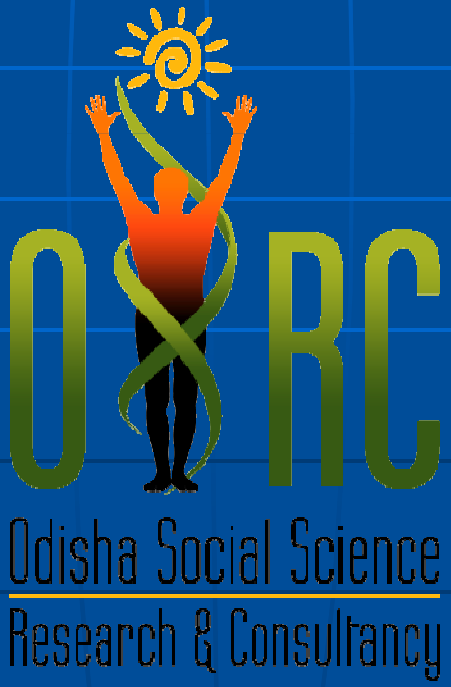
The chi-square distribution and the analysis of frequencies



CONTENTS

- Introduction
- The mathematical properties of the chi-square distribution
- Tests of Goodness-of-fit
- Test of Independence
- Tests of Homogeneity
- The fisher Exact test
- Relative risk, Odds ratio

Introduction

- The chi-square distribution is the most frequently employed statistical technique for the analysis of count or frequency data.
- Ex: a sample of hospitalized patients how many are male and how many are female.
- For the same sample we may also know how many have private insurance coverage, how many have medical insurance and how many are on medical assistance.

- The chi square distribution can be derived from normal distribution.
- Suppose that from a normally distributed random variable Y with mean μ and variance σ^2, we randomly and independently select samples of size n=1.
- By applying standard normal z transformation

$$z = \frac{y_i - \mu}{\sigma}$$

- $$z^2 = \chi^2_{(1)} = \left(\frac{y-\mu}{\sigma}\right)^2 = z^2$$ follows a χ^2 distribution with 1 d.f.

- If the sample size is ‘n’ then

$$\chi^2_{(n)} = z_1^2 + z_2^2 + \ldots\ldots + z_n^2$$ follows a χ^2 with n d.f.

- The mathematical form of the chi-square distribution is

$$f(\chi) = \frac{1}{\left(\frac{k}{2}-1\right)!}\frac{1}{2^{k/2}}\chi^{(k/2)-1}e^{-(\chi/2)}, u > 0$$

where e=the irrational number 2.71828.. And k is the number of degrees of freedom.

- The mean and variance of the chi-square distribution are k and 2k, respectively. The model value of the distribution is k-2 for values of k greater then or equal to 2 and is zero for k=1.

- **The Chi-square Test Statistic:** The test statistic for the chi-square tests is

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

When the null hypothesis is true, χ^2 is distributed approximately as χ^2 with k-r degrees of freedom.

- k is equal to the number of groups for which observed and expected frequencies are available and r is the number of restrictions or constraints imposed in the given comparison.
- In equation O_i is the observed frequency for the i^{th} category of variable of interest and E_i is the expected frequency for the i^{th} category.

- χ^2 is such that when there is close agreement between observed and expected frequencies it is small and when the agreement is poor it is large.
- Only a sufficiently large value of χ^2 will cause rejection of the null hypothesis.

The Decision Rule:

- The quantity $\Sigma[(O_i-E_i)^2/E_i]$ will be small if the observed and expected frequencies are close together and will be large if the differences are large.
- The decision rule is reject H_0 if χ^2 is greater than or equal to the tabulated χ^2 for the chosen value of α.

TESTS OF GOODNESS-OF-FIT

- A goodness-of-fit is appropriate when we want to decide whether an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution such as binomial, poisson, normal or any other distribution.

Example:

A research team making a study of hospitals in the United States collects data on a sample of 250 hospitals. The team computes for each hospital the inpatient occupancy ratio, a variable that shows, for a 12-month period, the ratio of average daily census to the average number of bed maintained. The sample yielded the distribution of ratios in Table.

We wish to know whether these data provide sufficient evidence to indicate that the sample did not come from a normally distributed population.

TABLE: Results of Study described in Example

Inpatient Occupancy ratio (%)	Number of hospital
0.0-39.9	16
40.0-49.9	18
50.0-59.9	22
60.0-69.9	51
70.0-79.9	62
80.0-89.9	55
90.0-99.9	22
100.0-109.9	4
Total	250

Solution:

Assumption: we assume that the sample available for analysis is a simple random sample.

Hypothesis:

H_0: In the population from which the sample was drawn, inpatient occupancy ratios are normally distributed.

H_A: The sampled population is not normally distributed.

Test statistic: The test statistic is

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

Distribution of test statistic: If H_0 is true the test statistic is distributed approximately as chi-square with k-r degrees of freedom. The values of k and r will be determined later.

Decision rule: We will reject H_0 if the computed value of χ^2 is equal to or greater than the critical value of chi-square.

Calculation of test statistic: The mean and standard deviation computed from the grouped data of Table are

$$\bar{x} = 69.91$$

$$s = 19.02$$

To obtain the relative frequency of occurrence of values in the interval 40.0 to 49.9 is

The z value corresponding to X=40.0 is

$$z = \frac{40.0 - 69.91}{19.02} = -1.57$$

The z value corresponding to X=50.0 is

$$z = \frac{50.0 - 69.91}{19.02} = -1.05$$

In Table D the area to the left of -1.05 is 0.1469, and the area to the left of -1.57 is 0.0582. the area between -1.05 and -1.57 is equal to 0.1469-0.0582=0.0887, which is equal to tha expected relative frequency of occurrence of values of occupancy ratio within the interval 40.0-49.9.

Class interval	Class interval	Z=(xi-x)/s	Expec R.f	Exp frequ NxR.f
<40	0-40	<1.57	0.0582	14.55
40.0-49.9	40-50	-1.57-1.05	0.0887	22.18
50.0-59.9	50-60	-1.05-0.52	0.1546	38.65
60.0-69.9	60-70	-0.52-.00	0.1985	49.62
70.0-79.9	70-80	0.00-0.53	0.2019	50.48
80.0-89.9	80-90	0.53-1.06	0.1535	38.38
90.0-99.9	90-100	1.06-1.58	0.0875	21.88
100.0-109.9	100-110	1.58-2.11	0.0397	9.92
>110.0	>=110	2.11-∞	0.0174	4.35
Total			1.0000	250.00

Class interval	Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i-E_i)^2/E_i$
<40	16	14.55	0.145
40.0-49.9	18	22.18	0.788
50.0-59.9	22	38.65	7.173
60.0-69.9	51	49.62	0.038
70.0-79.9	62	50.48	2.629
80.0-89.9	55	38.38	7.197
90.0-99.9	22	21.88	0.001
100.0-109.9	4	9.92	3.533
110.0 and greater	0	4.35	4.350
Total	250	250.00	25.854

$\chi^2 = \Sigma[(O_i-E_i)^2/E_i] = 25.854$. D.F 9(the number of groups or class intervals)-3(for the three restrictions: making $\Sigma E_i = \Sigma O_i$, and estimating μ and σ from the sample data)= 6

Statistical decision: χ^2 computed χ^2 =25.854 > tabulated $\chi^2_{.995}$=18.548, Hence H_0 is rejected 0.005 level of significance.

Conclusion: we conclude that in the sampled population, inpatient occupancy ratios are not normally distributed.

p value: Since 25.854> 18.548, p>0.005.

When any expected all frequency < 1, this is pooled with the adjacent cell 1 d.f is lost in the process.

TEST OF INDEPENDENCE

- It tests the null hypothesis that two criteria of classification of data, when applied to the same set of entities, are independent.
- Two criteria of classification are independent if the distribution of one criterion is the same no matter what the distribution of the other criterion.

The Contingency Table

- A contingency table is an array of ‘r’ rows and ‘k’ columns, where ‘r’ rows represent the various levels of one criterion of classification and the ‘k’ columns represent the various levels of the second criterion. The number in a cell of the array represents the frequency corresponding to the levels of the row and col.

Table: Two-way classification of a finite population of entities

Second criterion of classification Level	First criterion of classification level				
	1	2	3.....k		Total
1	N_{11}	N_{12}	N_{13}....	N_{1c}	$N_{1.}$
2	N_{21}	N_{22}	N_{23}	N_{2c}	$N_{2.}$
3.	N_{31}	N_{32}	N_{33}	N_{3c}	$N_{3.}$
.					
.					
r	N_{r1}	N_{r2}	N_{r3}	N_{rc}	$N_{r.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	$N_{.c}$	N

Table: Two-way classification of sample of entities

Second criterion of classification Level	First criterion of classification level				
	1	2	3.....	C	Total
1	n_{11}	n_{12}	n_{13}....	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	n_{2c}	$n_{2.}$
3.	n_{31}	n_{32}	n_{33}	n_{3c}	$n_{3.}$
.					
.					
r	n_{r1}	n_{r2}	n_{r3}	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.c}$	n

Example:

The purpose of a study by Vermund et al.(A-2) was to investigate the hypothesis that HIV-infected women who are also infected with human papillomavirus (HPV). detected by molecular hybridization, are more likely to have cervical cytologic abnormalities

than are women with only one or neither virus. The data shown in Table were reported by the investigators. We wish to know if we may conclude that there is a relationship between HPV status and stage of HIV infection.

Solution:

Assumption: We assume that the sample available for analysis is equivalent to a simple random sample drawn from the population of interest.

Hypothesis: H_0: HPV status and stage of HIV infection are independent.

H_A: The two variables are not independent.

Let $\alpha=0.05$

Test statistic: The test statistic is

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

Distribution of test statistic: When H0 is true $\chi 2$ with (r-1)(c-1)=(2-1)(3-1)=(1)(2)=2 degrees of freedom.

Decision rule: Reject H0 if the compound value of $\chi 2$ is equal to or greater than 5.991.

Table: HPV status and a stage of HIV infection among 96women.

	HIV			
HPV	Sero +Ve symptomatic	Sero+Ve asymptomatic	Sero -Ve	Total
+Ve	23	4	10	37
-Ve	10	14	35	59
Total	33	18	45	96

Table: Observed and Expected Frequencies

	HIV			
HPV	Sero +Ve symptomatic	Sero+Ve asymptomatic	Sero -Ve	Total
+Ve	23(12.72)	4(6.94)	10(17.34)	37
-Ve	10(20.28)	14(11.06)	35(27.66)	59
Total	33	18	45	96

Calculation of test statistic:

When two events are independent, probability of their joint occurrence is the product of their individual probability. Accordingly,

the expected frequency for the first cell is $(33/96)\times(37/96)) \times 96=17.72$. The other expected frequencies are calculated in a similar manner from the observed and expected value-

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

$$= \frac{(23-12.72)^2}{12.72} + \frac{(4-6.94)^2}{6.94} + \ldots\ldots + \frac{(35-27.66)^2}{27.66}$$

$$= 8.30805 + 1.24548 + \ldots + 1.94778$$

$$= 20.60081$$

Statistical decision: we reject H_0 since 20.60081>5.991

Conclusion: we conclude that H_0 is false, and that there is a relationship between HPV status and stage of HIV infection.

p value: Since 20.60081 is greater than 10.597, p<0.005.

Yates's Correction:

- The observed frequencies in a contingency Table are discrete and thereby give rise to a discrete statistic, X^2 , which is approximated by the χ^2 distribution, which is continuous.
- Yates proposed a procedure for correcting for this in the case of 2×2 tables.
- The correction consists of subtracting half the total number of observations from the absolute value of the quantity ad- bc before squaring.

$$X^2_{corrected} = \frac{n\left(|ad-bc|-.5n\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$

- It is generally agreed that no correction is necessary for larger contingency table.

Test of Independence-characteristics:

The characteristics of a chi-square test of independence that distinguish it from other chi-square tests are:

1. A single sample is selected from a population of interest and the subjects or objects are cross-classified on the basis of two variables of interest.
2. The rationale for calculating expected cell frequencies is based on the probability law, which states that if two events are independent, the probability of their joint occurrence is equal to the product of their individual probabilities.
3. The hypothesis and conclusions are stated in terms of the independence of two variables.

Small Expected Frequency

Cochran rule: For contingency table with > 1 d.f, a minimum expected frequency of 1 is allowable if no more than 20% of the cells have expected frequencies<5

TEST OF HOMOGENEITY

Under Chi square test of Independence the sample is assumed to have been drawn from a single population. The observed number of entities are then classified into two criteria of classification.

On occasion, however, either row or column totals may be under the control of the investigator; that is, the investigator may specify that independent samples be drawn from each of several populations. In this case one set of marginal totals is said to be fixed, while the other set, corresponding to the criterion of classification applied to the sample is random. This lead to Chi square test of homogeneity

Example:

Kodama et al.(A-8) studied the relationship between age and several prognostic factors in squamous cell carcinoma of the cervix. Among the data collected were the frequencies of histologic cell types in four age groups. The results are shown in Table. We wish to know if we may conclude that the populations represented by the four age-group samples are not homogeneous with respect to cell type.

	Cell type			
Age group (years)	No.of Patients	Large cell Nonkeratinizing Cell type	Keratinizing Cell type	Small cell Nonkeratinizing cell type
30-39	34	18	7	9
40-49	97	56	29	12
50-59	144	83	38	23
60-69	105	62	25	18
Total	380	219	99	62

Solution:

Assumption: We assume that we have a simple random sample from each one of the four populations of interest.

Hypothesis: H_0: The four populations are homogeneous with respect to cell type.

H_A: the four populations are not homogeneous with respect to cell type.

Let $\alpha=0.05$.

Test statistic: The test statistic is $X^2=\Sigma[(O_i-E_i)^2/E_i]$.

Distribution of test statistic: If H_0 is true X^2 is distributed approximately as χ^2 with $(4-1)(3-1)=(3)(2)=6$ degrees of freedom.

Decision rule: Reject H_0 if the computed value of X^2 is equal to or greater than 12.592.

Calculation of test statistic:

Statistical decision: Since 4.444 is less than the critical value of 12.592, we are unable to reject the null hypothesis.

Conclusion: We conclude that the four populations may be homogeneous with respect to cell type.

p value: Since 4.444 is less than 10.645, $p>0.10$.

The chi-square test homogeneity has the following characteristics:

1. Two or more populations are identified in advance and an independent sample is drawn from each.
2. Sample subjects or objects are placed in appropriate categories of the variable of interest.

3. The calculation of expected cell frequencies is based on the rationale that if the populations are homogeneous as stated in the null hypothesis , the best estimate of the probability that a subject or object will fall into a particular category of the variable of interest can be obtained by pooling the sample data.

4. The hypothesis and conclusions are stated in terms of homogeneity (with respect to the variable of interest) of populations.

Test of Homogeneity and $H_0: p_1 = p_2$:

- The chi-square test homogeneity for the two-sample case provides an alternative method for testing the null hypothesis that two population proportions are equal.

- To test $H_0: p_1 = p_2$ against $HA: p_1 \neq p_2$, by means of the statistic

$$z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - \left(p_1 - p_2\right)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1}} + \sqrt{\frac{\bar{p}(1-\bar{p})}{n_2}}}$$

Where $\bar{p}$ is obtained by pooling the data of the two independent samples available for analysis.

THE FISHER EXACT TEST

Fisher exact test may be used when the size requirements of the chi-square test are not met.

Data Arrangement:

Arrange data in the form of a 2×2 contingency table.

Arrange the frequencies such that A > B and choose the characteristic of interest such that a/ A > b/ B

Table: A 2×2 Contingency Table for the Fisher exact test.

Sample	With Characteristic	Without Characteristic	Total
1	a	A-a	A
2	b	B-b	B
Total	a+b	A+B-a-b	A+B

Assumptions: The assumptions for Fisher exact test are-

1. The data consists of two independent and random samples from population 1 & 2 with obsn. A & B respectively.
2. Each observation can be categorized as one of two mutually exclusive.

Hypothesis: The null hypothesis that may be tested and their alternatives.

1.(Two-sided)

H_0: The proportion with the characteristic of interest is the same in both population, that is. $P_1=P_2$.

HA: the proportion with the characteristic of interest is not the same in both populations; $P_1 \neq P_2$.

2.(One-sided)

H_0: The proportion with the characteristic of interest in population 1 is less than or the same as the proportion in population 2; $P_1 \leq P_2$.

H_A: the proportion with the characteristic of interest is greater in population 1 than in population 2; $P_1 > P_2$.

Test statistic: The test statistic is b, the number in sample 2 with the characteristic of interest.

Decision rule: The specific decision rules are

1. Two-sided Test: Enter Table J with A, B and a. If the observed value of b is equal to or less than the integer in a given column, reject H_0 at a level of significance equal to twice the significance level shown at the top of that column.
2. One-sided Test: Enter Table J with A, B and a. If the observed value of b is equal to or less than the integer in a given column, reject H_0, at a level of significance shown at the top of that column.

Approximation:

- For sufficiently large samples test the null hypothesis of the equality of two population proportions by normal approximation.

$$z = \frac{(a/A) - (b/B)}{\sqrt{\hat{p}(1-\hat{p})(1/A + 1/B)}}$$

$$where\ \hat{p} = (a+b)/(A+B)$$

Example:

The purpose of a study by Crozier et al.(A-12) was to document that patients with motor complete injury, but preserved pin appreciation, in additio to light touch, below the zone of injury have better prognoses with

Regard to ambulation than patients with only light touch preserved . Subjects were 27 patients with upper motor neuron lesions admitted for treatment within 72 hours of injury. They were divided into two groups. Group 1 consisted of patients who had touch sensation but no pin appreciation below the zone of injury. Group 2 consisted of patients who had partial or complete pin appreciation and light touch sensation below the zone of injury. Table shows the ambulatory status of these patients at time of discharge. We wish to know if we may conclude that patients classified as group 2 have a higher probability of ambulation at discharge than patients classified as group 1.

Table 1: Ambulatory Status at Discharge of Group 1 and Group 2 patients Described

	Ambulatory status		
Group	Total	Nonambulatory	Ambulatory
1	18	16	2
2	9	1	8
Total	27	17	10

Solution:

Assumption: We presume that the assumptions for application of the Fisher exact test are met.

Table 2: Data of Table 1 rearranged to confirm to the Layout

	Ambulatory status		
Group	Total	Nonambulatory	Ambulatory
1	18=A	16=a	2=A-a
2	9=B	1=b	8=B-b
Total	27=A+B	17=a+b	10=A+B-a-b

Hypothesis:

H_0: the rate of ambulation at discharge in a population of patients classified as group 2 is the same as or less than the rate of ambulation of discharge in a population of patients classified as group 1.

H_A: Group 2 patients have a higher rate of ambulation at discharge than group 1 patients.

Test statistic: The test statistic is the observed value of b as shown in Table 2.

Distribution of test statistic: We determine the significance of b by consulting Table J.

Decision rule: Suppose let α=0.01. the decision rule, then is to reject H_0 if the observed value of b is equal to or less than 3, the value of b in Table J for A=18, B=9, a=16 and α=0.01

Calculation of test statistic: The observed value of b is 1.

Statistical decision: Since 1<3, we reject H_0.

Conclusion: Since we reject H_0, we conclude that the alternative hypothesis is true. That is, we conclude that the probability of ambulation is higher in a population of group 2 patients than in population of group 1 patients.

p value: In Table J when a=18, B=9 and a=16, the value of b=2 has an exact probability of occurring by chance alone, when H_0 is true, of 0.001. since the observed value of b=1 is less than 2, its p value is less tan 0.001.

Relative Risk, Odds ratio

- Observational study is an Important class of scientific investigation that is widely used.
- It is a scientific investigation in which neither the subjects under study nor any of the variables of interest are manipulated in any way.
- It may be defined simply as an investigation that is not an experiment. Simplest form have only two variables of interest.
- One of the variables is called the risk factor, or independent variable, and the other variable is referred to as the outcome, or dependent variable.
- Risk factor is used to designate a variable that is thought to be related to some outcome variable.

- Risk factor may be suspected cause of some specific state of the outcome variable.

Ex: The outcome variable might be subjects' status relative to cancer and the risk factor might be their status with respect to cigarette smoking.

Types of Observational Studies: There are two basic types i.e. prospective studies and retrospective studies.

- A prospective study is an observational study in which two random sample of subjects are selected. One sample consists of subjects possessing the risk factor and the other consists of subjects who do not possess the risk factor.
- Data resulting from a prospective study involving two dichotomous variables can be displayed in a 2×2 contingency table.

Reospective study is the reverse of prospective study. The samples are selected from those falling into the categories of the outcome variable.

Relative Risk:

The risk of the disease among the subjects with the risk factor is a/(a+b). The risk of the development of the disease among the subjects without the risk factor is c/(c+d).

Table: Classification of a sample of subjects with Respect to Disease status and Risk Factor

	Disease status		
Risk factor	Present	Absent	Total at risk
Present Absent	a c	b d	a+b c+d
Total	a+c	b+d	n

Relative risk is defined as the ratio of the risk of developing a disease among subjects with the risk factor to the risk of developing the disease among subjects without the risk factor.

Relative risk from a prospective study symbolically as

$$\hat{R}\hat{R} = \frac{a/(a+b)}{c/(c+d)}$$

Where a, b, c and d are as defined in the Table and relative risk is computed from a sample to be used as an estimate of relative risk, RR, for the population from which the sample was drawn.

Confidence interval for RR

$$100(1-\alpha)\%CI = \hat{R}\hat{R}^{1\pm(z_\alpha/\sqrt{x^2})}$$

Where z_α is the two-sided z value corresponding to the chosen confidence coefficient and $\chi 2$ is computed.

Interpretation of RR:

- The value of RR may range anywhere between zero and infinity.
- A value of zero indicates that there is no association between the status of the risk factor and the status of the dependent variable.
- RR of 1 mean that the risk of acquiring the disease is the same for those subjects with the risk factor and those without the risk factor.

Example:

In a prospective study of postnatal depression in women, Boyce at al.(A-16) assessed women at four points in time, at baseline (during the second trimester of pregnancy), and at one, three, and six months postpartum. The subjects were primiparous women cohabiting in a

married or de facto stable relationship. Among the data collected were those shown in Table in which the risk factor is having a spouse characterized as being indifferent
and lacking in warmth and affection. A case is a women who became depressed according to an established criterion. From the sample of subjects in the study, we wish to estimate the relative risk of becoming a case of postnatal depression at one month postpartum when the risk factor is present.

Table : Subject with and without Risk Factor who became cases of postnatal depression at one month postpartum

Risk factor	Cases	Noncases	Total
Present	5	21	26
Absent	8	82	90
Total	13	103	116

Solution:

By equation we compute $\hat{R}\hat{R} = \frac{5/26}{8/90} = \frac{0.192308}{0.088889} = 2.2$

These data indicate that the risk of becoming a case of postnatal depression at one month postmartum when the spouse is indifferent and lacking in warmth and affection is 2.2 times as great as it is among women whose spouse do not exhibit these behaviors.

95% confidence interval for RR is

$$\chi^2 = \frac{116[(5)(82)-(21)(8)]^2}{(13)(103)(26)(90)} = 2.1682$$

The lower and upper confidence limits are, respectively,

$$2.2^{1-1.96/\sqrt{2.1682}} = 0.77$$

$$2.2^{1+1.96/\sqrt{2.1682}} = 6.28$$

Since the interval includes 1, we conclude, at the 0.05 level of significance, that the population risk may be 1. In other words, in the population there may not be an increased risk of becoming a case of postnatal depression at one month postpartum when the spouse is indifferent and lacking in warmth and affection.

Odds Ratio:

- When the data to be analyzed come from a retrospective study, relative risk is not , meaningful measure for comparing two groups.
- A retrospective study is based on a sample of subjects with the disease(cases) and a separate sample of subjects without the disease(controls or noncases)
- The appropriate measure for compairing cases and controls in a retrospective study is the odds ratio.

Table: Subjects of a retrospective study classified according to Status relative to a Risk factor and whether they are cases or controls

	Sample		
Risk factor	Cases	Controls	Total
Present Absent	a c	B d	a+b c+d
	a+c	b+d	n

- **To understand the concept of the odds ratio, we have to understand the terms odds.**
- **The odds for success are the ratio of the probability of success to the probability of failure.**

we can use this definition of odds to define two odds that we can calculate from the Table:

1. The odds of being a case (having the disease) to being a control (not having the disease) among subjects with the risk

factor is [a/(a+b)]/[b/(a+b)]=a/b.

2. The odds of being a case (having the disease) to being a control (not having the disease) among subjects without the risk factor is [c/(c+d)]/[d/(c+d)]=c/d.

- The estimate of the population odds ratio is

$$\hat{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

where a, b, c, and d are defined in Table.

- A confidence interval for OR is

$$100(1-\alpha)\%CI = \hat{OR}^{1\pm(z_\alpha/\sqrt{x^2})}$$

where z_α is the two sided z value corresponding to the chosen confidence coefficient and X^2 is computed by the equation.

Interpretation of the Odds Ratio:

- The odds ratio can assume values between zero and ∞.
- A value of zero indicates no association between the risk factor and disease status.
- A value less than 1 indicates reduced odds of the disease among subjects with the risk factor.
- A value greater than 1 indicates increased odds of having the disease among subjects in whom the risk factor is present.

Example

Cohen et al. (a-17) collected data on men who are booked through the Men's central jail, the main custody facility for men in Los Angels Country.table shows 150 subjects classified as cases or noncases of syphilis infection and according to number of sexual partners(the risk factor) in the preceding 90 days. We wish to compare the odds of syphilis infection among those with three or more sexual partners in the preceding 90 days with the odds of syphilis infection

among those with no sexual partners during the preceding 90 days.

Solution:

The odds ratio $\hat{OR} = \frac{(41)(49)}{(58)(10)} = 3.46$

Here cases are 3.46 times as likely as noncases to have had three or more sexual partners in the preceding 90 days.

95 percent confidence interval for OR is

$$X^2 = \frac{158[(41)(49) - (58)(10)]^2}{(51)(107)(99)(59)}$$

The lower and upper confidence limits for the population OR, are

$$3.46^{1-1.96/\sqrt{10.1223}} = 1.61$$

$$3.46^{1+1.96/\sqrt{10.1223}} = 7.43$$

We conclude with 95 percent confidence that the population

OR is somewhere between 1.61 and 7.43. Since the interval dose not include 1, we conclude that ,in the population, cases are more likely than noncases to have had three or more sexual partners in the preceding 90 days.

THANK YOU