12·9·1. Test for Single Proportion. If X is the number of successes in n independent trials with constant probability P of success for each trial (*c.f.* § 7·2·1)

$$E(X) = nP \quad \text{and} \quad V(X) = nPQ,$$

where $Q = 1 - P$, is the probability of failure.

It has been proved that for large n, the binomial distribution tends to normal distribution. Hence for large n, $X \sim N(nP, nPQ)$ *i.e.*,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \sim N(0, 1) \qquad \ldots(12\cdot4)$$

and we can apply the normal test.

Remarks 1. In a sample of size n, let X be the number of persons possessing the given attribute. Then

Observed proportion of successes $= X/n = p$, (say).

$$\therefore \qquad E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}nP = P$$

$$\Rightarrow \qquad E(p) = P \qquad \ldots(12\cdot4a)$$

Thus the sample proportion 'p' gives an unbiased estimate of the population proportion P.

Also $$V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}nPQ = \frac{PQ}{n}$$

$$\therefore \qquad \text{S.E.}(p) = \sqrt{PQ/n} \qquad \ldots(12\cdot4b)$$

Since X and consequently X/n is asymptotically normal for large n, the normal test for the proportion of successes becomes

$$Z = \frac{p - E(p)}{\text{S.E. }(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1) \qquad \ldots(12\cdot4c)$$

2. If we have sampling from a finite population of size N, then

$$S.E.(p) = \sqrt{\left(\frac{N - n}{N - 1}\right) \cdot \frac{PQ}{n}} \qquad \ldots(12\cdot4d)$$

3. Since the probable limits for a normal variate X are $E(X) \pm 3\sqrt{V(X)}$, the probable limits for the observed proportion of successes are :

$$E(p) \pm 3 \text{ S.E. } (p), \textit{ i.e., } P \pm 3\sqrt{PQ/n}.$$

If P is not known then taking p (the sample proportion) as an estimate of P, the probable limits for the proportion in the population are :

$$p \pm 3\sqrt{pq/n} \qquad \ldots(12\text{-}4e)$$

However, the limits for P at level of significance α are given by :

$$p \pm z_\alpha \sqrt{pq/n}, \qquad \ldots(12\text{-}4f)$$

where z_α is the significant value of Z at level of significance α.

In particular 95% confidence limits for P are given by :

$$p \pm 1{\cdot}96\sqrt{pq/n}, \qquad \ldots(12\text{-}4g)$$

and 99% confidence limits for P are given by

$$p \pm 2{\cdot}58\sqrt{pq/n} \qquad \ldots(12\text{-}4h)$$

Example 12·1. *A dice is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the dice cannot be regarded as an unbiased one and find the limts between which the probability of a throw of 3 or 4 lies.*

Solution. If the coming of 3 or 4 is called a success, then in usual notations we are given

$$n = 9{,}000;\ X = \text{Number of successes} = 3{,}240$$

Under the null hypothesis (H_0) *that the dice is an unbiased one*, we get

$$P = \text{Probability of success} = \text{Probability of getting a 3 or 4} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Alternative hypothesis, $H_1 : p \neq \frac{1}{3}$, (*i.e.*, dice is biased).

We have $Z = \dfrac{X - nP}{\sqrt{nQP}} \sim N(0, 1)$, since n is large.

Now $$Z = \frac{3240 - 9000 \times 1/3}{\sqrt{9000 \times (1/3) \times (2/3)}} = \frac{240}{\sqrt{2000}} = \frac{240}{44{\cdot}73} = 5{\cdot}36$$

Since $|Z| > 3$, H_0 is rejected and we conclude that the dice is almost certainly biased.

Since dice is not unbiased, $P \neq \frac{1}{3}$. The probable limits for 'P' are given by :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = p \pm 3\sqrt{pq/n},$$

where $\hat{P} = p = \dfrac{3240}{9000} = 0{\cdot}36$ and $\hat{Q} = q = 1 - p = 0{\cdot}64$.

Hence the probable limits for the population proportion of successes may be taken as

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}36 \pm 3\cdot\sqrt{\frac{0{\cdot}36 \times 0{\cdot}64}{9000}} = 0{\cdot}36 \pm 3 \times \frac{0{\cdot}6 \times 0{\cdot}8}{30\sqrt{10}}$$

$$= 0{\cdot}360 \pm 0{\cdot}015 = 0{\cdot}345 \text{ and } 0{\cdot}375.$$

Hence the probability of getting 3 or 4 almost certainly lies between 0·345 and 0·375.

Example 12·2. *A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the S.E. of the proportion of bad ones in a sample of this size is 0·015 and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8·5 and 17·5.*

Solution. Here we are given $n = 500$

X = Number of bad pineapples in the sample = 65

p = Proportion of bad pineapples in the sample $= \frac{65}{500} = 0{\cdot}13$

$\therefore \quad q = 1 - p = 0{\cdot}87$

Since P, the proportion of bad pineapples in the consignment is not known, we may take (as in the last example)

$$\hat{P} = p = 0{\cdot}13, \quad \hat{Q} = q = 0{\cdot}87$$

S.E. of proportion $= \sqrt{\hat{P}\hat{Q}/n} = \sqrt{0{\cdot}13 \times 0{\cdot}87/500} = 0{\cdot}015$

Thus, the limits for the proportion of bad pineapples in the consignment are :

$$\hat{P} \pm 3\sqrt{\hat{P}\hat{Q}/n} = 0{\cdot}130 \pm 3 \times 0{\cdot}015 = 0{\cdot}130 \pm 0{\cdot}045 = .(0{\cdot}085, 0{\cdot}175)$$

Hence the percentage of bad pineapples in the consignment lies almost certainly between 8·5 and 17·5.

Example 12·4. *In a sample of 1,000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this State at 1% level of significance ?*

Solution. In the usual notations we are given $n = 1{,}000$

X = Number of rice eaters = 540

$\therefore$ p = Sample proportion of rice eaters $= \frac{X}{n} = \frac{540}{1000} = 0{\cdot}54$

Null Hypothesis, H_0 : Both rice and wheat are equally popular in the State so that

P = Population proportion of rice eaters in Maharashtra = 0·5

$\Rightarrow$ $Q = 1 - P = 0{\cdot}5$

Alternative Hypothesis, $H_1 : P \neq 0{\cdot}5$ (two-tailed alternative).

Test Statistic. Under H_0, the test statistic is

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1), \text{ (since } n \text{ is large).}$$

Now $$Z = \frac{0{\cdot}54 - 0{\cdot}50}{\sqrt{0{\cdot}5 \times 0{\cdot}5/1000}} = \frac{0{\cdot}04}{0{\cdot}0138} = 2{\cdot}532$$

Conclusion. The significant or critical value of Z at 1% level of significance for two-tailed test is 2·58. Since computed $Z = 2{\cdot}532$ is less than 2·58, it is not significant at 1% level of significance. Hence the null hypothesis is accepted and we may conclude that rice and wheat are equally popular in Maharashtra State.

12·9·2. Test of Significance for Difference of Proportions. Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say A, among their members. Let X_1, X_2 be the number of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two populations respectively. Then sample proportions are given by

$$p_1 = X_1/n_1 \quad \text{and} \quad p_2 = X_2/n_2$$

If P_1 and P_2 are the population proportions, then

$$E(p_1) = P_1, \; E(p_2) = P_2 \qquad [c.f. \text{ Equation (12·4}a\text{)}]$$

and

$$V(p_1) = \frac{P_1Q_1}{n_1} \quad \text{and} \quad V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for large samples, p_1 and p_2 are asymptotically normally distributed, $(p_1 - p_2)$ is also normally dstributed. Then the standard variable corresponding to the difference $(p_1 - p_2)$ is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0, 1)$$

Under the *null hypothesis* $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference between the sample proportions, we have

$$E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2 = 0 \qquad \text{(Under } H_0\text{)}$$

Also $V(p_1 - p_2) = V(p_1) + V(p_2)$,

the covariance term $\text{Cov}(p_1, p_2)$ vanishes, since sample proportions are independent.

$$\therefore \qquad V(p_1 - p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} = PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right),$$

since under $H_0 : P_1 = P_2 = P$, (say), and $Q_1 = Q_2 = Q$.

Hence under $H_0 : P_1 = P_2$, the test statistic for the difference of proportions becomes

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots(12{\cdot}5)$$

In general, we do not have any information as to the proportion of A's in the populations from which the samples have been taken. Under $H_0 : P_1 = P_2 = P$, (say), an unbiased estimate of the population proportion P, based on both the samples is given by

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} \qquad \ldots(12{\cdot}5a)$$

The estimate is unbiased, since

$$E(\hat{P}) = \frac{1}{n_1 + n_2} E\left[n_1 p_1 + n_2 p_2\right] = \frac{1}{n_1 + n_2}\left[n_1 E(p_1) + n_2 E(p_2)\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 P_1 + n_2 P_2\right] = P \qquad [\because P_1 = P_2 = P, \text{ under } H_0]$$

Thus (12·5) along with (12·5*a*) gives the required test statistic.

Remarks 1. Suppose we want to test the significance of the difference between p_1 and p, where

$$p = \frac{(n_1 p_1 + n_2 p_2)}{(n_1 + n_2)}$$

gives a pooled estimate of the population proportion on the basis of both the samples. We have

$$V(p_1 - p) = V(p_1) + V(p) - 2\,\text{Cov}\,(p_1, p) \qquad \ldots(*)$$

Since p_1 and p are not independent, Cov $(p_1, p) \neq 0$.

$$\text{Cov}.(p_1, p) = E[\{p_1 - E(p_1)\}\{p - E(p)\}]$$

$$= E\left[\{p_1 - E(p_1)\}\left\{\frac{1}{n_1 + n_2}\{n_1p_1 + n_2p_2 - E(n_1p_1 + n_2p_2)\}\right\}\right]$$

$$= \frac{1}{n_1 + n_2} E\left[\{p_1 - E(p_1)\}\left\{n_1(p_1 - E(p_1)) + n_2(p_2 - E(p_2))\right\}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 E\left\{p_1 - E(p_1)\right\}^2 + n_2 E\left\{(p_1 - E(p_1))(p_2 - E(p_2))\right\}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 V(p_1) + n_2\,\text{Cov}\,(p_1, p_2)\right]$$

$$= \frac{1}{n_1 + n_2} n_1 V(p_1), \qquad [\because \text{Cov}\,(p_1, p_2) = 0]$$

$$= \frac{n_1}{n_1 + n_2} \cdot \frac{pq}{n_1} = \frac{pq}{n_1 + n_2}$$

$$\text{Also Var}\,(p) = \frac{1}{(n_1 + n_2)^2} E\left[(n_1 p_1 + n_2 p_2) - E(n_1 p_1 + n_2 p_2)\right]^2$$

$$= \frac{1}{(n_1 + n_2)^2}\left[n_1^2 \text{Var}\,(p_1) + n_2^2 \text{Var}\,(p_2)\right],$$

covariance term vanishes since p_1 and p_2 are independent.

$$\therefore \quad \text{Var}\,(p) = \frac{1}{(n_1 + n_2)^2}\left[n_1^2 \cdot \frac{pq}{n_1} + n_2^2 \cdot \frac{pq}{n_2}\right]$$

$$= \frac{pq}{n_1 + n_2}$$

Substituting in (*) and simplifying, we shall get

$$V(p_1 - p) = \frac{pq}{n_1} + \frac{pq}{n_1 + n_2} - 2\frac{pq}{n_1 + n_2} = pq\left[\frac{n_2}{n_1(n_1 + n_2)}\right]$$

Thus, the test statistic in this case becomes

$$Z = \frac{p_1 - p}{\sqrt{\frac{n_2}{(n_1 + n_2)} \cdot \frac{p\,q}{n_1}}} \sim N(0, 1) \qquad \ldots(12{\cdot}5b)$$

2. Suppose the population proportions P_1 and P_2 are given to be distinctly different, *i.e.*, $P_1 \neq P_2$ and we want to test if the difference $(P_1 - P_2)$ in population proportions is likely to be hidden in simple samples of sizes n_1 and n_2 from the two populations respectively.

We have seen that in the usual notations,

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\text{S.E.}(p_1 - p_2)} = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0, 1)$$

Here sample proportions are not given. If we set up the *null hypothesis* $H_0 : p_1 = p_2$, *i.e.* the samples will not reveal the difference in the population proportions or in other words the difference in population proportions is likely to be hidden in sampling, the test statistic becomes

$$|Z| = \frac{|P_1 - P_2|}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0, 1) \qquad \ldots(12{\cdot}5c)$$

Example 12·6. *Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportions of men and women in favour of the proposal, are same against that they are not, at 5% level.* **[Agra Univ. M.A., 1992]**

Solution. *Null Hypothesis* $H_0 : P_1 = P_2 = P$, (say), *i.e.*, there is no significant difference between the opinion of men and women as far as proposal of flyover is concerned.

Alternative Hypothesis, $H_1 : P_1 \neq P_2$ (two-tailed).

We are given :

$n_1 = 400$, X_1 = Number of men favouring the proposal = 200

$n_2 = 600$, X_2 = Number of women favouring the proposal = 325

$\therefore$ p_1 = Proportion of men favouring the proposal in the sample

$$= \frac{X_1}{n_1} = \frac{200}{400} = 0{\cdot}5$$

p_2 = Proportion of women favouring the proposal in the sample

$$= \frac{X_2}{n_2} = \frac{325}{600} = 0{\cdot}541$$

Test Statistic. Since samples are large, the test statistic under the Null Hypothesis, H_0 is :

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{200 + 325}{400 + 600} = \frac{525}{1000} = 0{\cdot}525$$

$$\Rightarrow \quad \hat{Q} = 1 - \hat{P} = 1 - 0{\cdot}525 = 0{\cdot}475$$

$$\therefore \quad Z = \frac{0{\cdot}500 - 0{\cdot}541}{\sqrt{0{\cdot}525 \times 0{\cdot}475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}}$$

$$= \frac{-0{\cdot}041}{\sqrt{0{\cdot}525 \times 0{\cdot}475 \times (10/2{,}400)}}$$

$$= \frac{-0{\cdot}041}{\sqrt{0{\cdot}001039}} = \frac{-0{\cdot}041}{0{\cdot}0323} = -1{\cdot}269$$

Conclusion. Since $|Z| = 1{\cdot}269$ which is less than 1·96, it is not significant at 5% level of significance. Hence H_0 may be accepte.. at 5% level of significance and we may conclude that men and women do not differ significantly as regards proposal of flyover is concerned.

Example 12·7. *A company has the head office at Calcutta and a branch at Bombay. The personnel director wanted to know if the workers at the two places would like the introduction of a new plan of work and a survey was conducted for this purpose. Out of a sample of 500 workers at Calcutta, 62% favoured the new plan. At Bombay out of a sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitude towards the new plan at 5% level ?*

Solution. In the usual notations, we are given :

$$n_1 = 500,\ p_1 = 0{\cdot}62 \text{ and } n_2 = 400,\ p_2 = 1 - 0{\cdot}41 = 0{\cdot}59$$

Null hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference between the two groups in their attitude towards the new plan.

Alternative hypothesis, $H_1 : P_1 \neq P_2$ (Two-tailed).

Test Statistic. Under H_0, the test statistic for large samples is :

$$Z = \frac{p_1 - p_2}{\text{S.E.}\ (p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

where $$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500 \times 0{\cdot}62 + 400 \times 0{\cdot}59}{500 + 400} = 0{\cdot}607$$

and $$\hat{Q} = 1 - \hat{P} = 0{\cdot}393$$

$$\therefore \quad Z = \frac{0{\cdot}62 - 0{\cdot}59}{\sqrt{0{\cdot}607 \times 0{\cdot}393 \times \left(\frac{1}{500} + \frac{1}{400}\right)}}$$

$$= \frac{0{\cdot}03}{\sqrt{0{\cdot}00107}} = \frac{0{\cdot}03}{0{\cdot}0327} = 0{\cdot}917.$$

Critical region. At 5% level of significance, the critical value of Z for a two-tailed test is 1·96. Thus the critical region consists of all values of $Z \geq 1{\cdot}96$ or $Z \leq -1{\cdot}96$.

Conclusion. Since the calculated value of $|Z| = 0{\cdot}917$ is less than the critical value of Z (1·96), it is not significant at 5% level of significance. Hence the data do not provide us any evidence against the null hypothesis which may be accepted, and we conclude that there is no significant difference between the two groups in their attitude towards the new plan.

Example 12·8. *Before an increase in excise duty on tea, 800 persons out of a sample of 1,000 persons were found to be tea drinkers. After an increase in duty, 800 people were tea drinkers in a sample of 1,200 people. Using standard error of proportion, state whether there is a significant decrease in the consumption of tea after the increase in excise duty ?*

Solution. In the usual notations, we have $n_1 = 1{,}000$; $n_2 = 1{,}200$

p_1 = Sample proportion of tea drinkers before increase in excise duty

$$= \frac{800}{1000} = 0{\cdot}80$$

p_2 = Sample proportion of tea drinkers after increase in excise duty

$$= \frac{800}{1200} = 0{\cdot}67$$

Null Hypothesis, $H_0 : P_1 = P_2$, *i.e.*, there is no significant difference in the consumption of tea before and after the increase in excise duty.

Alternative Hypothesis, $H_1 : P_1 > P_2$ (Right-tailed alternative).

Test Statistic. Under the null hypothesis, the test statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \text{(Since samples are large)}$$

where

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = \frac{16}{22}, \text{ and } \hat{Q} = 1 - \hat{P} = \frac{6}{22}$$

$$\therefore \quad Z = \frac{0{\cdot}80 - 0{\cdot}67}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \left(\frac{1}{1000} + \frac{1}{1200}\right)}}$$

$$= \frac{0{\cdot}13}{\sqrt{\frac{16}{22} \times \frac{6}{22} \times \frac{11}{6000}}} = \frac{0{\cdot}13}{0{\cdot}019} = 6{\cdot}842$$

Conclusion. Since Z s much greater than 1·645 as well as 2·33 (since test is one-tailed), it is highly significant at both 5% and 1% levels of significance. Hence,

we reject the null hypothesis H_0 and conclude that there is a significant decease in the consumption of tea after increase in the excise duty.