CHAPTER 8

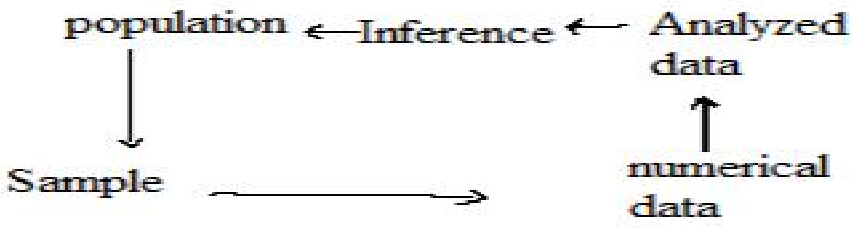
8. STATISTICAL INFERENCE ABOUT ONE POPULATION

8.1. Introduction

Statistical inference is the process of making interpretations or conclusions from sample data for the totality of the population. It is only the sample data that is ready for inference.

In statistics there are two ways though which inference can be made.

- Statistical estimation
- Statistical hypothesis testing.



Statistical estimation: This is one way of making inference about the population parameter where the investigator does not have any prior notion about values or characteristics of the population parameter.

It is the process by which sample data are used to indicate the value of an unknown quantity in a population.

There are two ways estimation.

i. **Point Estimation**: It is a procedure that results in a single value as an estimate for a parameter.

ii. **Interval estimation:** It is the procedure that results in the interval of values as an estimate for a parameter, which is interval that contains the likely values of a parameter. It deals with identifying the upper and lower limits of a parameter. The limits by themselves are random variable.

Estimator and Estimate

Estimator: is the rule or random variable that helps us to approximate a population parameter.

It is any quantity calculated from the data which is used to give information about an unknown quantity in a population.

Estimate: is the different possible value which an estimator can assume. It is an indication of the value of an unknown quantity based on observed data.

Or it is the particular value of an estimator that is obtained from a particular sample of data and used to indicate the value of a parameter.

Example: The sample mean $\overline{X} = \frac{\sum x_i}{n}$ is an estimator for the population mean and $\overline{X}$ =10 is an estimate, which is one of the possible value of $\overline{X}$.

Properties of best estimator

The following are some qualities of an estimator

- It should be unbiased.
- It should be consistent.
- It should be relatively efficient.
- Sufficiency

To explain these properties let $\widehat{\theta}$ be an estimator of θ

1. Unbiased Estimator: An estimator whose expected value is the value of the parameter being estimated. i.e. E($\widehat{\theta}$)= θ.
2. Consistent Estimator: An estimator which gets closer to the value of the parameter as the sample size increases. i.e. $\widehat{\theta}$ gets closer to θ as the sample size increases.
3. Relatively Efficient Estimator: The estimator for a parameter with the smallest variance. This actually compares two or more estimators for one parameter.

Sufficiency: an estimator that uses the entire information in estimating the parameter of our interest is called a sufficient estimator.

8.2 Point and interval estimation of the mean and proportion

Point and Interval estimation of the population mean: μ

☞ **Point Estimation of the population mean**

Another term for statistic is **point estimate**, since we are estimating the parameter value. A **point estimator** is the mathematical way we compute the point estimate. For instance, sum of x_i over n is the point estimator used to compute the estimate of the population means, μ .That is $\overline{X} = \frac{\sum x_i}{n}$ is a point estimator of the population mean.

☞ **Confidence interval estimation of the population mean**

Although $\overline{X}$ possesses nearly all the qualities of a good estimator, because of sampling error, we know that it's not likely that our sample statistic will be equal to the population parameter, but instead will fall into an interval of values. We will have to be satisfied knowing that the statistic is "close to" the parameter. That leads to the obvious question, what is "close"?

We can phrase the latter question differently: How confident can we be that the value of the statistic falls within a certain "distance" of the parameter? Or, what is the probability that the parameter's value is within a certain range of the statistic's value? This range is the confidence interval.

The **confidence level** is the probability that the value of the parameter falls within the range specified by the confidence interval surrounding the statistic.

- **There are different cases to be considered to construct confidence intervals.**

Case 1: If sample size is large or if the population is normal with known variance

Recall the Central Limit Theorem, which applies to the sampling distribution of the mean of a sample. Consider samples of size n drawn from a population, whose mean is μ and standard deviation is σ with replacement and order important. The population can have any frequency distribution. The sampling distribution of $\overline{X}$ will have a mean $\mu_{\bar{x}} = \mu$ and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, and approaches a normal distribution as n gets large. This allows us to use the normal distribution curve for computing confidence intervals.

$$\Rightarrow Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad has\, a\, normal\;\; distributi on\, with\, mean = 0\;\; and\; \mathrm{var}\, iance = 1$$

$$\Rightarrow \mu = \overline{X} \pm Z\,\sigma/\sqrt{n}$$

$$= \overline{X} \pm \varepsilon, \quad where\, \varepsilon\;\; is\, a\;\; measure\, of\;\; error.$$

$$\Rightarrow \varepsilon = Z\,\sigma/\sqrt{n}$$

- For the interval estimator to be good the error should be small. How it be small?
  - By making n large
  - Small variability
  - Taking Z small

To obtain the value of Z, we have to attach this to a theory of chance. That is, there is an area of size $1-\alpha$ such that

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$

$$Where\;\; \alpha = is\, the\, probability\, that\, the\;\; parameter\, lies\, outside\, the\, \mathrm{int}\, erval$$

$$Z_{\alpha/2} = s\tan ds\;\; for\, the\, s\tan dard\; normal\; \mathrm{var}\, iable\, to\, the\, right\; of\;\; which$$

$$\alpha/2\; probability\, lies, i.e\, P(Z > Z_{\alpha/2}) = \alpha/2$$

$$\Rightarrow P(-Z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}) = 1-\alpha$$

$$\Rightarrow P(\overline{X} - Z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \overline{X} + Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

$$\Rightarrow (\overline{X} - Z_{\alpha/2}\,\sigma/\sqrt{n},\ \ \overline{X} + Z_{\alpha/2}\,\sigma/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ \mathrm{int}erval\ \ for\ \mu$$

But usually σ^2 is not known, in that case we estimate by its point estimator S^2

$$\Rightarrow (\overline{X} - Z_{\alpha/2}\,S/\sqrt{n},\ \ \overline{X} + Z_{\alpha/2}\,S/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ \mathrm{int}erval\ \ for\ \mu$$

Here are the Z values corresponding to the most commonly used confidence levels.

$100(1-\alpha)$%	α	$\alpha/2$	$Z_{\alpha/2}$
90	**0.10**	**0.05**	**1.645**
95	**0.05**	**0.025**	**1.96**
99	**0.01**	**0.005**	**2.58**

<u>Case 2:</u> If sample size is small and the population variance, σ^2 is not known.

$$t = \frac{\overline{X}-\mu}{S/\sqrt{n}}\quad has\ t\ distribution\,with\ n-1\ \deg rees\,of\ freedom.$$

$$\Rightarrow (\overline{X} - t_{\alpha/2}\,S/\sqrt{n},\ \ \overline{X} + t_{\alpha/2}\,S/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ \mathrm{int}erval\ \ for\ \mu$$

The unit of measurement of the confidence interval is the standard error. This is just the standard deviation of the sampling distribution of the statistic.

Examples:

1. From a normal sample of size 25 a mean of 32 was found .Given that the population standard deviation is 4.2. Find

   a) A 95% confidence interval for the population mean.

   b) A 99% confidence interval for the population mean.

Solution:

$$\overline{X} = 32, \qquad \sigma = 4.2, \qquad 1-\alpha = 0.95 \Rightarrow \alpha = 0.05, \alpha/2 = 0.025$$

$$\Rightarrow Z_{\alpha/2} = 1.96 \quad from \quad table.$$

a)

$$\Rightarrow The\ \ required \operatorname{int} erval\ \ will\, be\, \overline{X} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$$

$$= 32 \pm 1.96 * 4.2/\sqrt{25}$$

$$= 32 \pm 1.65$$

$$= \underline{\underline{(30.35,\ 33.65)}}$$

b)

$$\overline{X} = 32, \qquad \sigma = 4.2, \qquad 1-\alpha = 0.99 \Rightarrow \alpha = 0.01,\ \alpha/2 = 0.005$$

$$\Rightarrow Z_{\alpha/2} = 2.58 \quad from \quad table.$$

$$\Rightarrow The\ \ required \operatorname{int} erval\ \ will\, be\, \overline{X} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$$

$$= 32 \pm 2.58 * 4.2/\sqrt{25}$$

$$= 32 \pm 2.17$$

$$= \underline{\underline{(29.83,\ 34.17)}}$$

2. A drug company is testing a new drug which is supposed to reduce blood pressure. From the six people who are used as subjects, it is found that the average drop in blood pressure is 2.28 points, with a standard deviation of .95 points. What is the 95% confidence interval for the mean change in pressure?

Solution:

$$\overline{X} = 2.28, \qquad S = 0.95, \qquad 1-\alpha = 0.95 \Rightarrow \alpha = 0.05,\ \alpha/2 = 0.025$$

$$\Rightarrow t_{\alpha/2} = 2.571\ \ with\ \ df = 5 \quad from\, table.$$

$$\Rightarrow The\ \ required \operatorname{int} erval\ \ will\, be\, \overline{X} \pm t_{\alpha/2}\, S/\sqrt{n}$$

$$= 2.28 \pm 2.571 * 0.95/\sqrt{6}$$

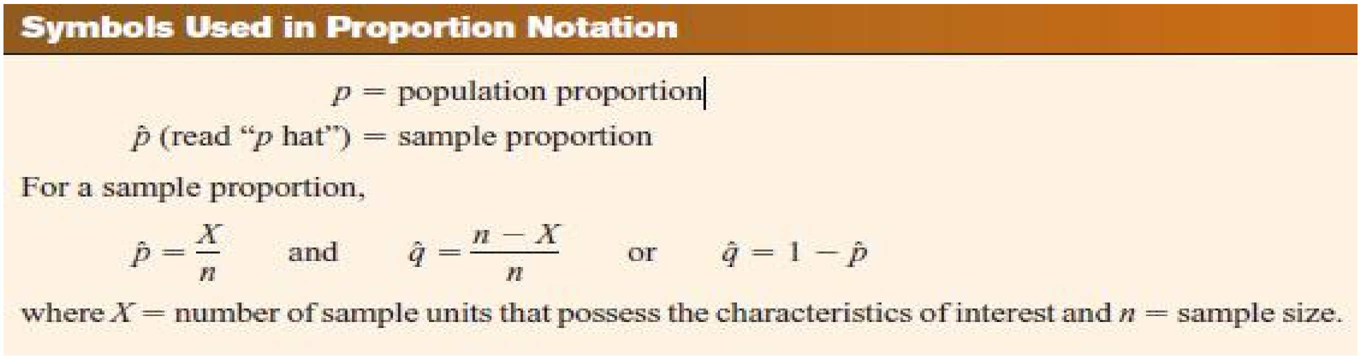
$$= 2.28 \pm 1.008$$

$$= \underline{\underline{(1.28,\ 3.28)}}$$

That is, we can be 95% confident that the mean decrease in blood pressure is between 1.28 and 3.28 points.

Point and Interval estimation of the population proportion P

- **Point Estimation of the population proportion**



Symbols Used in Proportion Notation

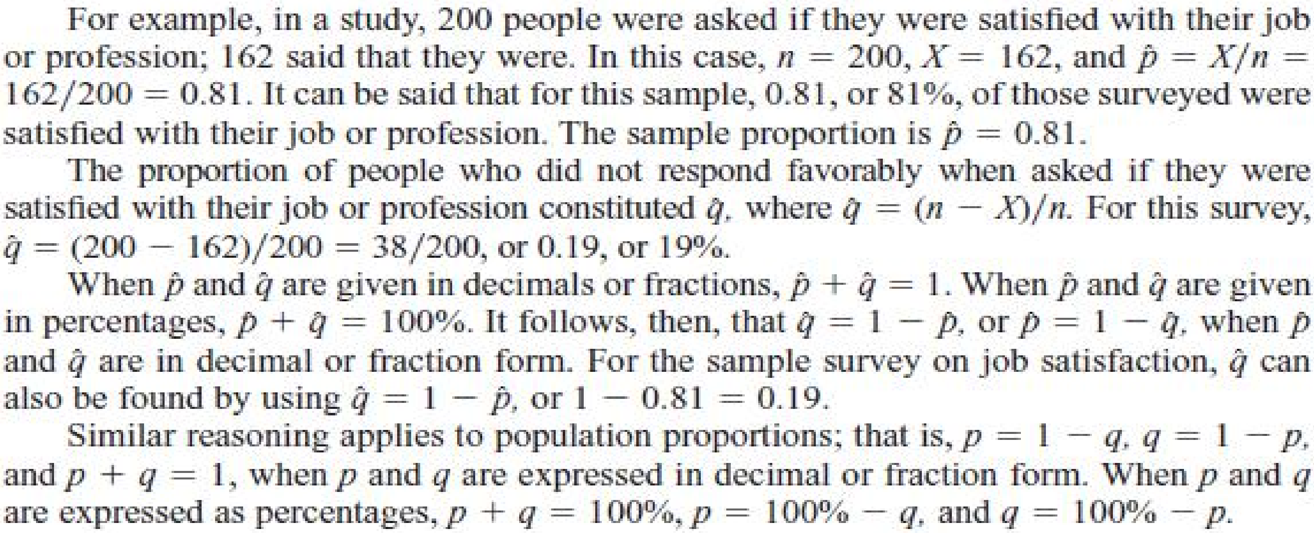
p = population proportion

$\hat{p}$ (read "p hat") = sample proportion

For a sample proportion,

$$\hat{p} = \frac{X}{n} \quad \text{and} \quad \hat{q} = \frac{n - X}{n} \quad \text{or} \quad \hat{q} = 1 - \hat{p}$$

where X = number of sample units that possess the characteristics of interest and n = sample size.

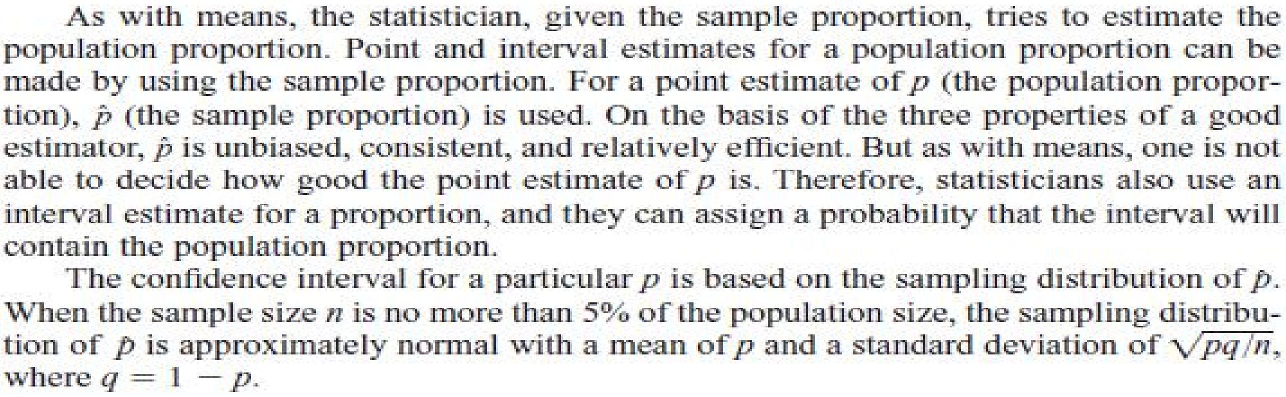


For example, in a study, 200 people were asked if they were satisfied with their job or profession; 162 said that they were. In this case, $n = 200$, $X = 162$, and $\hat{p} = X/n = 162/200 = 0.81$. It can be said that for this sample, 0.81, or 81%, of those surveyed were satisfied with their job or profession. The sample proportion is $\hat{p} = 0.81$.

The proportion of people who did not respond favorably when asked if they were satisfied with their job or profession constituted $\hat{q}$, where $\hat{q} = (n - X)/n$. For this survey, $\hat{q} = (200 - 162)/200 = 38/200$, or 0.19, or 19%.

When $\hat{p}$ and $\hat{q}$ are given in decimals or fractions, $\hat{p} + \hat{q} = 1$. When $\hat{p}$ and $\hat{q}$ are given in percentages, $\hat{p} + \hat{q} = 100\%$. It follows, then, that $\hat{q} = 1 - \hat{p}$, or $\hat{p} = 1 - \hat{q}$, when $\hat{p}$ and $\hat{q}$ are in decimal or fraction form. For the sample survey on job satisfaction, $\hat{q}$ can also be found by using $\hat{q} = 1 - \hat{p}$, or $1 - 0.81 = 0.19$.

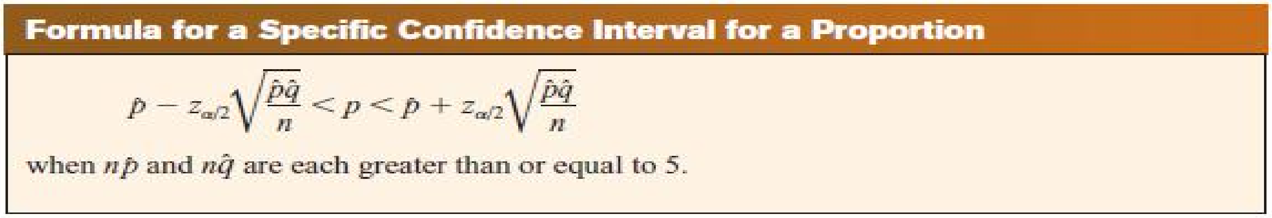
Similar reasoning applies to population proportions; that is, $p = 1 - q$, $q = 1 - p$, and $p + q = 1$, when p and q are expressed in decimal or fraction form. When p and q are expressed as percentages, $p + q = 100\%$, $p = 100\% - q$, and $q = 100\% - p$.



As with means, the statistician, given the sample proportion, tries to estimate the population proportion. Point and interval estimates for a population proportion can be made by using the sample proportion. For a point estimate of p (the population proportion), $\hat{p}$ (the sample proportion) is used. On the basis of the three properties of a good estimator, $\hat{p}$ is unbiased, consistent, and relatively efficient. But as with means, one is not able to decide how good the point estimate of p is. Therefore, statisticians also use an interval estimate for a proportion, and they can assign a probability that the interval will contain the population proportion.

The confidence interval for a particular p is based on the sampling distribution of $\hat{p}$. When the sample size n is no more than 5% of the population size, the sampling distribution of $\hat{p}$ is approximately normal with a mean of p and a standard deviation of $\sqrt{pq/n}$, where $q = 1 - p$.

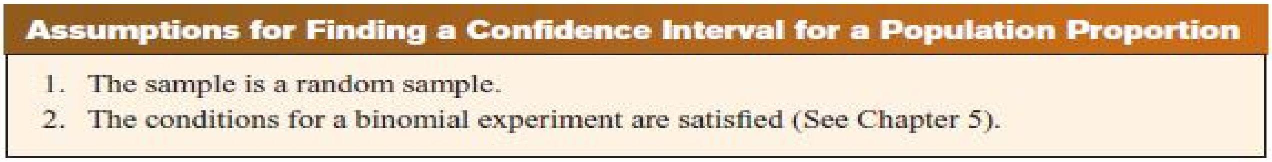
☞ **Confidence interval estimation of the population proportion**



Formula for a Specific Confidence Interval for a Proportion

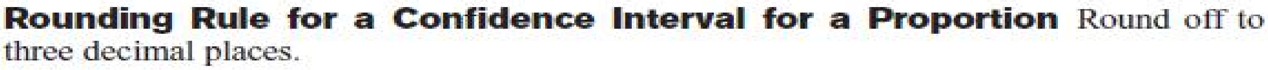
$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

when $n\hat{p}$ and $n\hat{q}$ are each greater than or equal to 5.

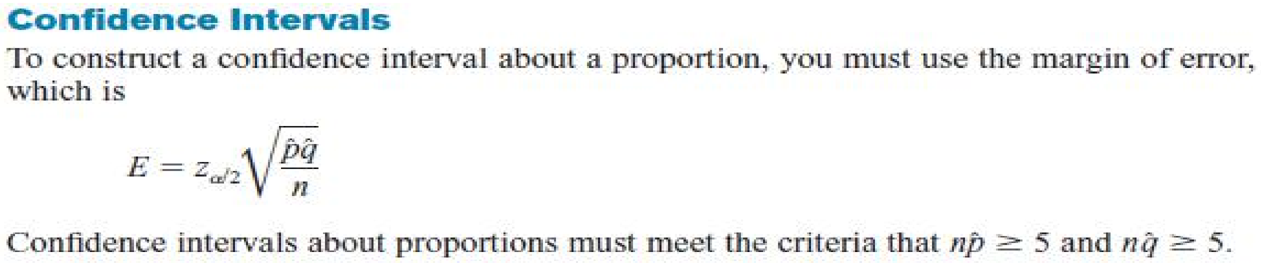


Assumptions for Finding a Confidence Interval for a Population Proportion

1. The sample is a random sample.
2. The conditions for a binomial experiment are satisfied (See Chapter 5).



Rounding Rule for a Confidence Interval for a Proportion Round off to three decimal places.



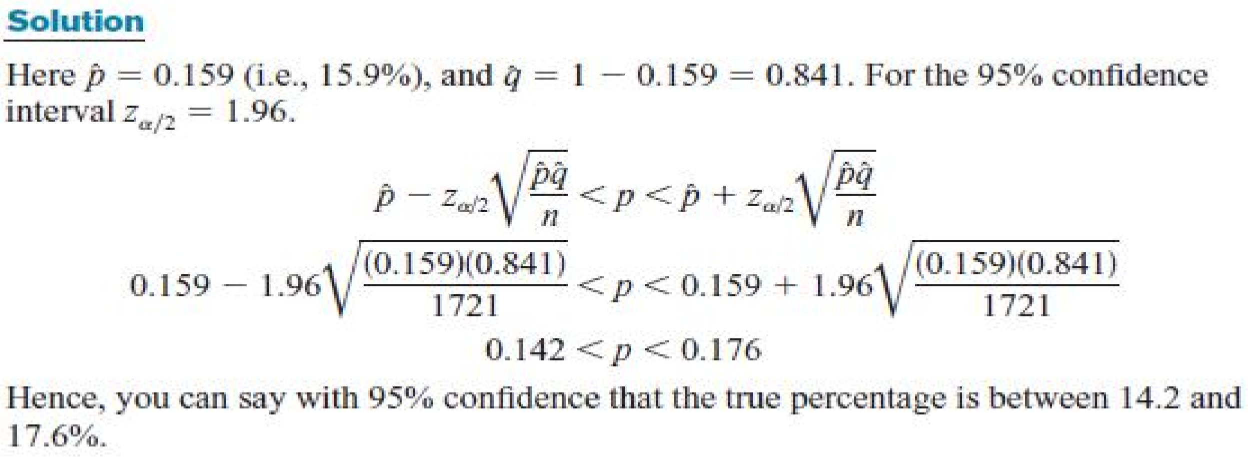
Confidence Intervals

To construct a confidence interval about a proportion, you must use the margin of error, which is

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence intervals about proportions must meet the criteria that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$.

Example: A survey of 1721 people found that 15.9% of individuals prefer a certain brand product. Find the 95% confidence interval of the true proportion of people who prefer that brand.



Solution

Here $\hat{p} = 0.159$ (i.e., 15.9%), and $\hat{q} = 1 - 0.159 = 0.841$. For the 95% confidence interval $z_{\alpha/2} = 1.96$.

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.159 - 1.96\sqrt{\frac{(0.159)(0.841)}{1721}} < p < 0.159 + 1.96\sqrt{\frac{(0.159)(0.841)}{1721}}$$

$$0.142 < p < 0.176$$

Hence, you can say with 95% confidence that the true percentage is between 14.2 and 17.6%.

8.3 Hypothesis testing about the mean and proportion

<u>Hypothesis Testing</u>

This is also one way of making inference about population parameter, where the investigator has prior notion about the value of the parameter.

Definitions:

- **Statistical hypothesis**: is an assertion or statement about the population whose plausibility is to be evaluated on the basis of the sample data.
- **Test statistic**: is a statistics whose value serves to determine whether to reject or accept the hypothesis to be tested. It is a random variable.
- **Statistic test**: is a test or procedure used to evaluate a statistical hypothesis and its value depends on sample data.

There are two types of hypothesis:

Null hypothesis:

- It is the hypothesis to be tested.
- It is the hypothesis of equality or the hypothesis of no difference.
- Usually denoted by H_0.

Alternative hypothesis:

- It is the hypothesis available when the null hypothesis has to be rejected.
- It is the hypothesis of difference.
- Usually denoted by H_1 or H_a.

Types and size of errors:

- Testing hypothesis is based on sample data which may involve sampling and non sampling errors.
- The following table gives a summary of possible results of any hypothesis test:

		Decision	
		Reject H_0	Don't reject H_0
Truth	H_0	Type I Error	Right Decision
	H_1	Right Decision	Type II Error

- **Type I error**: Rejecting the null hypothesis when it is true.
- **Type II error**: Failing to reject the null hypothesis when it is false.

NOTE:

1. There are errors that are prevalent in any two choice decision making problems.
2. There is always a possibility of committing one or the other errors.

3. Type I error (α) and type II error (β) have inverse relationship and therefore, can not be minimized at the same time.

- In practice we set α at some value and design a test that minimize β . This is because a type I error is often considered to be more serious, and therefore more important to avoid, than a type II error.

General steps in hypothesis testing:

1. Specify the null hypothesis (H_0) and the alternative hypothesis (H_1).
2. Specify the significance level, α
3. Identify the sampling distribution (if it is **Z** or **t**) of the estimator.
4. Identify the critical region.
5. Calculate a statistic analogous to the parameter specified by the null hypothesis.
6. Making decision.
7. Summarization of the result.

Hypothesis testing about the population mean, μ :

Suppose the assumed or hypothesized value of μ is denoted by μ_0 , then one can formulate two sided (1) and one sided (2 and 3) hypothesis as follows:

1. $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0$
2. $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu > \mu_0$
3. $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu < \mu_0$

Case 1: ***When sampling is from a normal distribution with σ^2 known***

- The relevant test statistic is

$$Z_{cal} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

- After specifying α we have the following regions (critical and acceptance) on the standard normal distribution corresponding to the above three hypothesis.

Summary table for decision rule:

H0	Reject H0 if	Accept H0 if	Inconclusive if
$\mu \neq \mu_0$	$\lvert Z_{cal}\rvert > Z_{\alpha/2}$	$\lvert Z_{cal}\rvert < Z_{\alpha/2}$	$Z_{cal} = Z_{\alpha/2}\, or\, Z_{cal} = -Z_{\alpha/2}$
$\mu < \mu_0$	$Z_{cal} < -Z_{\alpha}$	$Z_{cal} > -Z_{\alpha}$	$Z_{cal} = -Z_{\alpha}$
$\mu > \mu_0$	$Z_{cal} > Z_{\alpha}$	$Z_{cal} < Z_{\alpha}$	$Z_{cal} = Z_{\alpha}$

Case 2: ***When sampling is from a normal distribution with σ^2 unknown and small sample size***

- The relevant test statistic is

$$t_{cal} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t \;\; with\; n-1\; \text{deg}rees\; of\; freedom.$$

- After specifying α we have the following regions on the student t-distribution corresponding to the above three hypothesis.

H0	Reject H0 if	Accept H0 if	Inconclusive if
$\mu \neq \mu_0$	$\lvert t_{cal}\rvert > t_{\alpha/2}$	$\lvert t_{cal}\rvert < t_{\alpha/2}$	$t_{cal} = t_{\alpha/2}\, or\, t_{cal} = -t_{\alpha/2}$
$\mu < \mu_0$	$t_{cal} < -t_{\alpha}$	$t_{cal} > -t_{\alpha}$	$t_{cal} = -t_{\alpha}$
$\mu > \mu_0$	$t_{cal} > t_{\alpha}$	$t_{cal} < t_{\alpha}$	$t_{cal} = t_{\alpha}$

Case 3: ***When sampling is from a non- normally distributed population or a population whose functional form is unknown.***

- If a sample size is large one can perform a test hypothesis about the mean by using:

$$Z_{cal} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}, \quad if\;\; \sigma^2 \;\; is\;\; known.$$

$$= \frac{\overline{X} - \mu_0}{S/\sqrt{n}}, \;\; if\; \sigma^2\; is\;\; unknown.$$

- The decision rule is the same as **case I.**

Examples:

1. Test the hypotheses that the average height content of containers of certain lubricant is 10 liters if the contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 liters. Use the 0.01 level of significance and assume that the distribution of contents is normal.

Solution:

Let $\mu = Population\ mean.$, $\mu_0 = 10$

Step 1: Identify the appropriate hypothesis

$$H_0 : \mu = 10 \quad vs \quad H_1 : \mu \neq 10$$

Step 2: select the level of significance, $\alpha = 0.01(given)$

Step 3: Select an appropriate test statistics

t- Statistic is appropriate because population variance is not known and the sample size is also small.

Step 4: identify the critical region.

Here we have two critical regions since we have two tailed hypothesis

$$The\ critical\ region\ is\ |t_{cal}| > t_{0.005}(9) = 3.2498$$

$$\Rightarrow (-3.2498,\ 3.2498)\ \ is\ accep\tan ce\ \ region.$$

Step 5: Computations:

$$\overline{X} = 10.06,\ S = 0.25$$

$$\Rightarrow t_{cal} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{10.06 - 10}{0.25/\sqrt{10}} = 0.76$$

Step 6: Decision

Accept H$_0$, since $\mathbf{t_{cal}}$ is in the acceptance region.

Step 7: Conclusion

At 1% level of significance, we have no evidence to say that the average height content of containers of the given lubricant is different from 10 litters, based on the given sample data.

2. The mean life time of a sample of 16 fluorescent light bulbs produced by a company is computed to be 1570 hours. The population standard deviation is 120 hours. Suppose the hypothesized value for the population mean is 1600 hours. Can we conclude that the life time of light bulbs is decreasing?

(Use $\alpha = 0.05$ and assume the normality of the population)

Solution:

Let $\mu = Population\ mean.$, $\mu_0 = 1600$

Step 1: Identify the appropriate hypothesis

$$H_0 : \mu = 1600 \qquad vs \qquad H_1 : \mu < 1600$$

Step 2: select the level of significance, $\alpha = 0.05 (given)$

Step 3: Select an appropriate test statistics

Z- Statistic is appropriate because population variance is known.

Step 4: identify the critical region.

$$The\, critical\; region\;\; is\;\; Z_{cal} < -Z_{0.05} = -1.645$$

$$\Rightarrow (-1.645,\; \infty)\;\; is\, accep\tan ce\;\; region.$$

Step 5: Computations:

$$Z_{cal} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1570 - 1600}{120/\sqrt{16}} = -1.0$$

Step 6: **Decision**

Accept H_0, since Z_{cal} is in the acceptance region.

Step 7: **Conclusion**

At 5% level of significance, we have no evidence to say that that the life time of light bulbs is decreasing, based on the given sample data.

Hypothesis testing about the population proportion, P:

A hypothesis test involving a population proportion can be considered as a binomial experiment when there are only two outcomes and the probability of a success does not change from trial to trial.

Since a normal distribution can be used to approximate the binomial distribution when np ≥ 5 and nq ≥ 5, the standard normal distribution can be used to test hypotheses for proportions.

The estimate of P from a sample of size n is the sample proportion, $\hat{p}$ = x/n, where x is the number of successes in the sample. Using the normal approximation, the appropriate statistic to perform inferences on p is

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

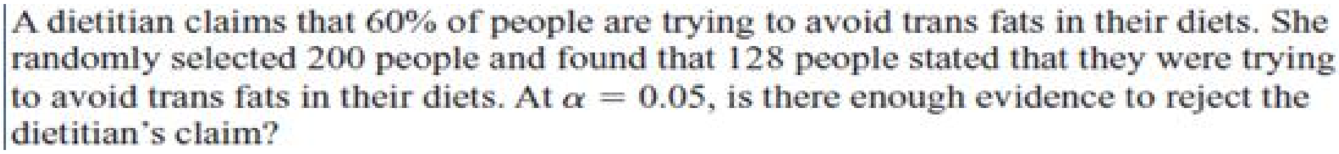
P = population proportion

n = sample size

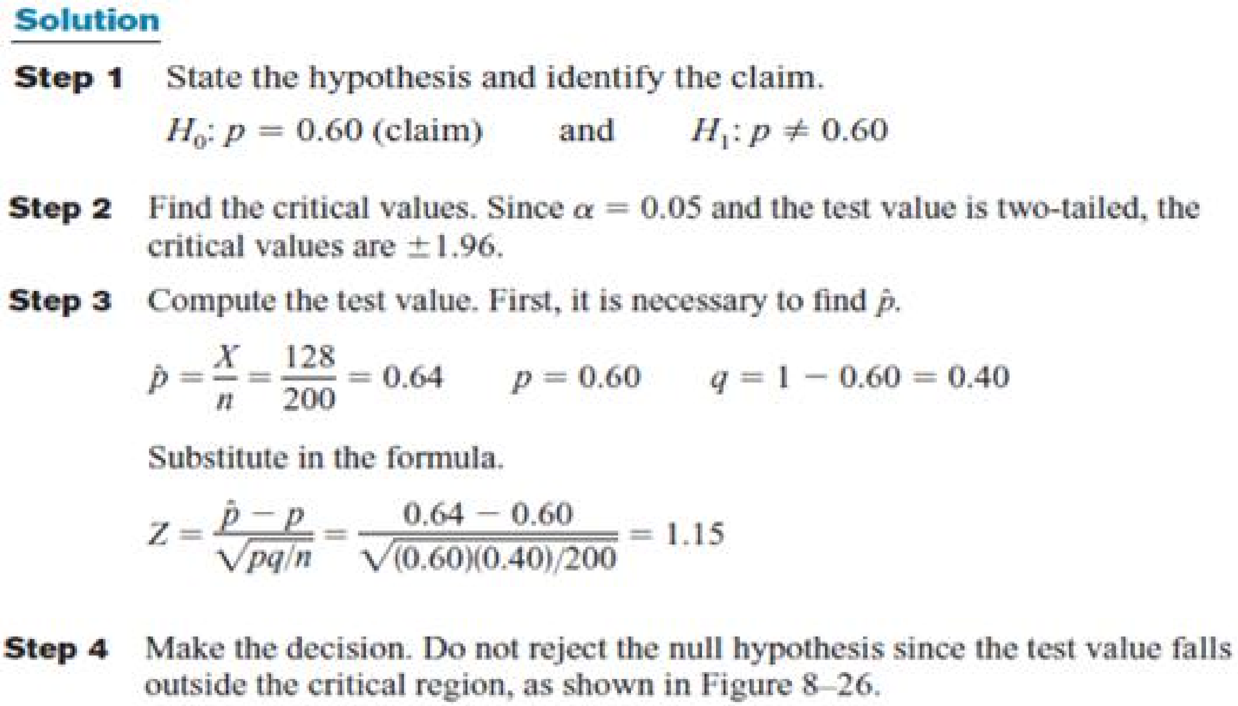
Procedures for test of hypothesis about a single population proportion (p)

One-Tailed Test		Two-Tailed Test
H_o: p = p_O		H_o: p = p_O
H_1: p > p_O or p < p_O		H_1: p ≠ p_O
Test statistic:	$z = \frac{\hat{p} - p_o}{\sqrt{p_o q_o / n}}$	
Rejection region:		Rejection region:
$z > z_\alpha$ or $z < -z_\alpha$		$\lvert z \rvert > z_{\alpha/2}$
Where $p(z > z_\alpha) = \alpha$	and	$p(z > z_{\alpha/2}) = \alpha/2$.

Example



A dietitian claims that 60% of people are trying to avoid trans fats in their diets. She randomly selected 200 people and found that 128 people stated that they were trying to avoid trans fats in their diets. At $\alpha = 0.05$, is there enough evidence to reject the dietitian's claim?



Solution

Step 1 State the hypothesis and identify the claim.

H_0: $p = 0.60$ (claim) and H_1: $p \neq 0.60$

Step 2 Find the critical values. Since $\alpha = 0.05$ and the test value is two-tailed, the critical values are ±1.96.

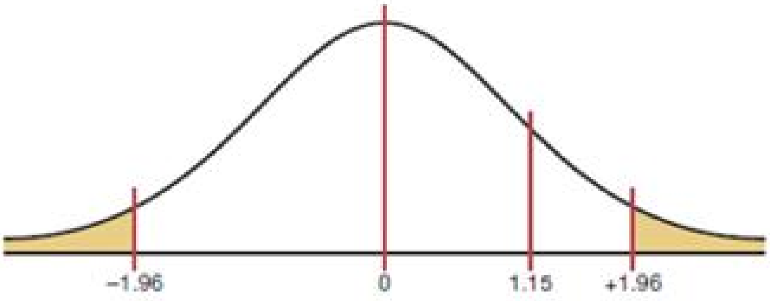
Step 3 Compute the test value. First, it is necessary to find $\hat{p}$.

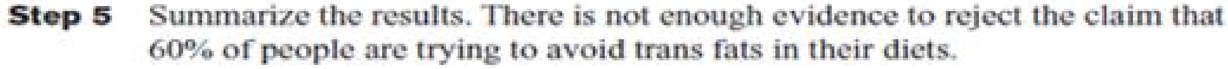
$$\hat{p} = \frac{X}{n} = \frac{128}{200} = 0.64 \qquad p = 0.60 \qquad q = 1 - 0.60 = 0.40$$

Substitute in the formula.

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0.64 - 0.60}{\sqrt{(0.60)(0.40)/200}} = 1.15$$

Step 4 Make the decision. Do not reject the null hypothesis since the test value falls outside the critical region, as shown in Figure 8–26.





Step 5 Summarize the results. There is not enough evidence to reject the claim that 60% of people are trying to avoid trans fats in their diets.

Example

A statistician claims that at most 77% of the population oppose replacing $1 bills with $1 coins. To see this claim is valid, the statistician selected a sample of 80 people and found that 55 were opposed to replacing the $ 1 bills. At α =0.01, test that at most 77% of the population are opposed to the change.

Solution

Step 1 State the hypotheses and identify the claim.

H_0: $p = 0.77$ (claim) and H_1: $p < 0.77$

Step 2 Find the critical value(s). Since $\alpha = 0.01$ and the test is left-tailed, the critical value is -2.33.

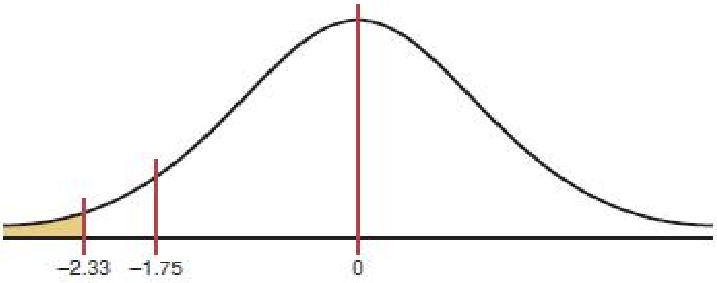
Step 3 Compute the test value.

$$\hat{p} = \frac{X}{n} = \frac{55}{80} = 0.6875$$

$$p = 0.77 \qquad \text{and} \qquad q = 1 - 0.77 = 0.23$$

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0.6875 - 0.77}{\sqrt{(0.77)(0.23)/80}} = -1.75$$

Step 4 Do not reject the null hypothesis, since the test value does not fall in the critical region



Step 5 There is not enough evidence to reject the claim that at least 77% of the population oppose replacing $1 bills with $1 coins.

8.4. Sample Size Determination

Sample size determination is closely related to statistical estimation. Quite often you ask, how large a sample is necessary to make an accurate estimate? The answer is not simple, since it depends on three things: the margin of error, the population standard deviation, and the degree of confidence.

The large the level of confidence selected, the large the sample size. A small allowable error will require a large sample. A large allowable error will permit a smaller sample. The maximum allowable error is one half the width of the corresponding CI. If the population is widely dispersed, large sample is required.

The formula for sample size is derived from the margin of error formula

$$E = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

and this formula is solved for n as follows:

$$E\sqrt{n} = z_{\alpha/2}(\sigma)$$

$$\sqrt{n} = \frac{z_{\alpha/2} \cdot \sigma}{E}$$

Hence, $$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Formula for the Minimum Sample Size Needed for an Interval Estimate of the Population Mean

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

where E is the margin of error. If necessary, round the answer up to obtain a whole number. That is, if there is any fraction or decimal portion in the answer, use the next whole number for sample size n.

Solution

Since $\alpha = 0.01$ (or $1 - 0.99$), $z_{\alpha/2} = 2.58$ and $E = 2$. Substituting in the formula,

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left[\frac{(2.58)(4.33)}{2}\right]^2 = 31.2$$

Round the value 31.2 up to 32. Therefore, to be 99% confident that the estimate is within 2 feet of the true mean depth, the scientist needs at least a sample of 32 measurements.

Depth of a River

A scientist wishes to estimate the average depth of a river. He wants to be 99% confident that the estimate is accurate within 2 feet. From a previous study, the standard deviation of the depths measured was 4.33 feet.

Solution

Since $\alpha = 0.01$ (or $1 - 0.99$), $z_{\alpha/2} = 2.58$ and $E = 2$. Substituting in the formula,

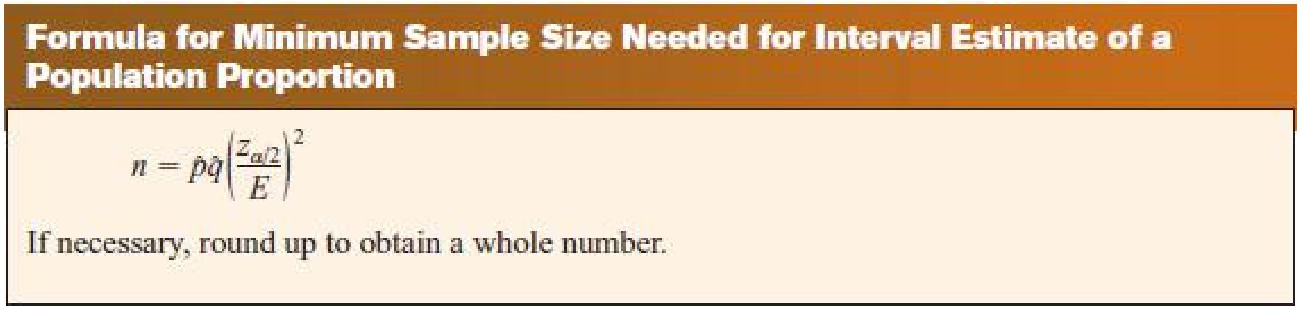
$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left[\frac{(2.58)(4.33)}{2}\right]^2 = 31.2$$

Round the value 31.2 up to 32. Therefore, to be 99% confident that the estimate is within 2 feet of the true mean depth, the scientist needs at least a sample of 32 measurements.

The formula for determining sample size requires the use of the population standard deviation. What happens when σ is unknown? In this case, an attempt is made to estimate σ. One such way is to use the standard deviation s obtained from a sample taken previously as an estimate for σ.

Sample Size for Proportions

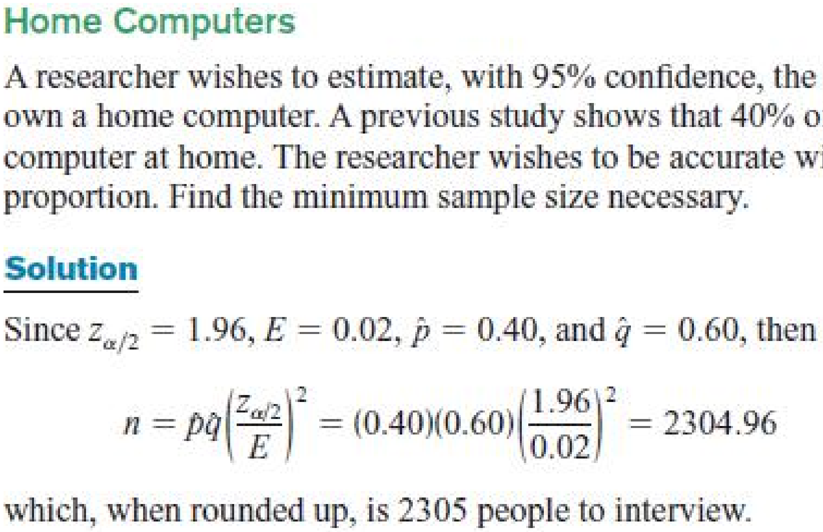
To find the sample size needed to determine a confidence interval about a proportion, use this formula:



Formula for Minimum Sample Size Needed for Interval Estimate of a Population Proportion

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$$

If necessary, round up to obtain a whole number.



Home Computers

A researcher wishes to estimate, with 95% confidence, the proportion of people who own a home computer. A previous study shows that 40% of those interviewed had a computer at home. The researcher wishes to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

Solution

Since $z_{\alpha/2} = 1.96$, $E = 0.02$, $\hat{p} = 0.40$, and $\hat{q} = 0.60$, then

$$n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2 = (0.40)(0.60)\left(\frac{1.96}{0.02}\right)^2 = 2304.96$$

which, when rounded up, is 2305 people to interview.

8.5. Tests of association

- Suppose we have a population consisting of observations having two attributes or qualitative characteristics say A and B.
- If the attributes are independent then the probability of possessing both A and B is P_A*P_B
  Where P_A is the probability that a number has attribute A.
  P_B is the probability that a number has attribute B.

- Suppose A has r mutually exclusive and exhaustive classes.
  B has c mutually exclusive and exhaustive classes

- The entire set of data can be represented using $r*c$ contingency table.

B								
A	B_1	B_2	.	.	B_j	.	B_c	**Total**
A_1	O_{11}	O_{12}			O_{1j}		O_{1c}	R_1
A_2	O_{21}	O_{22}			O_{2j}		O_{2c}	R_2
.								
.								
A_i	O_{i1}	O_{i2}			O_{ij}		O_{ic}	R_i
.								
.								
A_r	O_{r1}	O_{r2}			O_{rj}		O_{rc}	
Total	C_1	C_2			C_j			**n**

- The chi-square procedure test is used to test the hypothesis of independency of two attributes .For instance we may be interested
    - Whether the presence or absence of hypertension is independent of smoking habit or not.
    - Whether the size of the family is independent of the level of education attained by the mothers.
    - Whether there is association between father and son regarding boldness.
    - Whether there is association between stability of marriage and period of acquaintance ship prior to marriage.
- The χ^2 statistic is given by:

$$\chi^2{}_{cal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right] \sim \chi^2{}_{(r-1)(c-1)}$$

$$\begin{aligned} Where\quad O_{ij} &= the\,number\,of\;units\,that\,belong\,to\,category\,i\;of\;\,A\,and\;\,j\;of\;B.\\ e_{ij} &= Expected\;frequency\,that\,belong\,to\,category\,i\;of\;\,A\,and\;\,j\;of\;B.\end{aligned}$$

- The e_{ij} is given by :

$$e_{ij} = \frac{R_i * C_j}{n} \qquad \begin{aligned} Where\quad R_i &= the\;\,i^{th}\,row\,total.\\ C_j &= the\;j^{th}\;column\,total.\\ n &= total\;number\;of\;oservations\end{aligned}$$

Remark: $n = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} e_{ij}$

- The null and alternative hypothesis may be stated as:

$$H_0 : There\,is\,no\,association\,between\;A\,and\;B.$$
$$H_1 : not\;\;H_0(There\,is\,association\,between\;A\,and\;B).$$

Decision Rule: Reject H0 for independency at α level of significance if the calculated value of χ^2 exceeds the tabulated value with degree of freedom equal to $(r-1)(c-1)$.

$$\Rightarrow \text{Reject}\, H_0\;if\;\;\chi^2{}_{cal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right] > \chi^2{}_{(r-1)(c-1)}\;\;at\;\alpha$$

Examples:

1.A geneticist took a random sample of 300 men to study whether there is association between father and son regarding boldness. He obtained the following results.

Son		
Father	Bold	Not
Bold	85	59
Not	65	91

Using $\alpha = 5\%$, test whether there is association between father and son regarding boldness.

Solution:

$$H_0: \textit{There is no association between Father and Son regarding boldness.}$$

$$H_1: not\ H_0$$

- First calculate the row and column totals

$$R_1 = 144,\quad R_2 = 156,\quad C_1 = 150,\; C_2 = 150$$

- Then calculate the expected frequencies(e_{ij}'s)

$$e_{ij} = \frac{R_i * C_j}{n}$$

$$\Rightarrow e_{11} = \frac{R_1 * C_1}{n} = \frac{144*150}{300} = 72 \qquad e_{12} = \frac{R_1 * C_2}{n} = \frac{144*150}{300} = 72$$

$$e_{21} = \frac{R_2 * C_1}{n} = \frac{156*150}{300} = 78 \qquad e_{22} = \frac{R_2 * C_2}{n} = \frac{156*150}{300} = 78$$

- Obtain the calculated value of the chi-square.

$$\chi^2_{cal} = \sum_{i=1}^{2}\sum_{j=1}^{2}\left[\frac{(O_{ij} - e_{ij})^2}{e_{ij}}\right]$$

$$= \frac{(85-72)^2}{72} + \frac{(59-72)^2}{72} + \frac{(65-78)^2}{78} + \frac{(91-78)^2}{78} = 9.028$$

- Obtain the tabulated value of chi-square

$$\alpha = 0.05$$

$$\textit{Degrees of freedom} = (r-1)(c-1) = 1*1 = 1$$

$$\chi^2_{0.05}(1) = 3.841\ \textit{from table.}$$

- The decision is to reject H_0 since $\chi^2_{cal} > \chi^2_{0.05}(1)$

__Conclusion:__ At 5% level of significance we have evidence to say there is association between father and son regarding boldness, based on this sample data.

2. Random samples of 200 men, all retired were classified according to education and number of children is as shown below

Education level	***Number of children***		
	0-1	*2-3*	*Over 3*
Elementary	14	37	32
Secondary and above	31	59	27

Test the hypothesis that the size of the family is independent of the level of education attained by fathers. (Use 5% level of significance)

Solution:

$$H_0 : There\,is\,no\ \ association\,between\ \ the\,size\,of\ the\ family\,and\,the\,level\,of$$
$$education\,attained\,by\ fathers.$$
$$H_1 : not\ \ H_0.$$

- First calculate the row and column totals

$$R_1 = 83,\ \ R_2 = 117,\ \ \ C_1 = 45,\ C_2 = 96, C_3 = 59$$

- Then calculate the expected frequencies(**e**

$$e_{ij} = \frac{R_i * C_j}{n} \quad \Rightarrow e_{11} = 18.675,\ e_{12} = 39.84,\ e_{13} = 24.485$$
$$e_{21} = 26.325,\ e_{22} = 56.16,\ e_{23} = 34.515$$

- Obtain the calculated value of the chi-square.

$$\chi^2{}_{cal} = \sum_{i=1}^{2}\sum_{j=1}^{3}\left[\frac{(O_{ij} - e_{ij})^2}{e_{ij}}\right]$$
$$= \frac{(14-18.675)^2}{18.675} + \frac{(37-39.84)^2}{39.84} + ... + \frac{(27-34.515)^2}{34.515} = 6.3$$

- Obtain the tabulated value of chi-square

  $$\alpha = 0.05$$

  $$Degrees\ of\ \ freedom = (r-1)(c-1) = 1*2 = 2$$

  $$\chi^2_{0.05}(2) = 5.99\ \ from\ \ table.$$

- The decision is to reject H$_0$ since $\chi^2{}_{cal} > \chi^2_{0.05}(2)$

Conclusion: At 5% level of significance we have evidence to say there is association between the size of the family and the level of education attained by fathers, based on this sample data.