	Weight (y)	x^2	y^2	xy
7	12	49	144	84
6	8	36	64	48
8	12	64	144	96
5	10	25	100	50
6	11	36	121	66
9	13	81	169	117
$\sum x = 41$	$\sum y = 66$	$\sum x^2 = 291$	$\sum y^2 = 742$	$\sum xy = 461$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$r = \frac{6(461) - (41)(66)}{\sqrt{6(291) - (41)^2} \times \sqrt{6(742) - (66)^2}} = 0.75955$$

r =0.76 strong positive correlation between age and weight

Example

Relationship between Anxiety and Test Scores

Anxiety (x)	Test Scores (y)	x^2	y^2	xy
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum x = 32$	$\sum y = 32$	$\sum x^2 = 230$	$\sum y^2 = 204$	$\sum xy = 129$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$r = \frac{6(129) - (32)(32)}{\sqrt{6(230) - (32)^2} \times \sqrt{6(204) - (32)^2}} = -0.9369$$

r = -0.94 indirect(negative) strong correlation between anxiety and test scores

SESSION 4.2: REGRESSION

When the relationship has a straight-line pattern, the Pearson correlation coefficient describes it numerically. We can analyze the data further by finding an equation for the straight line that best describes the pattern. This equation predicts the value of the response(y) variable from the value of the explanatory variable(x).

Much of mathematics is devoted to studying variables that are deterministically related. Saying that x and y are related in this manner means that once we are told the value of x, the value of y is completely specified.

For example

Suppose the cost of ambulance service at the University Hospital is \$10 plus \$0.75 per number of number of nurses in the car. If we let x= # number of nurses and y = price of ambulance service, then y=10+.75x. If we order three nurses to accompany the ambiance, then y=10+.75(3)=12.25

There are many variables x and y that would appear to be related to one another, but not in a deterministic fashion.

Suppose we examine the relationship between x=sugar level and Y=blood pressure. The value of y cannot be determined just from knowledge of x, and two different patients could have the same x value but have very different y values. Yet there is a tendency for those patients who have high (low) high sugar level also to have high(low) blood pressure. Knowledge of a patient's sugar level should be quite helpful in enabling us to predict how that patient's blood pressure will be.

Regression analysis is the part of statistics that deals with investigation of the relationship between two or more variables related in a nondeterministic fashion.

Estimation of Regression Parameters

The least-squares line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Equation for Least Squares (Regression) Line

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$$

$\hat{\beta}_1$ denotes the slope. The slope in the equation equals the amount that $\hat{y}$ changes when x increases by one unit.

$$\hat{\beta}_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\hat{y} = \overline{y} + \hat{\beta}_1(x - \overline{x})$$

$\hat{\beta}_0$ denotes the y-intercept. The y-intercept is the predicted value of y when x=0.

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

When talking about regression equations, the following are terms used for x and y

x: predictor variable, explanatory variable, or independent variable

y: response variable or dependent variable

Example

The following data were collected at the University Hospital of new born babies; their and height (x), and weight(y) are given in the table below.

1. Find the regression equation
2. What is the slope and the intercept of the regression equation?
3. What is the predicted weight when a baby's height is 5.2?

Height (x)	Weight (y)
2	2
3	5
4	7
5	10
6	11

Solution

Height (x)	Weight (y)	x^2	y^2	xy
2	2	4	4	4
3	5	9	25	15
4	7	16	49	28
5	10	25	100	50
6	11	36	121	66
$\sum x = 20$	$\sum y = 35$	$\sum x^2 = 90$	$\sum y^2 = 299$	$\sum xy = 163$

$\bar{x} = \frac{20}{5} = 4$ $\bar{y} = \frac{35}{5} = 7$

$$\hat{\beta}_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(163) - (20)(35)}{5(90) - (20)^2} = 2.3$$

1. $\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x}) = 7 + 2.3(x - 4)$

$$= 7 + 2.3x - 9.2$$
$$= -2.2 + 2.3x$$

The regression equation is $\hat{y}$ = -2.2 + 2.3x

2. From the equation, the slope is 2.3 and the intercept is -2.2

3. The predicted weight when height is 5.2 ; $\hat{y}$ = -2.2 + 2.3(5.2)=9.76

Example

A paediatric at St. Patrick's Hospital selected a sample of 6 patients, the value of their age (x) and systolic blood pressure(y) is demonstrated in the table below. Find

1. The regression equation
2. **2.** The predicted systolic blood pressure when age is 8.5

Age (x)	Systolic blood pressure (y)
7	12
6	8
8	12
5	10
6	11
9	13

Solution

Age (x)	Blood Pressure (y)	x^2	y^2	xy
7	12	49	144	84
6	8	36	64	48
8	12	64	144	96
5	10	25	100	50
6	11	36	121	66
9	13	81	169	117
$\sum x = 41$	$\sum y = 66$	$\sum x^2 = 291$	$\sum y^2 = 742$	$\sum xy = 461$

$\bar{x} = \frac{41}{6} = 6.83$ $\bar{y} = \frac{66}{6} = 11$

$$\hat{\beta}_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{6(461) - (41)(66)}{6(291) - (41)^2} = 0.92$$

1. $\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x}) = 11 + 0.92(x - 6.83)$

$= 11 + 0.92x - 6.28$

$= 4.72 + 0.92x$

2. The predicted systolic blood pressure when age is 8.5 ; $\hat{y}$ = 4.72 + 0.92(8.5)=12.54

UNIT 5

ANALYSIS OF CATEGORICAL DATA

SESSION 5.1: TEST FOR GOODNESS OF FIT

In addition to being used to test a single variance, the chi-square statistic can be used to see whether a frequency distribution fits a specific pattern.

For example

A traffic engineer may wish to see whether accidents occur more often on some days than on others, so that she can increase police patrols accordingly.

An emergency service may want to see whether it receives more calls at certain times of the day than at others, so that it can provide adequate staffing.

When you are testing to see whether a frequency distribution fits a specific pattern, you can use the chi-square goodness-of-fit test.

Example

Suppose an insurance company wished to see whether patients have any preference among five hospitals in Kumasi. A sample of 100 patients provided these data:

KATH	University Hospital	Emina Hospital	County Hospital	Manhyia Hospital
32	28	16	14	10

If there were no preference, you would expect each Hospital to be selected with equal frequency. In this case, the equal frequency is 100/5 = 20. That is, *approximately* 20 people would select each Hospital.

Since the frequencies for each Hospital were obtained from a sample, these actual frequencies are called the observed frequencies. The frequencies obtained by calculation (as if there were no preference) are called the expected frequencies. A completed table for the test is shown.

frequency	KATH	University Hospital	Emina Hospital	County Hospital	Manhyia Hospital
Observed	32	28	16	14	10
Expected	20	20	20	20	20

The observed frequencies will almost always differ from the expected frequencies due to sampling error; that is, the values differ from sample to sample. But the question is: Are these differences significant (a preference exists), or are they due to chance? The chi-square goodness-of-fit test will enable the researcher to determine the answer. Before computing the test value, you must state the hypotheses.

The null hypothesis should be a statement indicating that there is no difference or no change. For this example, the hypotheses are as follows:

H_0 : Patients have no prefrence for Hospitals

H_1 : Patients have prefrence for Hospitals

In the goodness-of-fit test, the degrees of freedom are equal to the number of categories minus 1. For this example, there are five categories (KATH, University Hospital , Emina Hospital, County Hospital and Manhyia Hospital) hence, the degrees of freedom are $5-1=4$.This is so because the number of subjects in each of the first four categories is free to vary. But in order for the sum to be 100—the total number of subjects—the number of subjects in the last category is fixed.

Formula for the chi square goodness –of- fit- test

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

With degree of freedom equal to the number of categories minus 1.

$O = observed\,\text{frequency}$
$\text{E} = \text{Expected frequency}$

Two assumptions are needed for the goodness-of-fit test. These assumptions are given next

The data are obtained from a random sample.

The expected frequency for each category must be 5 or more.

This test is a right-tailed test, since when the $O-E$ values are squared, the answer will be positive or zero.

Example on Hospital preference

Is there enough evidence to reject the claim that there is no preference in the selection of Hospitals, using the data shown previously? Let $\alpha = 0.05$

Solution

State the hypotheses and identify the claim.

H_0 : Patients have no prefrence for Hospitals

H_1 : Patients have prefrence for Hospitals

Find the critical value. The degrees of freedom are $5-1=4$, and $\alpha=0.05$ Hence, the critical value from Table is 9.488

Compute the test value by subtracting the expected value from the corresponding observed value, squaring the result and dividing by the expected value, and finding the sum. The expected value for each category is 20, as shown previously

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(10-20)^2}{20} = 18.0$$

Make the decision. The decision is to reject the null hypothesis, since 18.0 > 9.488

Summarize the results. There is enough evidence to reject the claim that patients show no preference for hospitals.

The steps for the chi-square goodness-of-fit test are summarized in this Procedure

State the hypotheses and identify the claim.

Find the critical value. The test is always right-tailed

Compute the test value

Find the sum of the value $\chi^2 = \sum \frac{(O-E)^2}{E}$

Make the decision

Summarize the results

When there is perfect agreement between the observed and the expected values, $\chi^2 = 0$. Also, χ^2 can never be negative. Finally, the test is right-tailed because

“ H_0 : Good fit” and “ H_1 : Not a good fit” mean that χ^2 will be small in the first case and large in the second case

Example

A researcher read that firearm-related deaths for people aged 1 to 18 were distributed as follows: 74% were accidental, 16% were homicides, and 10% were suicides. In her district, there were 68 accidental deaths, 27 homicides, and 5 suicides during the past year. At a 0.10, test the claim that the percentages are equal.

Solution

State the hypotheses and identify the claim:

H_0: The deaths due to firearms for people aged 1 through 18 are distributed as follows: 74% accidental, 16% homicides, and 10% suicides (claim).

H_1: The distribution is not the same as stated in the null hypothesis.

Find the critical value. Since $\alpha = 0.05$and the degrees of freedom are$3-1=2$, the critical value is 4.605

Compute the test value. The expected values are as follows:

$0.74\times100=74$

$0.16\times100=16$

$0.10\times100=10$

$$\chi^2 = \sum\frac{(O-E)^2}{E} = \frac{(68-74)^2}{74} + \frac{(27-16)^2}{16} + \frac{(5-10)^2}{10} = 10.549$$

Reject the null hypothesis, since 10.549 > 4.605

Summarize the results. There is enough evidence to reject the claim that the distribution is 74% accidental, 16% homicides, and 10% suicides.

SESSION 5.2: TEST FOR INDEPENDENCE

The chi-square independence test can be used to test the independence of two variables. For example, suppose a new postoperative procedure is administered to a number of patients in a large hospital. The researcher can ask the question, Do the doctors feel differently about this procedure from the nurses, or do they feel basically the same way? Note that the question is not whether they prefer the procedure but whether there is a difference of opinion between the two groups. To answer this question, a researcher selects a sample of nurses and doctors and tabulates the data in table form, as shown

Group	prefer new procedure	prefer old procedure	no preference
Nurses	100	80	20
Doctors	50	120	30

As the survey indicates, 100 nurses prefer the new procedure, 80 prefer the old procedure, and 20 have no preference; 50 doctors prefer the new procedure, 120 like the old procedure, and 30 have no preference. Since the main question is whether there is a difference in opinion, the null hypothesis is stated as follows:

H_0 : The opinion about the procedure is independent of the profession

The alternative hypothesis is stated as follows:

H_1 : The opinion about the procedure is dependent on the profession

If the null hypothesis is not rejected, the test means that both professions feel basically the same way about the procedure and the differences are due to chance. If the null hypothesis is rejected, the test means that one group feels differently about the procedure from the other. Remember that rejection does *not* mean that one group favors the procedure and the other does not. Perhaps both groups favor it or both dislike it, but in different proportions. To test the null hypothesis by using the chi-square independence test, you must compute the expected frequencies, assuming that the null hypothesis is true. These frequencies are computed by using the observed frequencies given in the table.

When data are arranged in table form for the chi-square independence test, the table is called a contingency table. The table is made up of *R* rows and *C* columns. The table here has two rows and three columns

Group	prefer new procedure	prefer old procedure	no preference
Nurses	100	80	20
Doctors	50	120	30

Note that row and column headings do not count in determining the number of rows and columns. A contingency table is designated as an $R \times C$ (rows by columns) table. In this case, R=2 and C=3; hence, this table is a 2×3 contingency table. Each block in the table is called a *cell* and is designated by its row and column position. For example, the cell with a frequency of 80 is designated as $C_{1,2}$, or row 1, column 2. The cells are shown below.

	Column 1	Column 2	Column 3
Row 1	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$
Row 2	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$

The degrees of freedom for any contingency table are (rows -1) times (columns -1); that is, d.f. $(R-1)(C-1)$. In this case, $(2-1)(3-1) = (1)(2) = 2$.

The reason for this formula for *d.f.* is that all the expected values except one are free to vary in each row and in each column.

Using the previous table, you can compute the expected frequencies for each block (or cell), as shown next

Find the sum of each row and each column, and find the grand total, as shown

Group	prefer new procedure	prefer old procedure	no preference	Total
Nurses	100	80	20	200
Doctors	50	120	30	200
Total	150	200	50	400

For each cell, multiply the corresponding row sum by the column sum and divide by the grand total, to get the expected value:

$$\text{Expected value} = \frac{\text{row sum} \times \text{column sum}}{\mathit{grand}\ \text{total}}$$

For each cell, the expected values are computed as follows:

$$E_{1,1} = \frac{200 \times 150}{400} = 75 \qquad E_{1,2} = \frac{200 \times 200}{400} = 100 \qquad E_{1,3} = \frac{200 \times 50}{400} = 25$$

$$E_{2,1} = \frac{200 \times 150}{400} = 75 \qquad E_{2,2} = \frac{200 \times 200}{400} = 100 \qquad E_{2,3} = \frac{200 \times 50}{400} = 25$$

The expected values can now be placed in the corresponding cells along with the observed values, as shown.

Group	prefer new procedure	prefer old procedure	no preference	Total

Nurses	100 (75)	80(100)	20(25)	200
Doctors	50 (75)	120(100)	30(25)	200
Total	150	200	50	400

The rationale for the computation of the expected frequencies for a contingency table uses proportions. For $C_{1,1}$ a total of 150 out of 400 people prefer the new procedure. And since there are 200 nurses, you would expect, if the null hypothesis were true, (150/400)(200), or 75, of the nurses to be in favor of the new procedure. The formula for the test value for the independence test is the same as the one used for the goodness-of-fit test. It is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

For the previous example, compute the $(O-E)^2/E$ values for each cell, and then find the sum.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(100-75)^2}{75} + \frac{(80-100)^2}{100} + \frac{(20-25)^2}{25} + \frac{(50-75)^2}{75} + \frac{(120-100)^2}{100} + \frac{(30-25)^2}{25} = 26.67$$

The final steps are to make the decision and summarize the results. This test is always a right-tailed test, and the degrees of freedom are $(R-1)(C-1)$= $(2-1)(3-1)$ = (1)(2) = 2. If $\alpha = 0.05$, the critical value from Table is 5.991. Hence, the decision is to reject the null hypothesis, since 26.67 > 5.991

The conclusion is that there is enough evidence to support the claim that opinion is related to (dependent on) profession—that is, that the doctors and nurses differ in their opinions about the procedure

Example

A researcher wishes to determine whether there is a relationship between the gender of an individual and the amount of alcohol consumed. A sample of 68 people is selected, and the following data are obtained.

Alcohol consumption				
Gender	Low	Moderate	High	Total
Male	10	9	8	27
Female	13	16	12	41
Total	23	25	20	68

At $\alpha = 0.10$, can the researcher conclude that alcohol consumption is related to gender?

Solution

State the hypotheses and identify the claim

H_0 : The amount of alcohol that a person consumes is independent of the individual's gender

H_1 : The amount of alcohol that a person consumes is dependent on the individual's gender (claim)

Find the critical value. The critical value is 4.605, since the degrees of freedom are $(2-1)(3-1)=2$.

Compute the test value. First, compute the expected values.

$$E_{1,1}=\frac{27\times 23}{68}=9.13 \qquad E_{1,2}=\frac{27\times 25}{68}=9.93 \qquad E_{1,3}=\frac{27\times 20}{68}=7.94$$

$$E_{2,1} = \frac{41 \times 23}{68} = 13.87 \qquad E_{2,2} = \frac{41 \times 25}{68} = 15.07 \qquad E_{2,3} = \frac{41 \times 20}{68} = 12.06$$

	Alcohol consumption			
Gender	Low	Moderate	High	Total
Male	10 (9.13)	9(9.93)	8(7.94)	27
Female	13(13.87)	16(15.07)	12(12.06)	41
Total	23	25	20	68

Then the test value is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(10-9.13)^2}{9.13} + \frac{(9-9.93)^2}{9.93} + \frac{(20-25)^2}{25} + \frac{(8-7.94)^2}{7.94} + \frac{(13-13.87)^2}{13.87} + \frac{(16-15.07)^2}{15.07}$$
$$+ \frac{(12-12.06)^2}{12.06} = 0.283$$

Make the decision. The decision is to not reject the null hypothesis, since 0.283 < 4.605.

Summarize the results. There is not enough evidence to support the claim that the amount of alcohol a person consumes is dependent on the individual's gender.

SESSION 5.3: TEST FOR HOMOGENEITY OF PROPORTIONS

The second chi-square test that uses a contingency table is called the homogeneity of proportions test. In this situation, samples are selected from several different populations, and the researcher is interested in determining whether the proportions of elements that have a common characteristic are the same for each population. The sample sizes are specified in advance, making either the row totals or column totals in the contingency table known before the samples are selected.

For example, a researcher may select a sample of 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors and then find the proportion of students who are smokers in each level. The researcher will then compare the proportions for each group to see if they are equal. The hypotheses in this case would be

$H_0 : p_1 = p_2 = p_3 = p_4$

H_1 : At least one proportion is different from the others.

If the researcher does not reject the null hypothesis, it can be assumed that the proportions are equal and the differences in them are due to chance. Hence, the proportion of

students who smoke is the same for grade levels freshmen through senior. When the null hypothesis is rejected, it can be assumed that the proportions are not all equal. The computational procedure is the same as that for the test of independence shown in

Example

A researcher selected 100 patients from each of 3 Hospitals and asked them if the Hospital had lost their folder on their last visit. The data are shown in the table. At $\alpha = 0.05$, test the claim that the proportion of patients from each Hospital who lost folder on their last visit is the same for each Hospital.

	Hospital A	Hospital B	Hospital C	Total
Yes	10	7	4	21
No	90	93	96	279
Total	100	100	100	300

Solution

State the hypothesis

$H_0 : \mathrm{p}_1 = p_2 = p_3$

H_1 : At least one mean differs from the other.

Find the critical value. The formula for the degrees of freedom is the same as before: (rows – 1)(columns – 1) = (2 – 1)(3 – 1) = 1(2) = 2. The critical value is 5.991.

Compute the test value. First compute the expected values.

$\mathrm{E}_{1,1} = \frac{21\times100}{300} = 7$ $\mathrm{E}_{1,2} = \frac{21\times100}{300} = 7$ $\mathrm{E}_{1,3} = \frac{21\times100}{300} = 7$

$\mathrm{E}_{2,1} = \frac{279\times100}{300} = 93$ $\mathrm{E}_{2,2} = \frac{279\times100}{300} = 93$ $\mathrm{E}_{2,3} = \frac{279\times100}{300} = 93$

	Hospital A	Hospital B	Hospital C	Total
Yes	10 (7)	7(7)	4 (7)	21
No	90 (93)	93 (93)	96(93)	279
Total	100	100	100	300

$$\chi^2 = \sum\frac{(O-E)^2}{E} = \frac{(10-7)^2}{7} + \frac{(7-7)^2}{7} + \frac{(4-7)^2}{7} + \frac{(90-93)^2}{93} + \frac{(93-93)^2}{93} + \frac{(96-93)^2}{93}$$
$$= 2.765$$

Make the decision. Do not reject the null hypothesis since 2.765 < 5.991

Summarize the results. There is not enough evidence to reject the claim that the proportions are equal. Hence it seems that there is no difference in the proportions of the folder lost by each Hospital.

The steps for the chi-square independence and homogeneity tests are summarized

State the hypotheses and identify the claim

Find the critical value in the right tail. Using Table

Compute the test value. To compute the test value, first find the expected values. For each cell of the contingency table, use the formula

$$\mathrm{E} = \frac{(row\,\mathrm{sum})(\mathrm{column\ sum})}{\mathrm{grand\ total}}$$

to get the expected value. To find the test value, use the formula

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Make the decision.

Summarize the results

Assumptions for the chi-square independence and Homogeneity Tests

The data are obtained from a random sample

The expected value in each cell must be 5 or more.