

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2008, The Johns Hopkins University and Marie Diener-West. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Use of the Chi-Square Statistic

Marie Diener-West, PhD
Johns Hopkins University



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Use of the Chi-Square Statistic in a Test of Association
Between a Risk Factor and a Disease

The Chi-Square (χ^2) Statistic

- Categorical data may be displayed in **contingency tables**
- The **chi-square statistic** compares the observed count in each table cell to the count which would be expected **under the assumption of no association** between the row and column classifications
- The chi-square statistic may be used to test the hypothesis of no association between two or more groups, populations, or criteria
- Observed counts are compared to expected counts

Displaying Data in a Contingency Table

Criterion 2	Criterion 1					
	1	2	3	...	C	Total
1	n_{11}	n_{12}	n_{13}	...	n_{1c}	r_1
2	n_{21}	n_{22}	n_{23}	...	n_{2c}	r_2
3	n_{31}				\vdots	
\vdots	\vdots				\vdots	
\vdots	\vdots				\vdots	
r	n_{r1}	n_{rc}	r_r
Total	C_1	C_2			C_c	n

Chi-Square Test Statistic

- The test statistic is:

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

- The degrees of freedom are:

- $(r-1)(c-1)$

- $r = \#$ of rows and $c = \#$ of columns

- Where:

- $O_i =$ the observed frequency in the i^{th} cell of the table

- $E_i =$ the expected frequency in the i^{th} cell of the table

Example: Is Disease Associated With Exposure?

- The relationship between disease and exposure may be displayed in a contingency table
- We can see that:
 - 37/54 = 68 % of **diseased** individuals were exposed
 - 13/66 = 20 % of **non-diseased** were exposed
- Do these data suggest an association between disease and exposure?

Disease			
Exposure	Yes	No	Total
Yes	37	13	50
No	17	53	70
Total	54	66	120

Observed Counts

- The **observed numbers or counts** in the table are:

Disease			
Exposure	Yes	No	Total
Yes	37	13	50
No	17	53	70
Total	54	66	120

Test of No Association

- **Question of interest: is disease associated with exposure?**
- Calculate what numbers of “exposed” and “non-exposed” individuals would be **expected** in each disease group **if** the probability of disease were the same in both groups
- If there was **no association** between exposure and disease, then the expected counts should nearly equal the observed counts, and the value of the chi-square statistic would be small
- In this example, we can calculate:
Overall proportion with exposure = $50/120 = 0.42$
Overall proportion without exposure = $70/120 = 0.58 = 1 - 0.42$

Expected Counts

- Under the assumption of no association between exposure and disease, the **expected numbers or counts** in the table are:

Disease			
Exposure	Yes	No	Total
Yes	$50/120 \times 54 = 22.5$	$50/120 \times 66 = 27.5$	50
No	$70/120 \times 54 = 31.5$	$70/120 \times 66 = 38.5$	70
Total	54	66	120

Chi-Squared (χ^2) Statistic

$$\begin{aligned}\chi^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(37 - 22.5)^2}{22.5} + \frac{(13 - 27.5)^2}{27.5} \\ &\quad + \frac{(17 - 31.5)^2}{31.5} + \frac{(53 - 38.5)^2}{38.5}\end{aligned}$$

Test of Association

- The **test statistic** is:
 - $\chi^2 = 29.1$ with 1 degree of freedom
- **Assumption:** no association between disease and exposure
- A small value of the χ^2 statistic supports this assumption (observed counts and expected counts would be similar)
- A large value of the χ^2 statistic would not support this assumption (observed counts and expected counts would differ)
- What is the probability of obtaining a statistic of this magnitude or larger when there is no association?

Probability Associated with a χ^2 Statistic

- If the assumption of no association is true, then what is the probability of observing this value of the χ^2 statistic?
- Table A.8 of the Pagano text provides the probability (area in upper tail of the distribution) associated with values of the chi-square statistic for varying degrees of freedom
- Degrees of freedom = 1 for a 2x2 table:

	Area in Upper Tail			
df	0.100	0.0500	...	0.0010
1	2.71	3.84	...	10.83

Conclusion: Test of (No) Association

- For the data in this example, $\chi^2 = 29.1$ with 1 degree of freedom
- From the chi-squared table, the probability obtaining a statistic of this magnitude or larger when there is no association is < 0.001
- In other words, the probability of obtaining discrepancies between observed and expected counts of this magnitude is < 0.001 (unlikely to occur by chance alone)
- Conclude that our finding is unlikely to occur if there is no association between disease and exposure
 - Thus, we conclude that there appears to be an association

Guidelines for Interpreting the χ^2 Statistic

- The χ^2 statistic is calculated under the assumption of no association
- **Large value of χ^2 statistic** \Rightarrow small probability of occurring by chance alone ($p < 0.05$) \Rightarrow conclude that **association** exists between disease and exposure
- **Small value of χ^2 statistic** \Rightarrow large probability of occurring by chance alone ($p > 0.05$) \Rightarrow conclude that **no association** exists between disease and exposure

Short-Cut χ^2 Formula for 2x2 Table

- Suppose:

Disease			
Exposure	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

- Then we can write:

$$\chi_1^2 = \frac{n(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Short-Cut χ^2 Formula for the Example

- Using this formula for the previous example gives

$$\chi_1^2 = \frac{120[(37)(53) - (13)(17)]^2}{54(66)(50)(70)}$$
$$= 29.1$$

Disease			
Exposure	Yes	No	Total
Yes	37	13	50
No	17	53	70
Total	54	66	120

— Same as before!

Helpful Hints Regarding the Chi-Square Statistic

- The calculations use expected and observed counts or frequencies, not proportions
- The χ^2 short-cut formula applies only to 2x2 tables
- Probabilities are available from tables and computing packages

- The χ^2 statistic provides a statistical test for ascertaining whether an association exists between disease and exposure
- A **large value of the χ^2 statistic** indicates that the observed data are unlikely under an assumption of no association between disease and exposure \Rightarrow **small probability (p-value)** \Rightarrow **association**
- A **small value of the χ^2 statistic** indicates that the observed data are likely under an assumption of no association between disease and exposure \Rightarrow **large probability (p-value)** \Rightarrow **no association**



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

Applications of the Chi-Square Statistic in Epidemiology

Applications of the X^2 Statistic in Epidemiology

- Cohort study (2 samples)
- Case-control study (2 samples)
- Matched case-control study (paired cases and controls)

Cohort Study 2x2 Table

Disease			
Factor	Present (D)	Absent (\bar{D})	Total
Present (F)	a	b	a+b
Absent (\bar{F})	c	d	c+d
Total	a+c	b+d	N

Cohort Study: Measure of Association

- Assumptions:

The two samples are independent

- ▶ Let $a+b$ = number of people **exposed** to the risk factor
- ▶ Let $c+d$ = number of people **not exposed** to the risk factor

- Assess whether there is association between exposure and disease by calculating the relative risk (RR)

Cohort Study: Relative Risk

- We can define the relative risk of disease:

$$p_1 = P(\text{disease} | \text{factor present}) = P(D|F)$$

$$p_2 = P(\text{disease} | \text{factor absent}) = P(D|\bar{F})$$

$$RR = \frac{p_1}{p_2}$$

$$= \frac{P(D|F)}{P(D|\bar{F})}$$

is called the **relative risk**

Cohort Study: Test of Association

- For these samples, we can estimate the relative risk as:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

- We can test the hypothesis that $RR=1$ by calculating the chi-square test statistic

$$\chi_1^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Example: Test of Association in a Cohort Study

Develop CHD			
Smoke	Yes	No	Total
Yes	84	2916	3000
No	87	4913	5000
Total	171	7829	8000

- RR = 1.61
- Chi-square statistic = $\chi^2_1 = 10.1 = \frac{8000(84(4913) - 2916(87))^2}{(84 + 87)(2916 + 4913)(84 + 2916)(87 + 4913)}$

Example: Test of Association in a Cohort Study

- Using Table A.8 in the Pagano text, the probability is less than 0.010 (between 0.001 and 0.010)
- This supports an association between exposure and disease

Area in Upper Tail				
df	0.100	0.0500	...	0.0010
1	2.71	3.84	...	10.83

Case-Control Study 2x2 Table

Disease			
Factor	Present (case)	Absent (control)	Total
Present	a	b	a+b
Absent	c	d	c+d
Total	a+c	b+d	N

Case-Control Study: Measure of Association

- Assumptions

- The samples are independent

- ▶ **Cases** = diseased individuals = $a+c$

- ▶ **Controls** = non-diseased individuals = $b+d$

- We are interested in whether:

$$P(F|D) = P(F|\bar{D})$$

- We cannot estimate $P(D)$, the prevalence of the disease and, hence, cannot estimate the RR

- Assess whether there is association between exposure and disease by calculating the odds ratio (OR)

Case-Control Study: Odds for Diseased Group

- The **odds of exposure for the diseased group** is:

$$\frac{p_1}{1-p_1} = \frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{c}$$

Disease			
Factor	Present (case)	Absent (control)	Total
Present	a	b	a+b
Absent	c	d	c+d
Total	a+c	b+d	N

Case-Control Study: Odds for Non-Diseased Group

- The **odds of exposure for the non-diseased group** is:

$$\frac{p_2}{1-p_2} = \frac{\frac{b}{b+d}}{\frac{d}{b+d}} = \frac{b}{d}$$

Disease			
Factor	Present (case)	Absent (control)	Total
Present	a	b	a+b
Absent	c	d	c+d
Total	a+c	b+d	N

Case-Control Study: Odds Ratio

- The odds ratio is:

$$\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

- And is estimated by $OR =$

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

Case-Control Study: Test of Association

- We can test whether $OR=1$ by calculating the chi-square statistic:

$$\chi_1^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

Example: Test of Association in a Case-Control Study

Disease			
Past Smoking	CHD Cases	Controls	Total
Yes	112	176	288
No	88	224	312
Total	200	400	600

- OR = 1.62
- Chi-square statistic = $\chi^2_1 = 7.69 =$

$$\frac{600(112(224) - 176(88))^2}{(112 + 88)(176 + 224)(112 + 176)(88 + 224)}$$

Example: Test of Association in a Case-Control Study

- Using Table A.8 in the Pagano text, the probability is between 0.001 and 0.01
- This supports association between exposure and disease

Area in Upper Tail				
df	0.100	0.0500	...	0.0010
1	2.71	3.84	...	10.83

Matched Case-Control Study Table

		Controls	
		Exposed	Not Exposed
Cases	Exposed	aa	bb
	Not exposed	cc	dd

Matched Study: Measure of Association

- Assumptions
 - Case-control pairs are matched on characteristics such as age, race, sex
 - Samples are not independent
- The discordant pairs are case-control pairs with different exposure histories
 - The matched odds ratio is estimated by bb/cc
 - ▶ **Pairs in which cases exposed but controls not = bb**
 - ▶ **Pairs in which controls exposed but cases not = cc**
 - Assess whether there is association between exposure and disease by calculating the matched odds ratio (OR)

Matched Case-Control Study: Test of Association

- We can test whether $OR = 1$ by calculating McNemar's statistic
- McNemar's test statistic:

$$\chi_1^2 = \frac{(|bb - cc| - 1)^2}{(bb + cc)}$$

Example: Test of Association in a Matched Study

		Controls	
		Exposed	Not Exposed
Cases	Exposed	2	4
	Not exposed	1	3

- OR = bb/cc = 4
- McNemar's test statistic = $\chi^2_1 = 0.80$

$$\chi^2_1 = \frac{(|4 - 1| - 1)^2}{(4 + 1)}$$

Example: Test of Association in a Matched Study

- Using Table A.8 in the Pagano text, the probability is greater than 0.100
- This supports no association between exposure and disease

	Area in Upper Tail			
df	0.100	0.0500	...	0.0010
1	2.71	3.84	...	10.83

Review: Uses of the Chi-Squared Statistic

- The **chi-squared statistic** provides a test of the association between two or more groups, populations, or criteria
- **The chi-square test** can be used to test the strength of the association between exposure and disease in a cohort study, an unmatched case-control study, or a cross-sectional study
- **McNemar's test** can be used to test the strength of the association between exposure and disease in a matched case-control study

Review: Association between Exposure and Disease

- The χ^2 statistic is calculated under the assumption of no association
- A **large** value of the chi-square or McNemar's test statistic \Rightarrow **small probability** \Rightarrow **supports an association**
- A **small** value of the chi-square or McNemar's test statistic \Rightarrow **large probability** \Rightarrow **does not support an association** between exposure and disease