Dr. Suresh Kumar, Department of Mathematics, BITS-Pilani, Pilani Campus

Note: *Some concepts of Probability & Statistics are briefly described here just to help the students. Therefore, the following study material is expected to be useful but not exhaustive for the Probability & Statistics course. For detailed study, the students are advised to attend the lecture/tutorial classes regularly, and consult the text book prescribed in the hand out of the course.*

Chapter 5

So far we have studied a single random variable either discrete or continuous. Such random variables are called univariate. Problems do arise where we need to study two random variables simultaneously. For example, we may wish to study the heights and weights of a group of students up to the age of 20 years. Typical questions to ask are, "What is the average height of students of age less than or equal to 18 years?" or, "Is the height independent of weight?". To answer this type of questions, we need to study what are called two-dimensional or bivariate random variables.

Discrete Bivariate Random Variable

Let X and Y be two discrete random variables. Then the ordered pair (X, Y) is called a two dimensional or bivariate discrete random variable.

Joint density function

A function f_{XY} such that

$$f_{XY}(x,y) \geq 0,\ f_{XY}(x,y) = P[X = x, Y = y],\ \sum_{X=x}\sum_{Y=y} f_{XY}(x,y) = 1$$

is called joint density function of (X, Y).

Distribution function

The distribution function of (X, Y) is given by

$$F(x,y) = \sum_{X\leq x}\sum_{Y\leq y} f_{XY}(x,y).$$

Marginal density functions

The marginal density of X, denoted by f_X, is defined as

$$f_X(x) = \sum_{Y=y} f_{XY}(x,y).$$

Similarly, the marginal density of Y, denoted by f_Y, is defined as

$$f_Y(y) = \sum_{X=x} f_{XY}(x,y).$$

Independent random variables

The discrete random variables X and Y are said to be independent if and only if

$$f_{XY}(x,y) = f_X(x)f_Y(y) \text{ for all } (x,y).$$

Expectation

The expectation or mean of X is defined as

$$E[X] = \sum_{X=x} \sum_{Y=y} x f_{XY}(x, y) = \mu_X.$$

In general, the expectation of a function of X and Y, say $H(X, Y)$, is defined as

$$E[H(X, Y)] = \sum_{X=x} \sum_{Y=y} H(x, y) f_{XY}(x, y).$$

Covariance

If μ_X and μ_Y are the means of X and Y respectively, then covariance of X and Y, denoted by $\text{Cov}(X, Y)$ is defined as

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y].$$

Ex. In an automobile plant, two tasks are performed by robots, the welding of two joints and tightening of three bolts. Let X denote the number of defective bolts and Y denote the number of improperly tightened bots produced per car. The probabilities of (X, Y) are given in the following table.

X/Y	0	1	2	3	$f_X(x)$
0	0.84	0.03	0.02	0.01	0.9
1	0.06	0.01	0.008	0.002	0.08
2	0.01	0.005	0.004	0.001	0.02
$f_Y(y)$	0.91	0.45	0.032	0.013	1

(i) Is it a density function?
(ii) Find the probability that there would be exactly one error made by the robots.
(iii) Find the probability that there would be no improperly tightened bolts.
(iv) Are the variables X and Y independent?
(v) Find $\text{Cov}(X, Y)$.

Sol. (i) We have

$$\begin{aligned}
\sum_{X=0}^{2} \sum_{Y=0}^{3} f_{XY}(x, y) &= f_{XY}(0,0) + f_{XY}(0,1) + f_{XY}(0,2) + f_{XY}(0,3) + f_{XY}(1,0) + f_{XY}(1,1) \\
&\quad + f_{XY}(1,2) + f_{XY}(1,3) + f_{XY}(2,0) + f_{XY}(2,1) + f_{XY}(2,2) + f_{XY}(2,3) \\
&= 0.84 + 0.03 + 0.02 + 0.01 + 0.06 + 0.01 + 0.008 + 0.002 + 0.01 + 0.005 + 0.004 + 0.001 \\
&= 1
\end{aligned}$$

This shows that f_{XY} is a density function.

(ii) The probability that there would be exactly one error made by the robots, is given by

$$P[X = 1, Y = 0] + P[X = 0, Y = 1] = f_{XY}(1, 0) + f_{XY}(0, 1) = 0.06 + 0.03 = 0.09.$$

(iii) The probability that there would be no improperly tightened bolts, reads as

$$P[Y = 0] = \sum_{X=0}^{2} f_{XY}(x, 0) = f_{XY}(0, 0) + f_{XY}(1, 0) + f_{XY}(2, 0) = 0.84 + 0.06 + 0.01 = 0.91.$$

It is the marginal density $f_Y(y)$ of Y at $y = 0$, that is, $f_Y(0) = 0.91$.

(iv) From the given Table, we notice that $f_{XY}(0,0) = 0.84$, $f_X(0) = 0.9$ and $f_Y(0) = 0.91$. So we have

$$f_X(0)f_Y(0) = 0.819 \neq f_{XY}(0,0).$$

This shows that X and Y are not independent.

(v) We find

$$E[X] = \sum_{X=0}^{2}\sum_{Y=0}^{3} x f_{XY}(x,y) = 0.12,$$

$$E[Y] = \sum_{X=0}^{2}\sum_{Y=0}^{3} y f_{XY}(x,y) = 0.148,$$

$$E[XY] = \sum_{X=0}^{2}\sum_{Y=0}^{3} xy f_{XY}(x,y) = 0.064.$$

Hence, $\text{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0.046$.

Continuous Bivariate Random Variable

Let X and Y be two continuous random variables. Then the ordered pair (X,Y) is called a two dimensional or bivariate continuous random variable.

Joint density function

A function f_{XY} such that

$$f_{XY}(x,y) \geq 0,\ P[a \leq X \leq b, c \leq Y \leq d] = \int_a^b \int_c^d f_{XY}(x,y)dxdy,\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1,$$

for real a, b, c and d, is called joint density function of (X,Y).

Distribution function

The distribution function of (X,Y) is given by

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(x,y)dxdy.$$

Marginal density functions

The marginal density of X, denoted by f_X, is defined as

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

Similarly, the marginal density of Y, denoted by f_Y, is defined as

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

Independent random variables

The continuous random variables X and Y are said to be independent if and only if

$$f_{XY}(x,y) = f_X(x)f_Y(y).$$

Expectation

The expectation or mean of X is defined as

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x,y)dxdy = \mu_X.$$

In general, the expectation of a function of X and Y, say $H(X,Y)$, is defined as

$$E[H(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y) f_{XY}(x,y)dxdy.$$

Covariance

If μ_X and μ_Y are the means of X and Y respectively, then covariance of X and Y, denoted by $\text{Cov}(X,Y)$ is defined as

$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - E[X]E[Y].$$

Ex. Let X denote a person's blood calcium level and Y, the blood cholesterol level. The joint density function of (X,Y) is

$$f_{XY}(x,y) = k,\ 8.5 \le x \le 10.5, 120 \le y \le 240.$$

(i) Find the value of k.
(ii) Find the marginal densities of X and Y.
(iii) Find the probability that a healthy person has a cholesterol level between 150 to 200.
(iv) Are the variables X and Y independent?
(v) Find $\text{Cov}(X,Y)$.

Sol. (i) $f_{XY}(x,y)$ being joint density function, we have

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = \int_{120}^{240}\int_{8.5}^{10.5} k dxdy = 240k.$$

So $k = 1/240$ and $f_{XY}(x,y) = 1/240$

(ii) The marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \int_{120}^{240} \frac{1}{240}dy = \frac{1}{2},\ \ 8.5 \le x \le 10.5.$$

Similarly, the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx = \int_{8.5}^{10.5} \frac{1}{240}dy = \frac{1}{120},\ \ 120 \le y \le 240.$$

(iii) The probability that a healthy person has a cholesterol level between 150 to 200, is

$$P[150 \le Y \le 200] = \int_{150}^{200} f_Y(y)dy = \frac{5}{12}.$$

(iv) We have

$$f_X(x)f_Y(y) = \frac{1}{2} \times \frac{1}{120} = \frac{1}{240} f_{XY}(x,y).$$

This shows that X and Y are independent.

(v) We find

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x,y)dxdy = \int_{120}^{240}\int_{8.5}^{10.5} \frac{x}{240}dxdy = 9.5,$$
$$E[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{XY}(x,y)dxdy = \int_{120}^{240}\int_{8.5}^{10.5} \frac{y}{240}dxdy = 180,$$
$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)dxdy = \int_{120}^{240}\int_{8.5}^{10.5} \frac{xy}{240}dxdy = 1710.$$

Hence, $\text{Cov}(X,Y) = E[XY] - E[X]E[Y] = 1710 - 9.5 \times 180 = 0$.

Theorem: If X and Y are two independent random variables with joint density f_{XY}, then show that $E[XY] = E[X]E[Y]$, that is, $\text{Cov}(X,Y) = 0$.

Proof. We have

$$\begin{aligned}
E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x) f_Y(y)dxdy \quad (\because f_{XY}(x,y) = f_X(x)f_Y(y) \text{ as } X \text{ and } Y \text{ are given independent.}) \\
&= \int_{-\infty}^{\infty} y f_Y(y) \left[\int_{-\infty}^{\infty} x f_X(x)dx\right] dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) E[X]dy \\
&= E[X]\int_{-\infty}^{\infty} y f_Y(y)dy \\
&= E[X]E[Y].
\end{aligned}$$

Note. Converse of the above result need not be true, that is, if $E[XY] = E[X]E[Y]$, then X and Y need not be independent. For instance, see the following table for the joint density function of a two dimensional discrete random variable (X,Y).

X/Y	-2	-1	1	2	$f_X(x)$
1	0	1/4	1/4	0	1/2
4	1/4	0	0	1/4	1/2
$f_Y(y)$	1/4	1/4	1/4	1/4	1

We find that $E[X] = 5/2$, $E[Y] = 0$ and $E[XY] = 0$. So $E[XY] = E[X]E[Y]$. Next, we see that $f_X(1) = 1/2$, $f_Y(-1) = 1/4$ and $f_{XY}(1,-1) = 1/4$. So $f_X(1)f_Y(-1) \neq f_{XY}(1,-1)$, and hence X and Y are not independent. In fact, we can easily observe the dependency $X = Y^2$. Thus, covariance between X and Y gives only a rough indication of any association that may exist between X and Y. Also it does not describe the type or strength of the association. The linear relationship between X and Y can be predicted by using a measure known as Pearson coefficient of correlation.

Pearson coefficient of correlation

If X and Y are two random variables with means μ_X, μ_Y, and variances σ_X^2 and σ_Y^2, then correlation between X and Y is given by

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

It can be proved that ρ_{XY} lies in the range $[-1,1]$. Further, $|\rho_{XY}| = 1$ if and only if $Y = \beta_0 + \beta_1 X$ for some real numbers β_0 and $\beta_1 \neq 0$.

Note that if $\rho_{XY} = 0$, we say that X and Y are uncorrelated. It does not imply that X and Y are unrelated. Of course, the relationship, if exists, would not be linear.

In Robot's example, $\sigma_X^2 = 0.146$, $\sigma_Y^2 = 0.268$, $\text{Cov}(X,Y) = 0.046$ and therefore $\rho_{XY} = 0.23$.

Conditional densities and regression

Let (X,Y) be a two dimensional random variable with joint density f_{XY} and marginal densities f_X and f_Y. Then the conditional density of X given $Y = y$, denoted by $f_{X/y}$, is defined as

$$f_{X/y} = \frac{f_{XY}(x,y)}{f_Y(y)}, \quad (f_Y(y) > 0).$$

Similarly, the conditional density of Y given $X = x$, denoted by $f_{Y/x}$, is defined as

$$f_{Y/x} = \frac{f_{XY}(x,y)}{f_X(x)}, \quad (f_X(x) > 0).$$

The mean value of X given $Y = y$, denoted by $\mu_{X/y}$, is given by

$$\mu_{X/y} = E[X/Y = y] = \int_{-\infty}^{\infty} x f_{X/y} dx.$$

The graph of $\mu_{X/y}$ versus y is called regression line of X on Y. Similarly, the graph of

$$\mu_{Y/x} = E[Y/X = x] = \int_{-\infty}^{\infty} y f_{Y/x} dy,$$

versus x is called regression line of Y on X.

Ex. The joint density function of (X,Y) is

$$f_{XY}(x,y) = c/x, \ 27 \le y \le x \le 33.$$

(i) Find the value of c.
(ii) Find the marginal densities and hence check the independence of X and Y
(iii) Evaluate $P[X \le 30, \ Y \le 28]$.
(iv) Find conditional densities $f_{X/y}$ and $f_{Y/x}$. Evaluate $P[X > 32|y = 30]$ and $\mu_{X/y=30}$.
(v) Find the curves of regression of X on Y and Y on X.

Sol. (i) To find c, we use $\int_{27}^{33}\int_{y}^{33} f_{XY}(x,y)dxdy = 1$ and we get $c = \frac{1}{6-27\ln 33/27}$.

(ii) $f_X(x) = \int_{27}^{x} \frac{c}{x} dx = c(1 - 27/x)$, $27 \le x \le 33$
$f_Y(y) = \int_{y}^{33} \frac{c}{x} dx = c(\ln 33 - \ln y)$, $27 \le y \le 33$.
We observe that $f_{XY}(x,y) = c/x \ne f_X(x) f_Y(y)$. So X and Y are not independent.

(iii) $P[X \le 30, \ Y \le 28] = \int_{27}^{28}\int_{y}^{30} \frac{c}{x} dxdy = 0.15$.

(iv) We have

$$f_{X/y} = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{1}{x(\ln 33 - \ln y)}, \quad y \le x \le 33.$$

$$f_{Y/x} = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{1}{x - 27}, \quad 27 \le y \le x.$$

$$P[X > 32|y = 30] = \int_{32}^{33} f_{X/y=30} dx = \int_{32}^{33} \frac{1}{x(\ln 33 - \ln 30)} dx = 0.32.$$

$$\mu_{X/y=30} = \int_{30}^{33} x f_{X/y=30} dx = \int_{30}^{33} \frac{1}{\ln 33 - \ln 30} dx = 31.48.$$

(v) Curve of regression of X on Y is

$$\mu_{X/y} = \int_y^{33} x f_{X/y} dx = \int_y^{33} \frac{1}{\ln 33 - \ln y} dx = \frac{33 - y}{\ln 33 - \ln y}.$$

Curve of regression of Y on X is

$$\mu_{Y/x} = \int_{27}^{x} y f_{Y/x} dx = \int_{27}^{x} \frac{y}{x - 27} dx = \frac{1}{2}(x + 27).$$

Chapter 6

The inferential statistics is essentially based on random sampling from the population. So it is important to understand the meaning of random sample.

Random Sample

A random sample of size n from the distribution of X is a collection n independent random variables, each with the same distribution as of X.

It may noted that the term "random sample" is used in three different but closely related ways in applied statistics. It may refer to objects for study or to the random variables associated with the selected objects for study or to the numerical values assumed by the associated random variables as illustrated in the following example.

Suppose we wish to find the mean effective life of lithium batteries used in a particular model of pocket calculator so that a limited warranty can be placed on the product. For this purpose, we randomly choose n batteries from the population of batteries. Here, prior to the actual selection of the batteries, the life span X_i $(i = 1, 2, ..., n)$ of the ith battery is a random variable. It has the same distribution as X, the life span of batteries in the population. The random variables X_i are independent in the sense that the value assumed by one has no effect on the value assumed by any other variable. Thus, the random variables $X_1,\ X_2,\ .....,\ X_n$ constitute a random sample. For the selected sample of n batteries, the random variables $X_1,\ X_2,\ .....,\ X_n$ shall assume n real values $x_1,\ x_2,\ .....,\ x_n$.

In the above example, the selected n batteries, the associated n random variables $X_1,\ X_2,\ .....,\ X_n$ and the values $x_1,\ x_2,\ .....,\ x_n$ assumed by the n random variables, all refer to random sample in the context under consideration.

Statistics

A statistic is a random variable whose numerical value can be determined from the random sample. In other words, a statistic is a random variable that is a function of the variables $X_1,\ X_2,\ .....,\ X_n$ in the random sample. Some statistics are described in the following.

Sample Mean

Let $X_1,\ X_2,\ ....,\ X_n$ be a random sample from the distribution of X. Then the statistic $\sum_{i=1}^{n} X_i/n$ is called the sample mean and is denoted by $\overline{X}$. So $\overline{X} = \sum_{i}^{n} X_i/n$ is the mean of the sample $X_1,\ X_2,\ ....,\ X_n$.

Sample Median

Let $x_1,\ x_2,\ ....,\ x_n$ be a random sample of observations arranged in the order from the smallest to the largest. The sample mean is the middle observation if n is odd otherwise it is the average of the two middle observations.

Sample Variance

Let X_1, X_2, ..., X_n be a random sample of size n from the distribution of X. Then the statistic

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

is called the sample variance. The statistic $S = \sqrt{S^2}$ is called the sample standard deviation.

Important Remark: It can be shown that the statistics S^2 tends, on the average, to underestimate σ^2, the population variance. To improve the situation, $\sum_{i=1}^{n}(X_i - \overline{X})^2$ is divided by $n-1$ in place of n. In this way, S^2 is unbiased for σ^2, that is, centred at the right spot. In case, X_1, X_2, ..., X_n constitute the entire population, then

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

A computational formula for the sample variance is

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

Sample Range

The sample range is defined as the difference between the largest and the smallest observations.

Ex. A random sample of 9 observations is given as follows:

310 400 406 410 450 395 401 408 415

Find sample mean, median, variance, standard deviation and range. (Ans. Sample mean= 408.3, Median= 406, Variance= 303.25, Standard deviation= 17.4, Range= 60).

Chapter 7

Unbiased point estimator

A parameter $\hat{\theta}$ is an unbiased estimator for a parameter θ if and only if $E[\hat{\theta}] = \theta$. For example, if X_1, X_2, ..., X_n is a random sample of size n from a distribution with mean μ, then the sample mean $\overline{X}$ is an unbiased estimator for μ. For,

$$E[\overline{X}] = E[(X_1 + X_2 + ... + X_n)/n] = (E[X_1] + E[X_2] + ... + E[X_n])/n = (\mu + \mu + ... + \mu)/n = (n\mu)/n = \mu$$

since X_1, X_2, ..., X_n constitute the random sample from the distribution having mean μ, so each of the random variables X_i has mean μ.

It is desirable that the unbiased estimator $\hat{\theta}$ has a small variance for large sample sizes.

Standard error of mean

Let $\overline{X}$ denote the sample mean of a sample of size n drawn from a distribution with standard deviation σ. Then standard deviation of $\overline{X}$ can be proved to be $\sigma/\sqrt{n}$ and is called standard error of mean.

Unbiased estimator for variance

Let S^2 be the sample variance based on a random sample of size n from a distribution with mean μ and variance σ^2. Then it can be proved that

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

is an unbiased estimator for the population variance σ^2. Also, it can be shown that S is not unbiased for σ. This emphasizes the fact that unbiasedness is desirable but not essential in an estimator.

Method of moments for finding point estimators

In many cases, the moments involve the parameter θ to be estimated. We can often obtain a reasonable estimator for θ by replacing the theoretical moments by their estimates based drawn sample and solving the resulting equations for the estimator $\overline{\theta}$, as illustrated in the following example.

Ex. A forester plants 5 rows of pine seedlings with 20 pine seedlings in each row. Let X denotes the number of seedlings per row that survive the first winter. Then X follows a binomial distribution with $n = 20$ and unknown p. Find an estimate of p given that $X_1 = 18$, $X_2 = 17$, $X_3 = 15$, $X_4 = 19$, $X_5 = 20$.

Sol. We have $E[X] = np$. So $\overline{X} = 20\hat{p}$. It follows that

$$\sum_{i=1}^{5} X_i/5 = 20\hat{p} \quad \text{or} \quad 17.8 = 20\hat{p}$$

Finally, we get $\hat{p} = 0.89$, the estimate for p.

Note: If there are two parameters to be estimated, then we need to search two equations involving the parameters and moments.

Method of maximum likelihood

Obtain a random sample X_1, X_2, ..., X_n from the distribution of a random variable X with density f and associated parameter θ. Define a function $L(\theta)$ given by

$$L(\theta) = \sum_{i=1}^{n} f(x_i)$$

known as the likelihood function for the sample. Find the expression for θ that maximizes the likelihood function. Note that the likelihood function gives the probability of getting the sample X_1, X_2, ..., X_n from the distribution of the random variable X. So we find the value of θ that maximizes the value of the likelihood function. This value of θ serves as an estimate for the parameter θ.

Ex. Let X_1, X_2, ..., X_n be a random sample from a normal distribution with mean μ and variance σ^2. The density for X is

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

Therefore, the likelihood function reads as

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

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

Putting the partial derivatives of $\ln L(\mu,\sigma)$ equal to 0, we find

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

Thus, the maximum likelihood estimators for the parameters μ and σ^2 are

$$\hat{\mu} = \overline{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2.$$

Note: The estimator obtained from the method of moments often agrees with the one obtained from the method of maximum likelihood. If it does not happen in some case, then the maximum likelihood estimator is preferred.

Theorem: Let X_1 and X_2 be independent random variables with mgf $m_{X_1}(t)$ and $m_{X_2}(t)$ respectively. Let $Y = X_1 + X_2$. Then the mgf of Y is given by

$$m_Y(t) = m_{X_1}(t)m_{X_2}(t).$$

Proof: We have

$$m_Y(t) = E[e^{tY}] = E[e^{tX_1+tX_2}] = E[e^{tX_1}]E[e^{tX_2}] = m_{X_1}(t)m_{X_2}(t),$$

since e^{tX_1} and e^{tX_2} are independent as X_1 and X_2 are independent.

Ex. The mgf of a normal random variable with mean μ and variance σ^2 is $m_X(t) = e^{\mu t+\frac{1}{2}\sigma^2t^2}$. Let X_1, X_2, ..., X_n be independent normal variables with means μ_1, μ_2, ..., μ_n and variance σ_1^2, σ_2^2, ..., σ_n^2, respectively. Let $Y = X_1 + X_2 + ... + X_n$. Then, we have

$$m_Y(t) = \prod_{i=1}^{n} m_{X_i}(t) = e^{\left(\sum_{i=1}^{n}\mu_i\right)+\frac{1}{2}\left(\sum_{i=1}^{n}\sigma_i^2\right)t^2}$$

Theorem: Let X be a random variable with mgf $m_X(t)$, and $Y = \alpha + \beta X$. Then the mgf of Y is given by

$$m_Y(t) = e^{\alpha t}m_X(\beta t).$$

Proof: We have

$$m_Y(t) = E[e^{tY}] = E[e^{\alpha t+\beta tX}] = E[e^{\alpha t}e^{\beta tX}] = e^{\alpha t}E[e^{\beta tX}] = e^{\alpha t}m_X(\beta t).$$

Theorem: Let X_1, X_2, ..., X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $\overline{X}$ is normally distributed with mean μ and variance σ^2/n.
Proof: We know that

$$m_X(t) = e^{\mu t+\frac{1}{2}\sigma^2t^2}.$$

So by the previous theorem, we have

$$m_{\frac{X}{n}}(t) = e^{\left(\frac{\mu}{n}\right)t+\frac{1}{2}\left(\frac{\sigma^2}{n^2}\right)t^2}.$$

It follows that

$$m_{\overline{X}}(t) = m_{\frac{X_1+X_2+.....+X_n}{n}}(t) = m_{\frac{X_1}{n}}(t)m_{\frac{X_2}{n}}(t)....m_{\frac{X_n}{n}}(t) = e^{\left(\frac{\mu}{n}+....\frac{\mu}{n}\right)t+\frac{1}{2}\left(\frac{\sigma^2}{n^2}+....+\frac{\sigma^2}{n^2}\right)t^2} = e^{\mu t+\frac{1}{2}\left(\frac{\sigma^2}{n}\right)t^2}.$$

Thus, $\overline{X}$ is normally distributed with mean μ and variance σ^2/n.

Confidence interval

A $100(1-\alpha)\%$ confidence interval for a parameter θ is a random interval $[L_1, L_2]$ such that $P[L_1 \leq \theta \leq L_2] = 1-\alpha$, regardless the value of θ.

Confidence interval for μ of normal distribution with known σ

Let X_1, X_2, ..., X_n be a random sample of size n from a normal distribution with unknown mean μ and known variance σ^2. Then $\overline{X}$ is normally distributed with mean μ and variance σ^2/n. Therefore, $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ follows a standard normal distribution. We utilize this fact to find confidence intervals for the unknown μ. Let us find 95% confidence interval for μ. From the normal probability distribution table, we have

$$P[-1.96 \leq Z \leq 1.96] = F(1.96) - F(-1.96) = 0.95.$$

$$\text{or} \quad P[-1.96 \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96] = 0.95.$$

$$\therefore \quad P[\overline{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \overline{X} + 1.96\sigma/\sqrt{n}] = 0.95.$$

Thus, 95% confidence interval for μ is $[L_1, L_2] = [\overline{X} - 1.96\sigma/\sqrt{n}, \overline{X} + 1.96\sigma/\sqrt{n}]$.

In general, $100(1-\alpha)\%$ confidence interval for μ is $[L_1, L_2] = [\overline{X} - z_{\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{\alpha/2}\sigma/\sqrt{n}]$. Here $z_{\alpha/2}$ is the value of $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ such that $P[Z > z_{\alpha/2}] = P[Z < -z_{\alpha/2}] = \alpha/2$. Obviously, $P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1-\alpha$.

Note: If the sample is drawn from a non-normal distribution, then the following theorem helps us in getting the confidence intervals for μ.

Ex. Find the 95% confidence interval for mean of population given a sample

8.0	13.6	13.2	13.6
12.5	14.2	14.9	14.5
13.4	8.6	11.5	16.0
14.2	19.0	17.9	17.0

and population variance $\sigma^2 = 9$.

Sol. Here $n = 16$, $\overline{X} = 13.88$ and $\sigma = 3$. So 95% confidence limits are given by

$$L_1 = \overline{X} - 1.96\sigma/\sqrt{n} = 13.88 - 1.96(3/4) = 12.41,$$

$$L_2 = \overline{X} + 1.96\sigma/\sqrt{n} = 13.88 + 1.96(3/4) = 15.35.$$

Central limit theorem

Let X_1, X_2, ..., X_n be a random sample of size n from a distribution with mean μ and variance σ^2. Then for large n, $\overline{X}$ is approximately normal with mean μ and variance σ^2/n. Furthermore, for large n, $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is approximately normal.

Empirical studies have shown that the above theorem holds for $n \geq 25$.

Chapter 8

We have seen how to estimate both mean and variance of a distribution via point estimation. We have also seen how to construct a confidence interval for the mean of a normal distribution when its variance is assumed to be known. Unfortunately, in most of the statistical studies, the assumption that σ^2 is known is unrealistic. If it is necessary to estimate the mean of a distribution, then its variance is usually unknown. In what follows we shall learn how to make inferences on the mean and variance when both of these parameters are unknown.

Confidence interval for σ^2 of normal distribution with unknown μ

Let X_1, X_2, ..., X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then it can be proved that the random variable

$$(n-1)S^2/\sigma^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2$$

has a chi-squared distribution[1]with $n-1$ degrees of freedom.

Denoting $X^2_{n-1} = (n-1)S^2/\sigma^2$, let us find 95% confidence interval for σ^2. Let $\chi^2_{0.025}$ and $\chi^2_{0.975}$ denote the values of X^2_{n-1} such that $P[X^2_{n-1} \geq \chi^2_{0.025}] = 0.025$ and $P[X^2_{n-1} \geq \chi^2_{0.975}] = 0.975$. Obviously, we have

$$P[\chi^2_{0.975} \leq X^2_{n-1} \leq \chi^2_{0.025}] = 0.95.$$

or $\quad P[\chi^2_{0.975} \leq (n-1)S^2/\sigma^2 \leq \chi^2_{0.025}] = 0.95.$

$\therefore \quad P[(n-1)S^2/\chi^2_{0.025} \leq \sigma^2 \leq (n-1)S^2/\chi^2_{0.975}] = 0.95.$

Thus, 95% confidence interval for σ^2 is $[L_1, L_2] = [(n-1)S^2/\chi^2_{0.025}, (n-1)S^2/\chi^2_{0.975}]$.

In general, $100(1-\alpha)\%$ confidence interval for σ^2 is $[L_1, L_2] = [(n-1)S^2/\chi^2_{\alpha/2}, (n-1)S^2/\chi^2_{1-\alpha/2}]$.

Ex. Find the 95% confidence interval for σ^2 of a normal population based on the following sample:

3.4	3.6	4.0	0.4	2.0
3.0	3.1	4.1	1.4	2.5
1.4	2.0	3.1	1.8	1.6
3.5	2.5	1.7	5.1	0.7
4.2	1.5	3.0	3.9	3.0

Sol. Here $n = 25$ and $S^2 = 1.408$. From the χ^2 probability distribution table, for 24 degrees of freedom, we have $\chi^2_{0.025} = 39.4$ and $\chi^2_{0.975} = 12.4$. So 95% confidence limits are given by

$$L_1 = (n-1)S^2/\chi^2_{0.025} = 24(1.408)/39.4 = 0.858,$$

$$L_2 = (n-1)S^2/\chi^2_{0.975} = 24(1.408)/12.4 = 2.725.$$

---

[1]A random variable X is said to follow chi-square distribution with γ degrees of freedom if its density function is given by

$$f(x) = \frac{1}{\Gamma(\gamma/2)2^{\gamma/2}} x^{\gamma/2-1}e^{-x/2}, \quad x > 0.$$

. We denote the chi-square random variable with n degrees of freedom by X^2_γ.

If X is a normal random variable with mean μ and variance σ^2. Then it can be proved that the square of the standard normal variable, that is, $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ follows a chi-square distribution with one degree of freedom. Also, it can be proved that the sum of independent chi-square random variables is also a chi-square random variable with degrees of freedom equal to the sum of degrees of freedom of all the independent random variables. It follows that if X_1, X_2, ..., X_n is a random sample of size n from a normal distribution with mean μ and variance σ^2, then $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ is a chi-square random variable with n degrees of freedom.

Confidence interval for μ of normal distribution with unknown σ^2

Let X_1, X_2, ..., X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then it can be proved that the random variable

$$\frac{\overline{X}-\mu}{S/\sqrt{n}}$$

follows a T distribution[2] with $n-1$ degrees of freedom.

Denoting $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$, let us find 95% confidence interval for μ. Let $t_{0.025}$ and $t_{-0.025}$ denote the values of T_{n-1} such that $P[T_{n-1} \le t_{-0.025}] = 0.025 = P[T_{n-1} \ge t_{0.025}]$. Obviously, we have

$$P[t_{-0.025} \le T_{n-1} \le t_{0.025}] = 0.95.$$

$$\text{or} \quad P[t_{-0.025} \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{0.025}] = 0.95.$$

$$\text{or} \quad P[-t_{0.025} \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t_{0.025}] = 0.95. \quad \text{(Because of symmetry of T distribution, } t_{-0.025} = -t_{0.025})$$

$$\therefore \quad P[\overline{X} - t_{0.025}S/\sqrt{n} \le \mu \le \overline{X} + t_{0.025}S/\sqrt{n}] = 0.95.$$

Thus, 95% confidence interval for μ is $[L_1, L_2] = [\overline{X} - t_{0.025}S/\sqrt{n}, \overline{X} + t_{0.025}S/\sqrt{n}]$.

In general, $100(1-\alpha)\%$ confidence interval for μ is $[L_1, L_2] = [\overline{X} - t_{\alpha/2}S/\sqrt{n}, \overline{X} + t_{\alpha/2}S/\sqrt{n}]$.

Ex. Find the 95% confidence interval for μ of a normal population based on the following sample:

52.7	43.9	41.7	71.5	47.6	55.1
62.2	56.5	33.4	61.8	54.3	50.0
45.3	63.4	53.9	65.5	66.6	70.0
52.4	38.6	46.1	44.4	60.7	56.4

Sol. Here $n = 24$, $\overline{X} = 53.92$ and $S = 10.07$. From the T probability distribution table, for 23 degrees of freedom, we have $t_{0.025} = 2.069$. So 95% confidence limits are given by

$$L_1 = \overline{X} - t_{0.025}S/\sqrt{n} = 53.92 - 2.069(10.07)/\sqrt{24} = 49.67,$$

$$L_2 = \overline{X} + t_{0.025}S/\sqrt{n} = 53.92 + 2.069(10.07)/\sqrt{24} = 58.17.$$

[2]If Z is a standard normal variable and X_γ^2 is an independent chi-squared random variable with γ degrees of freedom, then the random variable $T_\gamma = Z/\sqrt{X_\gamma^2/\gamma}$ is said to follow a T distribution with γ degrees of freedom.

The density function of a T_γ random variable reads as

$$f(t) = \frac{\Gamma(\gamma+1)/2}{\Gamma(\gamma/2)\sqrt{\pi\gamma}}\left(1+\frac{t^2}{\gamma}\right)^{-(\gamma+1)/2}, \quad -\infty < t < \infty.$$

The graph of this density function is symmetric about the line $t = 0$ and tends to the standard normal curve as the number of degrees of freedom γ increases.

Hypothesis testing

In the theory of hypothesis testing, the experimenter/researcher proposes a hypothesis on population parameter θ. The hypothesis proposed by the experimenter/researcher is known as alternative or research hypothesis and is denoted by H_1. Negation of H_1 is called null hypothesis and is denoted by H_0. While testing a hypothesis on population parameter θ, the statement of equality $\theta = \theta_0$ (known as null value of θ), is always included in H_0. Further, H_1 being the research hypothesis, it is expected that the evidence leads us to reject H_0 and thereby to accept H_1.

Ex. Highway engineers think that the reflective highway signs do not perform properly because more than 50% of the automobiles on the road have misaimed headlights. If this contention is supported statistically, a tougher inspection program will be put into operation. Let p denote the proportion of automobiles with misaimed headlights. Since the engineers wish to support $p > 0.5$, so the research hypothesis H_1 and the null hypothesis H_0 are

$$H_1 : \ p > 0.5$$

$$H_0 : \ p \leq 0.5$$

Note that $p = 0.5$, the null value of p, is included in H_0.

Errors in Hypothesis testing

In hypothesis testing, we consider a test statistic whose values that lead to the rejection of H_0 are set before the experiment is conducted. These values of the test statistic constitute critical or rejection of the test.

Type I Error: The probability that the observed value of the test statistic will fall in the critical region when $\theta = \theta_0$ is called the size of the test or level of significance of the test, and is denoted by α. If this occurs, a Type I error is committed. Thus, α is the probability of committing Type I error. It is possible that the observed value of the test statistic falls into the H_1 region even though H_0 is true and should not be rejected. The probability α of getting the text statistic value into the H_1 region is, therefore, the Type I error.

Ex. A random sample of 20 cars is selected and the headlights are tested. Let us design a test so that α, the probability of rejecting H_0 when $p = 0.5$, is about 0.05. The test statistic that we shall use is X, the number of cars in the sample with misaimed headlights. Then X follows a binomial distribution with $n = 20$, $p = 0.5$ and $E[X] = np = 10$. So on an average 10 of every 20 cars tested are expected to have misaimed headlights. So logically we should reject H_0 if the observed value of X is somewhat greater than 10. Also, from the binomial probability distribution table, we observe that $P[X \leq 13 : p = 0.5] = 0.9423$. Therefore,

$$P[X \geq 14 : p = 0.5] = 1 - P[X \leq 13 : p = 0.5] = 1 - 0.9423 = 0.0577.$$

Let us choose $\alpha = 0.0577$. It implies that we agree to reject H_0 in favor of H_1 if the observed value of $X \geq 14$. In this way, we have split the values of X into two sets $C = \{14, 15, ..., 20\}$ and $C' = \{0, 1, 2, ..., 13\}$. Here C is the critical or rejection region of the test. If the observed value of X lies in C, we reject H_0 and conclude that majority of cars have misaimed headlights.

Type II Error: It is possible that the observed value of the test statistic falls into the H_0 region even though H_0 is not true and should be rejected. So we fail to reject H_0 even though it is not true. If this occurs a Type II error is committed. The probability of committing Type II error is denoted by β. The value of β is harder to predict directly. It is usually calculated subject to some give value of alternative. It is obviously the probability of getting the test statistic value in the H_0 region subject to the given alternative.

Note: When we fail to reject H_0, we say that the observed sample for the test statistic is not enough to support H_1.

Ex. Suppose that, unknown to the researcher, the true proportion of the cars with misaimed headlights is $p = 0.7$. What is the probability that the $\alpha = 0.0577$ test, as designed in the previous example, is unable to detect this situation? For this, we calculate

$$\beta = P[X \leq 13 : p = 0.7] = 0.392.$$

Power: Suppose a researcher puts a great deal of time, effort and money into designing and carrying out an experiment to gather evidence to support a research theory. Therefore, the researcher would like to have the probability of rejecting the null hypothesis when the research theory is true. This probability is called power of the test.

Note that both the probabilities, power and β, are calculated under the assumption that the research theory is true. The researcher will either fail to reject the null hypothesis with probability β or will reject the null hypothesis with probability power.

$$\therefore \quad \beta + \text{power} = 1 \quad \text{or} \quad \text{power} = 1 - \beta.$$

In the previous example, we found $\beta = 0.392$ under the assumption that the research theory is true ($p = 0.7$). Therefore, power$= 1 - 0.392 = 0.608$.

Significance testing

Suppose we want to test

$$H_0 : p \leq 0.1$$

$$H_1 : p > 0.1$$

based on a sample of size 20. Let the test statistic is X, the number of successes that are observed in 20 trials. If $p = 0.1$, the null value of p, then X follows a binomial distribution with mean $E[X] = 20(0.1) = 2$. So values of X somewhat greater than 2 will lead to the rejection of null hypothesis. Suppose we want α to be very small, say 0.0001. From the binomial probability distribution table, we have

$$P[X \geq 9 : p = 0.1] = 1 - P[X \leq 8 : p = 0.1] = 1 - 0.9999 = 0.0001.$$

So the critical region of the test is $\{C = 9, 10, ..., 20\}$. Now suppose we conduct the test and observe 8 successes. It does not fall into C. So via our rigid rule of hypothesis testing we are unable to reject H_0. However, a little thought should make us a bit uneasy with this decision. We find

$$P[X \geq 8 : p = 0.1] = 1 - P[X \leq 7 : p = 0.1] = 1 - 0.9996 = 0.0004.$$

It means we are willing to tolerate 1 chance in 10000 of making a Type I error. But we shall declare 4 chances in 10000 of making such an error too large to risk. There is so little difference between these probabilities that it seems a bit silly to insist with our original cut off value 9.

Such a problem can be avoided by adopting a technique known as significance testing where we do not preset α and hence do not specify a rigid critical region. Rather, we evaluate the test statistic and then determine the probability of observing a value of the test statistic at least as extreme as the value noted, under the assumption $\theta = \theta_0$. This probability is known as critical level or descriptive level of significance or P value of the test. We reject H_0 if we consider this P value to be small. In case, an α level has been preset to ensure that a traditional or industry maximum acceptable level is met, we compare the P value with the preset α value. If $P \leq \alpha$, then we can reject the null hypothesis atleast at the stated level of significance.

Ex. Automotive engineers are using more and more aluminium in manufacturing the automobiles in hopes of reducing the cost and improving the petrol mileage. For a particular model, the mileage on highway has a mean 26 kmpl with a standard deviation of 5 kmpl. It is hoped that a new design manufactured by using more aluminium will increase the mean petrol mileage on highway maintaining the standard deviation of 5 kmpl. So we test the hypothesis

$$H_0 : \mu \leq 26$$

$$H_1 : \mu > 26 \text{ (the new design incereases the petrol mileage on highway)}$$

Suppose 36 vehicles with new design are tested on highway and the mean petrol mileage is found to be 28.04 kmpl. Here, $n = 36$ and sample mean is $\overline{X} = 28.04$. We choose $\overline{X}$ as the test statistic since $\overline{X}$ is an unbiased estimator for the population mean μ. We know $\overline{X}$ is approximately normally distributed with mean $\mu = 26$ and

$\sigma = 5/\sqrt{36} = 5/6$. Therefore $Z = (\overline{X} - \mu)/\sigma/\sqrt{n} = (\overline{X} - 26)/(5/6$ is the corresponding standard normal variate. So P value of the test is given by

$$P[\overline{X} \geq 28.04 : \mu = 26] = P[Z \geq 2.45] = 1 - P[Z \leq 2.45] = 1 - 0.9929 = 0.0071.$$

There are two explanations of this very small probability. First, the null hypothesis H_0 is true and we have observed a very rare sample that by chance has a large mean. Second, the new design with more aluminium has, in fact, resulted in a higher mean petrol mileage. We prefer the second explanation as it supports our research hypothesis H_1. That is, we shall reject H_0 and report that P value of our test is 0.0071. In case, there is some prestated level of significance say $\alpha = 0.05$. We can safely reject H_0 at this level of significance since the P value 0.0071 of our test is less than $\alpha = 0.05$.

Note: Significance testing is a widely used concept as it is more appealing than hypothesis testing. For a right tailed test ($H_1 : \theta > \theta_0$), the P value is the area under the probability curve to the right of the observed value of the test statistic while for a left-tailed test ($H_1 : \theta < \theta_0$), it is the area to the left. For a two-tailed test ($H_1 : \theta \neq \theta_0$), where the distribution is symmetric as it is for Z or T statistic, it is logical to double the apparent one-tailed P value. If the distribution is not symmetric as it is for chi-squared statistic, then presumably the two-tailed P value is nearly double the one-tailed value.

Hypothesis and significance tests on the mean

There are three forms for the tests of hypotheses on the mean μ of a distribution.

I	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$ (Right-tailed test)	II	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$ (Left-tailed test)	III	$H_0 : \mu = \mu_0$ $H_0 : \mu \neq \mu_0$ (Two-tailed test)

Tests of hypothesis on μ are actually conducted by testing $H_0 : \mu = \mu_0$ against one of the alternatives $\mu > \mu_0$, $\mu < \mu_0$ and $\mu \neq \mu_0$. In particular, the values of the test statistic that lead us to reject μ_0 and to conclude that $\mu > \mu_0$ will also lead us to reject any value less than μ_0. Similarly, the values of the test statistic that lead us to reject μ_0 and to conclude that $\mu < \mu_0$ will also lead us to reject any value greater than μ_0. For this reason many statisticians prefer to write the above three tests as

I	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$ (Right-tailed test)	II	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$ (Left-tailed test)	III	$H_0 : \mu = \mu_0$ $H_0 : \mu \neq \mu_0$ (Two-tailed test)

This emphasizes the fact that while performing a hypothesis test on μ, α is computed assuming that $\mu = \mu_0$. Similarly, while performing a significance test on μ, P value is computed under the assumption $\mu = \mu_0$.

Ex. The maximum acceptable level for exposure to microwave radiation in Mumbai is an average of 10 microwatts per square centimeter. It is feared that a large television transmitter may be polluting the air nearby by pushing the level of microwave radiation above the safe limit. So we want to test

$H_0 : \mu \leq 10$

$H_1 : \mu > 10$ (unsafe)

Obviously, a right-tailed test is applicable here. Suppose a sample of 25 readings is to be obtained. Then our test statistic $(\overline{X} - 10)/(S/\sqrt{25})$ follows a T_{24} distribution when $\mu = 10$. Let us preset α. If we make a Type I error, we shall shut down the transmitter unnecessarily. On the other hand, if we make a Type II error, we shall fail to detect potential health hazard. We want α small but not so small as to force β very large. Let us choose $\alpha = 0.1$. From the T distribution probability table, we find that the critical point of the test is 1.318. Suppose the sample of 25 readings gives $\overline{X} = 10.3$ and $S = 2$. So the observed value of the test statistic is $(\overline{X} - 10)/(S/\sqrt{25}) = (10.3 - 10)/(2/5) = 0.75$, which is less than the critical value 1.318. Therefore, we are unable to reject H_0 and conclude that the observed data do not support the contention that the transmitter is forcing the average microwave level level above the safe limit.

Now, let us find the P value of the test, that is, $P[T_{24} \geq 0.75]$. From the T distribution probability table, we find that $P[T_{24} > 0.685] = 1 - P[T_{24} \leq 0.685] = 1 - 0.75 = 0.25$. Also, $P[T_{24} > 1.318] = 1 - P[T_{24} \leq 1.318] = 1 - 0.9 = 0.1$. Next, the observed value of the test statistic is 0.75, which lies between 0.685 and 1.318. It follows that the P value of the test, given by $P[T_{24} \geq 0.75]$, is greater than 0.1 but less than 0.25. Since the P value of the test is greater than the preset α value 0.1. So we are unable to reject H_0 in favor of H_1 at the stated level of significance.

Ex. See example 8.5.5 from the text book for a two-tailed test on mean.

Chapter 9 (9.1, 9.2)

Estimator for proportions

Consider a population of interest where a particular trait is being studied, and each member of the population can be classified as either having or failing to have the trait. Let p be proportion of the population with the trait. In order to find a logical estimator for p, we select a random sample X_1, X_2, ..., X_n of size n from the population where $X_i = 1$ if the ith member of the population has the trait otherwise $X_i = 0$. Then $X = X_1 + X_2 + ... + X_n$ is equal to the number in sample with trait. Therefore, the sample proportion with trait is given by

$$\hat{p} = \frac{X}{n}.$$

It serves as an estimator for the population proportion p. Note that $\hat{p} = \overline{X}$, that is, $\hat{p}$ is the mean of the selected random sample. Therefore, by Central Limit Theorem, $\hat{p}$ is approximately normally distributed with same mean as of each X_i and variance equal to $(\text{Var}X_i)/n$. Now, the density of X_i is given by

$$\begin{array}{rcc} x_i: & 1 & 0 \\ f(x_i): & p & 1-p \end{array}$$

So mean of X_i is $E[X_i] = p$ and $\text{Var}(X_i) = E[X_i^2] - (E[X_i])^2 = p - p^2 = p(1-p)$. Hence, $\hat{p}$ is approximately normal with mean p and variance $p(1-p)/n$. It implies that $Z = (\hat{p} - p)/\sqrt{p(1-p)/n}$ is standard normal variable, and the $100(1-\alpha)\%$ confidence interval on p is $[L_1, L_2]$, where

$$L_1 = \hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n}$$

$$L_2 = \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}$$

Note that we do not know p. So we replace p by its unbiased estimator $\hat{p}$. So we have

$$L_1 = \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

$$L_2 = \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

Ex. In a randomly selected sample of 100 bulbs from the output of a factory, 91 bulbs are found to be working fine without any defect. Find 95% confidence interval on the population proportion of non-defective bulbs.
Sol. Here, $n = 100$ and $\hat{p} = 91/100 = 0.91$. Also, from the normal table, $z_{0.05/2} = 1.96$. So 95% confidence limits on the population proportion of non-defective bulbs are

$$L_1 = \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = 0.91 - 1.96\sqrt{0.91(1-0.91)100} = 0.091 - 0.056 = 0.854$$

$$L_2 = \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = 0.91 + 1.96\sqrt{0.91(1-0.91)100} = 0.091 + 0.056 = 0.966$$

So with 95% confidence, we expect production of non-defective bulbs from the factory between 85.4% and 96.6%.

Sample size for estimating p

A sample based experiment may yield a lengthy confidence interval on p which is virtually useless. The $100(1-\alpha)\%$ confidence interval $[L_1, L_2] = [\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}]$ on p tells us that we are $100(1-\alpha)\%$ sure that p lies in this interval, and consequently $\hat{p}$ differs from p at most by $d = z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. This in turn implies that the sample size n for the confidence interval of desired length $2d$ is given by

$$n = \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{d^2}.$$

Note that this formula can be used if the prior estimate $\hat{p}$ for p is available. Otherwise, we use the formula

$$n = \frac{z^2_{\alpha/2}}{4d^2},$$

since it can be shown that $\hat{p}(1-\hat{p})$ will never exceed $1/4$.

Ex. In the previous example where $\hat{p} = 0.91$, if we desire the length of the confidence interval to be 0.02, that is, $d = 0.01$, then the number of bulbs in the sample should be

$$n = \frac{(1.96)^2(0.91)(1-0.91)}{(0.01)^2} = 3147.$$

In case, the prior estimate $\hat{p}$ for p is not available, then for the 95% confidence interval with length 0.02, we need to select a sample of size

$$n = \frac{(1.96)^2}{4(0.01)^2} = 9604.$$

Testing hypothesis on proportion

Let the null value of population proportion p be p_0. Then for testing the hypothesis $H_0 : p = p_0$ against one of the alternatives $H_1 : p > p_0$, $H_1 : p < p_0$ and $H_1 : p \neq p_0$, we use the test statistic $(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n}$. This statistic is logical since it compares the unbiased point estimator $\hat{p}$ for p to the null value p_0. Furthermore, by Central Limit Theorem, this statistic follows standard normal distribution when $p = p_0$.

Ex. The majority of faults on transmission lines are the result of external influences and are usually transitory. It is thought that more than 70% of the faults are caused by lightening. To gain evidence to support this contention, we test

$$H_0 : p = 0.7$$

$$H_1 : p > 0.7$$

Data gathered over a year-long period show that 151 of 200 faults are observed due to lightening. So the observed value of the test statistic is

$$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n} = (151/200 - 0.7)/\sqrt{0.7(1-0.7)/200} = 1.697$$

From the normal table, we see that

$$P[Z \geq 1.69] = 0.0455, \quad P[Z \geq 1.70] = 0.0446.$$

It implies that the P value of our test lies between 0.0446 and 0.0455. Considering this small p value, we reject H_0 and conclude that $p > 0.7$.

Note. In the above example, the hypothesis testing on p does not assume that the sample size is large. In fact, the criteria $p_0 = 0.7 > 0.5$ and $n(1-p_0) = 200(1-0.7) = 60 > 5$ is met. So the Binomial distribution is approximated by normal distribution.

MORE CONCEPTS SHALL BE ADDED SOON.
Cheers!