r x c Contingency Table

Chi-square Test for Independence or Homogeneity

Purpose: comparing percentages or testing of association.

Study of the effectiveness of antidepressant

	Relapse		
	No	Yes	Row Total
Desipramine	14	10	24
Lithium	6	18	24
Placebo	4	20	24
Column Total	24	48	72

Hypothesis:

Ho: There is **NO** relation between variable 1 (treatment) and variable 2 (outcome variables).

H*a*: There is relation between two variables.

Compare Observed and Expected Frequencies

	Relapse		
	No	Yes	Row Total
Desipramine	14 (8)	10 (16)	24
Lithium	6 (8)	18 (16)	24
Placebo	4 (8)	20 (16)	24
Column Total	24	48	72

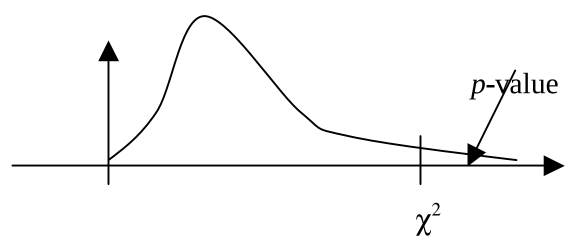
Numbers in (..) :
(i,j)th cell expected freq. = $\dfrac{M_i \times N_j}{T}$

M_i: i-th column total
N_j: j-th **row** total
T : grand total

Test Statistic:

Test Statistics:

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(\{r\text{-}1\}\{c\text{-}1\})$$



Cochran's guidelines: (Assumption: Large sample.)

- None of the expected cell counts less than 1
- No more than 20% of the expected cell frequencies are less than 5.

Decision Rule:

If $\chi^2 > \chi^2_\alpha$ or *p*-value < α, the null hypothesis is rejected.

Test Statistics: $\chi^2 = \dfrac{(14-8)^2}{8} + \dfrac{(10-16)^2}{16} + \dfrac{(6-8)^2}{8} + \dfrac{(18-16)^2}{16} + \dfrac{(4-8)^2}{8} + \dfrac{(20-16)^2}{16}$

$= 10.5$

d.f.= (3-1)(2-1) = 2 => $\chi^2_{.05} = 5.99$ (Chi-square table)

C.V. approach: Since $\chi^2 = 10.5 > \chi^2_{.05} = 5.99$, so we reject null hypothesis. (See Table A.8, page A-26.)

p-value approach: With $\chi^2 = 10.5 > 9.210$, the ***p*-value** of the test is less than 0.01, null hypothesis is rejected.

Conclusion: The relation between treatment and outcome variables is statistically significant.

2 x 2 Contingency Table (A special case of r x c table)

Test Statistics:

$$\chi^2 = \sum_{i=1}^{rc} \frac{(|O_i - E_i| - 0.5)^2}{E_i} \sim \chi^2(1), \text{ with } \textit{Yate's correction}, \text{ "}-0.5\text{"}$$

Example:

Is there a relationship between treatment and heart disease?
(Is there a difference in the percentages of heart disease between people who took Placebo and those who took Aspirin?)

Group	Heart Disease Yes +	Heart Disease No –	Total
Placebo	20 (14)	80 (86)	100
Aspirin	15 (21)	135 (192)	150
Total	35	215	250

35x100/250 = 14,
35x150/250 = 21,
215x100/250 = 86,
215x150/250 = 192

Test Statistic:

$$\chi^2 = \frac{(|20-14|-.5)^2}{14} + \frac{(|80-86|-.5)^2}{86} + \frac{(|15-21|-.5)^2}{21} + \frac{(|135-192|-.5)^2}{192}$$
$$= 4.19$$

d.f.= (2-1)(2-1) = 1 => $\chi_{.05}^2 = 3.84$ (Chi-square table)

C.V. approach:

Since $\chi^2 = 4.19 > \chi^2_{.05} = 3.84$, reject null hypothesis.

p-value approach:

With $\chi^2 = 4.19$, $.025 < p\text{-value} < .05$, null hypothesis is rejected.

Conclusion: There is significant association between the use of Aspirin and heart disease.

An equivalent formula:

Group	Heart Disease Yes +	Heart Disease No –	Total
Placebo	*a*	*b*	*a* + *b*
Aspirin	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*n*

Test Statistics: $\chi^2 = \dfrac{n[|ad - bc| - (n/2)]^2}{(a+c)(b+d)(a+b)(c+d)} \sim \chi^2(1)$, (computational convenient)

Example:

In Aspirin example : $\chi^2 = \dfrac{250[|(20)(135)-(80)(15)|-(250/2)]^2}{(20+15)(80+135)(20+80)(15+135)} = 4.19$

(For small sample, **Fisher's Exact Test** can be used for 2x2 contingency table.)

Example: Suppose we want to determine if people with a rare brain tumor are more likely to have been exposed to benzene than people without a brain tumor. One experimental design used to answer this question. First, we start with cases, people with a disease or condition (brain tumor) and find people who are as similar as possible but who do not have brain tumors. Those people are called controls.

Exposure	Outcome Case	Outcome Control	Total
Yes	50	20	70
No	100	130	230
Total	150	150	300

At the level of significance $\alpha = 0.05$, are "exposure to benzene" and "have brain tumors" independent?

McNemar's Test (Paired-sample test)

Example: A program is designed to promote people to join public health profession. Is there a significant change in the percentage of people who wish to join the public health profession.

Hypothesis:

Ho: There is no association between the **promotion program** and the people who **wish to join the public health profession** or not.
(There is no association between two categorical variables.)

H*a*: There is association between two variables.

(Pairs of **dichotomous** observations were collected.)

After	Before Yes	Before No	Total
Yes	9	**37**	46
No	**16**	82	98
Total	25	119	144

Concordant pairs – provide no information for testing a null hypothesis about the difference in willing to join public health profession status. (i.e. 9 , 82)

Discordant pairs – provide information for testing a null hypothesis about the difference in willing to join public health profession status. (i.e. r = 37, s = 16)

(If null hypothesis is true the discordant pairs should be almost equal to each other.)

Test Statistic: (based on discordant pairs)

$$\chi^2 = \frac{[|r - s| - 1]^2}{(r + s)} \sim \chi^2(1)$$

he example has a test statistic $\chi^2 = \dfrac{[|37 - 16| - 1]^2}{(37 + 16)} = 7.5$

Decision Rule: $\chi^2 = 7.5 > \chi^2_{.05} = 3.84$, or p-value < .05, therefore, reject the null hypothesis.

The Odds Ratio
A method for estimating the effect of the exposure effect.

Risk factor is a variable that is thought to be related to some outcome variable, and it may be a suspected cause of some specific state of this outcome variable.

	Risk Factor		
Outcome	Exposed	Unexposed	Total
Disease No Disease	a c	b d	$a+b$ $c+d$
Total	$a+c$	$b+d$	n

($a + b + c + d = n$)

	Risk Factor		
Outcome	Exposed	Unexposed	Total
Disease No Disease	a c	b d	$a+b$ $c+d$
Total	$a+c$	$b+d$	n

The odds of getting the disease, given that one has the **exposure**, are

$O_+ = P$ [disease | exposed] / P [no disease | exposed] ,
can be estimated by $[a/(a+c)]/[c/(a+c)]$ or a / c

The odds of getting the disease, given that one has **no exposure**, are

$O_- = P$ [disease | unexposed] / P [no disease | unexposed] ,
can be estimated by $[b/(b+d)]/[d/(b+d)]$ or b / d

	Risk Factor		
Outcome	Exposed	Unexposed	Total
Disease No Disease	a c	b d	$a+b$ $c+d$
Total	$a+c$	$b+d$	n

The **odds ratio**, ***OR***, is then defined to be $\frac{O_+}{O_-}$, and its estimate $\hat{OR} = \frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]} = \frac{a/c}{b/d} = \boxed{\frac{ad}{bc}}$

Example: Suppose we want to determine if people with a rare brain tumor are more likely to have been exposed to benzene than people without a brain tumor. One experimental design used to answer this question. First, we start with cases, people with a disease or condition (brain tumor) and find people who are as similar as possible but who do not have brain tumors. Those people are called controls.

	Exposure		
Outcome	Yes	No	Total
Case	50	100	150
Control	20	130	150
Total	70	230	300

Odds ratio = (50/20) / (100/130) = (50x130) / (20x100) = 3.25
(Is the odds ratio different from 1?)

	Risk Factor		
Outcome	Exposed	Unexposed	Total
Disease	a	b	$a + b$
No Disease	c	d	$c + d$
Total	$a + c$	$b + d$	n

Relative risk, ***RR***, is a standard measure of strength of the exposure effect and is defined to be

$RR = P$ [disease | exposed] / P [disease | unexposed]

and its estimate $$\hat{RR} = \frac{a/(a+c)}{b/(b+d)} = \frac{a(b+d)}{b(a+c)} \approx \frac{ad}{bc} = \hat{OR}$$

When a and b are small relative to the values of c and d Odds Ratio is a good estimate of the relative risk.

Example: Suppose we conducted a prospective cohort study to investigate the effect of aspirin on heart disease. A group of patients who are at risk for a heart attack are randomly assigned to either a placebo or aspirin. At the end of one year, the number of patients suffering a heart attack is recorded.

	Group		
Heart Disease	Placebo	Aspirin	Total
Yes +	20	15	35
No –	80	135	215
Total	100	150	250

Relative risk = (20/100)/(15/150) = .2/.1 = 2
(The risk of a heart attack for people on placebo is twice that of people on aspirin.)

	Risk Factor		
Outcome	Exposed	Unexposed	Total
Disease	a	b	$a + b$
No Disease	c	d	$c + d$
Total	$a + c$	$b + d$	n

The (1 – α)100% confidence interval estimate for the Odds Ratio is

$$\left(e^{\ln(\hat{OR}) - z_{\alpha/2}\cdot s^*}, e^{\ln(\hat{OR}) + z_{\alpha/2}\cdot s^*}\right)$$

where $\hat{OR} = \frac{ad}{bc}$, standard error of $\ln(\hat{OR})$ is $s^* = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$, and *a*, *b*, *c* and *d* should not be zero.

The modified estimate is $s^* = \sqrt{\frac{1}{a+.5}+\frac{1}{b+.5}+\frac{1}{c+.5}+\frac{1}{d+.5}}$

Example: In brain tumor example, the 95% confidence interval estimate for the odds ratio of getting brain tumor for person exposed to benzene versus not is $\left(e^{\ln(3.25)-1.96\cdot s^*}, e^{\ln(3.25)+1.96\cdot s^*}\right)$,

where $\hat{OR} = \frac{ad}{bc}$ =3.25,

(People exposed to benzene are more than 3 times as likely to get brain tumor.)

$s^* = \sqrt{\frac{1}{50+.5}+\frac{1}{100+.5}+\frac{1}{20+.5}+\frac{1}{130+.5}}$ =.294.

The 95% confidence interval is $\left(e^{\ln(3.25)-1.96(.294)}, e^{\ln(3.25)+1.96(.294)}\right)$ $\Rightarrow$ **(1.83, 5.78)**,

and it **does not contain 1**.

This implies that there is significant association between benzene exposure and brain tumor.

(There is also confidence interval estimate for RR.)

Odds Ratio Estimation for Paired-sample

The **sample Odds Ratio** of getting disease (or getting result) from exposed to the risk (or improvement) factor versus not for **paired dichotomous data** is $\hat{OR}$ **= r/s** (= 37/16 = 2.31.)

The (1 – α)100% confidence interval estimate of the odds ratio for "paired dichotomous data" is

$$\left(e^{\ln(\hat{OR})-z_{\alpha/2}\cdot s^*}, e^{\ln(\hat{OR})+z_{\alpha/2}\cdot s^*}\right)$$

where $s^* = \sqrt{\frac{r+s}{rs}} = \sqrt{\frac{37+16}{(37)(16)}} = .299$

Example: (Promotion for public health program) The 95% confidence interval estimate of the odds ratio of wishing to join public health profession after promotion program versus before promotion program is

$\hat{OR}$ **= r/s** (= 37/16 = 2.31.)

$s^* = \sqrt{\frac{r+s}{rs}} = \sqrt{\frac{37+16}{(37)(16)}} = .299$

$$\left(e^{\ln(2.31)-1.96(.299)}, e^{\ln(2.31)+1.96(.299)}\right)$$

$\Rightarrow$ (1.29, 4.15).

This interval does not cover 1, it implies that there is significant effect from the promotional program.

Berkson's Fallacy

An investigation surveyed 2784 individuals, 257 of them were hospitalized and examined to determine whether each subject suffered from a disease of the circulatory system or a respiratory illness or both. From only those 257 patients, the chi-square test indicates that there is significant association between having respiratory illness and having circulatory disease.

(Table with 257 individuals)

Circulatory Disease	Respiratory Disease Yes	No	Total
Yes	7	29	36
No	13	208	221
Total	20	237	257

odds ratio = (7)(208)/(29)(13) = 3.86, *p*-value < .025

(Table with 2784 individuals)

Circulatory Disease	Respiratory Disease Yes	No	Total
Yes	22	171	193
No	202	2389	2591
Total	224	2560	2784

Odds ratio = 1.52, p-value > 0.1 ?????

Simpson's Paradox

Example: (City College Admissions)
Overall: *Admission rate for men is higher than women.*

The Whole School

	Admitted	No admitted	Total
Men	198	162	360
Women	88	112	200
Total	286	274	560

Men admitted = 55%
Women admitted = 44%
Sample OR of men versus women = (198)(112) / (162)(88) = 1.56

In separate schools: *Admission rate for women is higher than men.* **??? Lurking variable "schools"**

Business School

	Admitted	No admitted	Total
Men	18	102	120
Women	24	96	120
Total	42	198	240

Men admitted = 15%, Women admitted = 20%
Sample OR of men versus women = (18)(96) / (102)(24) = 0.71

Law School

	Admitted	No admitted	Total
Men	180	60	240
Women	64	16	80
Total	244	76	320

Men admitted = 75%, Women admitted = 80%
Sample OR of men versus women = (180)(16) / (60)(64) = 0.75

The Mantel-Haenszel Method
This same technique can also be used to combine results from several studies identified in a literature search on a specific topic. This technique is sometimes referred to as **meta-analysis**.

Steps:

1. **Test of Homogeneity of Odds Ratios for all contingency tables.**

2. **Summary Odds Ratio:** $\hat{OR} = \dfrac{\sum_{i=1}^{g} a_i d_i / T_i}{\sum_{i=1}^{g} b_i c_i / T_i}$ **, T_i total of the *i*-th table.**

3. **Test of Association: Ho: OR = 1 v.s. Ha: OR ≠ 1.**

$$\chi^2 = \frac{\left[\sum_{i=1}^{g} a_i - \sum_{i=1}^{g} m_i\right]^2}{\sum_{i=1}^{g} \sigma_i^2} \quad \sim \chi^2(1),$$

where a_i = count in the *a* cell count on the *i*-th table,

$$m_i = \frac{M_{1i} N_{1i}}{T_i} \qquad \sigma_i^{\,2} = \frac{M_{1i} N_{1i} M_{2i} N_{2i}}{T_i^2 (T_i - 1)}$$

M_{1i} = the 1-th column total of the *i*-th table, N_{1i} = the 1-th row total of the *i*-th table,
M_{2i} = the 2-nd column total of the *i*-th table, N_{2i} = the 2-nd row total of the *i*-th table,
T_i = the grand total of the *i*-th table.

		Test results, Boys			Test results, Girls	
		Fail	Pass		Fail	Pass
Sleep	Low	20	100	Low	30	100
	High	15	150	High	25	200

Sleep variable is the **risk factor**

(Low => less than 8 hours, High => more than 8 hours)
The Breslow-Day test for homogeneity of the odds ratio is not significant (p-value = .698), so we can be comfortable in combining these two tables.
The Odds Ratio of failing the test for low sleep hours v.s. high sleep hours can be estimated with confidence interval.

Following is SAS output (Output from another statistical software)

Mantel-Haneszel Chi-square Test

SUMMARY STATISTICS FOR SLEEP BY RESULTS
CONTROLLING FOR GENDER

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	12.477	0.001
2	Row Mean Scores Differ	1	12.477	0.001
3	General Association	1	12.477	0.001

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control	**Mantel-Haenszel**	**2.229**	**1.429**	**3.477**
(Odds Ratio)	Logit	2.232	1.421	3.506
Cohort	Mantel-Haenszel	1.977	1.355	2.887
(Col1 Risk)	Logit	1.982	1.351	2.909
Cohort	Mantel-Haenszel	0.889	0.833	0.949
(Col2 Risk)	Logit	0.894	0.833	0.958

The confidence bounds for the M-H estimates are test-based.

Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square = 0.150 DF = 1 Prob = 0.698

Total Sample Size = 640