Unit 4
Categorical Data Analysis

"Don't ask what it means, but rather how it is used"
- L. Wittgenstein

Is frequency of exercise associated with better health? Is the proportion of adults who visit their doctor more than once per year, significantly lower among the frequent exercisers than among the non-exercisers? Is alcohol associated with higher risk of lung cancer? Is the apparent association misleading because we have failed to account for the relationship between drinking and smoking? Is greater exposure to asbestos associated with the development of pleural plaques? Is more exposure associated with more pleural plaques? All of these questions are about **counts**.

Units 2 (*Discrete Distributions*) and 4 (*Categorical Data Analysis*) are a two-part introduction to the analysis of **count data** that can be represented in a **contingency table** (a two-way cross-tabulation of the counts of individuals with each profile of traits; eg non-drinker and lung cancer). Data that are counts are **categorical** data (*note: R calls these factors*).

Recall. A categorical variable is discrete. It might be nominal (e.g. – religion), ordinal (e.g. – diagnosis coded as "benign", "suspicious", or "malignant"), or integeter (e.g., 0, 1, 2, visits to the dentist).

Unit 4 (*Categorical Data Analysis*) is an introduction to basic methods for the analysis of categorical data: (1) association in a 2x2 table; (2) variation of a 2x2 table association, depending on the level of another variable; and (3) trend in outcome in a contingency table.

Nice … These methods require minimal assumptions for their validity. In addition, the contingency table approaches introduced here have the advantage of giving us a much closer look at the data than is generally afforded by regression techniques.

Tip **– Always precede a logistic regression analysis with contingency table analyses**.

Table of Contents

Topics

1. Learning Objectives 3
2. Examples of Categorical Data 4
3. Hypotheses of Independence, No Association, Homogeneity..... 9
4. The Chi Square Test of No Association in an RxC Table 10
5. Rejection of Independence: The Chi Square Residual..... 16
6. Confidence Interval Estimation of RR and OR 20
7. Strategies for Controlling Confounding 24
8. Multiple 2x2 Tables - Stratified Analysis of Rates 26
   - A. Woolf Test of Homogeneity of Odds Ratios..... 31
   - B. Breslow-Day-Tarone Test of Homogeneity of Odds Ratios 33
   - C. How to estimate the Mantel-Haenszel Odds Ratio 36
   - D. Mantel Haenszel Test of No Association 37
9. The R x C Table – Test for (Monotone) Trend 40
10. The Chi Square Goodness-of-Fit Test 46

Appendices

A. The Chi Square Distribution 54
B. Probability Models for the 2x2 Table 58
C. Concepts of Observed and Expected 60
D. Review: Measures of Association in a 2x2 Table 64
E. Review: Confounding of Rates 70

Learning Objectives

When you have finished this unit, you should be able to:

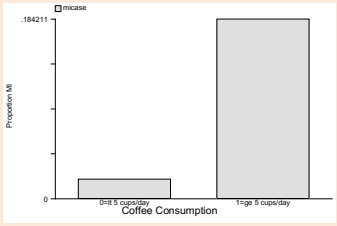
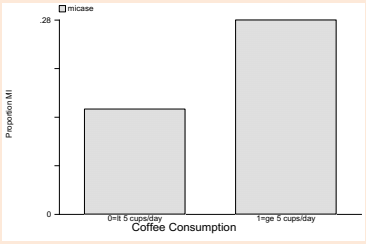
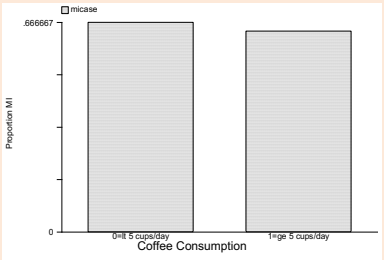
- Perform and interpret the ***chi square test*** of association in a single 2x2 table.
- Define and distinguish between exposure-outcome associations that are ***confounded*** versus ***effect modified***.
- Perform and interpret an analysis of stratified 2x2 tables, using ***Mantel-Haenszel methods***.
- Perform and interpret the ***test of trend*** for RxC tables of counts of ordinal data that are suitable for explorations of dose-response.
- Perform and interpret a ***chi square goodness-of-fit (GOF)*** test.

Note -
Currently, this unit does not discuss matched pairs or matched data.

2. Examples of Categorical Data

Source:
Fisher LD and Van Belle G. Biostatistics: A Methodology for the Health Sciences New York: John Wiley, 1993, page 235, problem #14.

Is there a relationship between coffee consumption and cardiovascular risk? What about the possibility that many coffee drinkers are also smokers and **smoking** is itself a risk factor for heart disease? But suppose we wish to estimate the nature and strength of a coffee-MI relationship ***independent of*** the role of smoking. We can do this by looking at coffee-heart disease data separately within groups (strata) defined by separate levels of the 3rd variable smoking. Consider the following bar graph summaries. On the Y-axis is the proportion with MI. On the X-axis are the two groups: low coffee drinkers (left bar) and high coffee drinkers (right bar). Separately in each stratum of smoking, low coffee drinkers are compared with high coffee drinkers with respect to proportion suffering a myocardial infarction (MI).

Never Smoked	
Former Smoker	
….	*… Some strata of smoking omitted …*
45+ cigarettes/day	

- Stratum=never smokers: Among **never** smokers, high coffee consumption is associated with MI
- Stratum=former smokers: Among **former smokers**, the coffee-MI association is ***less strong***.
- Stratum=45+cigarettes/day: Among frequent smokers, there is ***no longer evidence*** of a coffee-MI association. → We have an interaction; that is

What do you see? The association of coffee consumption with myocardial infarction (MI) is different (modified by), depending on smoking status.

Nature _____ **Population/ Sample** _____ **Observation/ Data** _____ **Relationships/ Modeling** _____ **Analysis/ Synthesis**

In Unit 2 *(Discrete Distributions),* we learned three probability distribution models for discrete data: Binomial**, Poisson,** and **Hypergeometric**. Here are some examples.

Example - Binomial Model for One Group Count of Events of Success

Does minoxidil show promise for the treatment of hair loss?

n=13 volunteers
↓
Administer minoxidil
↓
Wait 6 months
↓
Count occurrences of new hair growth. Call this X.

Suppose we observe X=12.
Possible values of X=count of occurrences of new hair growth are 0, 1, 2, …, 13. Thus,

IF: (1) π = probability [new hair growth] for all 13 volunteers, and the
(2) outcomes for each of the 13 volunteers are independent

THEN: X is distributed Binomial (n=13, π)

The likelihood L (*think "chances of"*) of the outcomes in the one group intervention study design data is modeled as a binomial probability:

$$L_X(x) = \Pr[X=x] = \binom{13}{x} \pi^x (1-\pi)^{13-x}$$

Example -

The probability of X=12 events of "new hair growth" in N=13 trials ("study participants") = $\binom{13}{12} \pi^{12} (1-\pi)^1$

Example - The product of 2 binomials is used to model 2 independent outcome counts in a cohort study

In a randomized controlled trial, is minoxidil better than standard care for the treatment of hair loss?

Consent of N=30 volunteers

↓

Randomization

Minoxidil	Standard Care
$n_1 = 13$	$n_2 = 17$
Administer minoxidil	Administer standard care
↓	↓
Wait 6 months	Wait 6 months
↓	↓
Count occurrences of new hair growth. Call this X_1.	Count occurrences of new hair growth. Call this X_2.

This design produces a 2x2 table array of count data that is appropriately modeled using the ***product of two binomial*** distributions (one in each row).

	New Growth	Not	
Minoxidil	$X_1 = 12$		$n_1 = 13$ trials
Standard care	$X_2 = 6$		$n_2 = 17$ trials

IF: (1) π_1 = probability [new hair growth] among minoxidil recipients
(2) π_2 = probability [new hair growth] among standard care recipients
(3) The outcomes for all 30 trial participants are independent

THEN: (1) X_1 is distributed Binomial (n_1 =13, π_1)
(2) X_2 is distributed Binomial (n_2 =17, π_2)

The likelihood L ("chances of") of the outcomes in the two-group **cohort** study design data is modeled as the product of 2 independent binomial likelihoods:

$$L_{X_1X_2}(x_1,x_2) = \Pr[X_1=x_1 \text{ and } X_2=x_2] = \left\{\binom{13}{x_1}\pi_1^{x_1}(1-\pi_1)^{13-x_1}\right\} * \left\{\binom{17}{x_2}\pi_2^{x_2}(1-\pi_2)^{17-x_2}\right\}$$

The probability of X_1=6 and X_2=13 events in the standard and minoxidil groups is =

$$\binom{13}{12}\pi_1^{12}(1-\pi_1)^{1} * \binom{17}{6}\pi_2^{6}(1-\pi_2)^{11}$$

Example - The product of 2 binomials is used to model 2 exposure history counts in a case-control study
Is a history of oral contraceptive (OC) use associated with thromboembolism?

Enroll cases of thromboembolism $n_1 = 100$	Enroll controls: w/o thromboembolism $n_2 = 200$
Query history of OC use ↓	Query history of OC use ↓
↓	↓
Count histories of OC use. Call this X_1.	Count histories of OC use. Call this X_2.

This design also produces a 2x2 table array of count data that is correctly modeled using two binomial distributions (now it is one in each column).

	Case	Control
History of OC Use	$X_1 = 65$	$X_2 = 118$
Not		
	$n_1 = 100$ trials	$n_2 = 200$ trials

Reminder: A case-control design does not permit the estimation of probabilities of disease.

IF:
(1) π_1 = probability [history of OC use] among cases
(2) π_2 = probability [history of OC use] among controls
(3) The histories for all 300 observations are independent

THEN:
(1) X_1 is distributed Binomial (N=100, π_1)
(2) X_2 is distributed Binomial (N=200, π_2)

The likelihood ("chances of") L of the outcomes in the two group **case-control** study design data is modeled as the product of 2 independent binomial likelihoods:

$$L_{X_1X_2}(x_1,x_2) = \Pr[X_1=x_1 \text{ and } X_2=x_2] = \left\{\binom{100}{x_1}\pi_1^{x_1}(1-\pi_1)^{100-x_1}\right\} * \left\{\binom{200}{x_2}\pi_2^{x_2}(1-\pi_2)^{200-x_2}\right\}$$

Example

The probability of $X_1=65$ and $X_2=118$ counts of OC use history is = $\binom{100}{65}\pi_1^{65}(1-\pi_1)^{35} * \binom{200}{118}\pi_2^{118}(1-\pi_2)^{82}$

Example - A Hypergeometric Distribution is used for the Cross-Tabulation of Counts in a Cross-Sectional Prevalence Study.

WHO investigated the variability in the prevalence of Alzheimer's Disease among distinct populations.

	Alzheimer's Disease	No Alzheimer's Disease	
African Black	$X_1 = 115$	22,885	$n_1 = 23{,}000$
Native Japan	$X_2 = 7{,}560$	46,440	$n_2 = 54{,}000$
European White	$X_3 = 105{,}930$	857,070	$n_3 = 963{,}000$
South Pacific	$X_4 = 21$	8,479	$n_4 = 8{,}500$
North American Indian	$X_5 = 44$	10,956	$n_5 = 11{,}000$
	113,670	945,830	1,059,500

An analysis of these data might test the "no association" null hypothesis that the prevalence of Alzheimer's Disease is the same in all race/ethnicity groups. The correct null hypothesis probability distribution to use is the following multiple hypergeometric probability:

$$\frac{\binom{n_1}{x_1}\binom{n_2}{x_2}\binom{n_3}{x_3}\binom{n_4}{x_4}\binom{n_5}{x_5}}{\binom{n_1+n_2+n_3+n_4+n_5}{x_1+x_2+x_3+x_4+x_5}}$$

$$= \frac{\binom{23{,}000}{115}\binom{54{,}000}{7{,}560}\binom{963{,}000}{105{,}930}\binom{8{,}500}{21}\binom{11{,}000}{44}}{\binom{1{,}059{,}500}{113{,}670}}$$

Whew! Note to class – You will **not** be required to work with this distribution in this unit.

3. Hypothesis of Independence, No Association, Homogeneity of Proportions

There are multiple wordings of the same null hypothesis. "Independence", "No Association", "Homogeneity of Proportions" are all the same null hypothesis. For example,

(1) ***"Length of time since last visit to physician" is independent of "income"*** says that income has no bearing on the elapsed time between visits to a physician. The expected elapsed time is the same regardless of income level.

(2) ***"There is no association between coffee consumption and lung cancer"*** says that an individual's likelihood of lung cancer is not affected by his or her coffee consumption.

(3) The equality of probability of success on treatment (experimental versus standard of care) in a randomized trial of two groups is a test of ***homogeneity of proportions*****.**

We use the same procedure (the chi square test) to test hypotheses of "independence", "no association", "homogeneity of proportions" in an analysis of contingency table data.

4. The Chi Square Test of No Association in an R x C Table

Example

Is there an association between income level and the time elapsed since last visit to a physician (H_A)? Or is the elapsed time independent of income level (H_O), meaning that there is no association?

Last Consulted Physician

Income	≤ 6 months	7-12 months	>12 months	Total
< $6000	O_{11} = 186	38	35	$O_{1.}$ = 259
$6000-$9999	227	54	45	326
$10,000-$13,999	219	78	78	375
$14,000-$19,999	355	112	140	607
≥ $20,000	653	285	259	1197
Total	$O_{.1}$ = 1640	567	557	$N = O_{..}$ = 2764

Key to subscripts "i" and "j" and so on. The subscript "i" tells you the row (e.g., i=1 tells you that the row is row 1). If cell entries are summed over all the rows, the row subscript goes away and is replaced by a dot (".."). Similarly, if the cell entries are summed over all the column, the column subscript goes away and is replaced by a dot (".")

Columns, "j"

Rows, "i"	$j = 1$	…	$j = C$	
$i = 1$	$O_{11}=n_{11}$	…	$O_{1C}=n_{1C}$	$N_{1.} = O_{1.}$
…	…			
$i = R$	$O_{R1}=n_{R1}$	…	$O_{RC}=n_{RC}$	$N_{R.} = O_{R.}$
	$N_{.1} = O_{.1}$	…	$N_{.C} = O_{.C}$	$N=O_{..}$

Key

Row 1, i=1: $O_{11} = n_{11}$ = 186 Row 1 total (taken over all j) is $N_{1.} = O_{1.}$ = 259

Column 1, j=1: $O_{11} = n_{11}$ = 186 Column 1 total (taken over all i) is $N_{.1} = O_{.1}$ = 1640

The grand total (taken over all i and all j) is $N=O_{..}$ = 2764

Preliminary: How shall we estimate the overall probability of the occurrence of income < $6000?

Occurrence of income < $6000 is an example of a **marginal event**, "marginal" because it is taken over all possibilities of elapsed time (all the "j's"). By convention we say $\pi_{1.}$ **= Pr [income < $6000].** It makes sense to estimate $\pi_{1.}$ **= Pr [Income < $6000]** using the proportion of occurrences in the entire table that correspond to income < $6000:

$$= \frac{\text{total \# instances} < \$6000}{\text{total of table}} = \frac{259}{2764}$$

$$= \frac{\text{row 1 total}}{\text{total of table}} = \frac{O_{1.}}{N}$$

$$= \hat{\pi}_{1.}$$

Preliminary: How shall we estimate the overall probability of the occurrence of elapsed time < 6 months

By the same reasoning, the event of elapsed time ≤ 6 months is also an example of a marginal event, here because it is elapsed time ≤ 6 months regardless of income. Here, by convention we say **$\pi_{.1}$ = Pr [elapsed time ≤ 6 months].** It makes sense to estimate this marginal event probability using the observed *column* 1 proportion

$$= \frac{\text{total \# instances elapsed time is } \le \text{ 6 months}}{\text{total of table}} = \frac{1640}{2764}$$

$$= \frac{\text{column 1 total}}{\text{total of table}} = \frac{O_{.1}}{N}$$

$$= \hat{\pi}_{.1}$$

The **chi square test of no association** is a test of the null hypothesis that income and elapsed time are independent. We assume the null hypothesis is true, apply it to the data and, in particular we obtain the null hypothesis expected proportions *(these are then compared to the observed proportions)*

Recall the meaning of independence of two coin tosses

You are probably already familiar with the idea that the probability of 2 heads in 2 coin tosses is (.5 for the first toss) × (.5 for the second toss) = .25, representing a 25% chance of 2 heads in 2 tosses. This is independence:

Pr[heads toss #1 and heads toss #2 | H_O: independence] = Pr[heads toss #1] x Pr[heads toss #2]

These are marginal event probabilities

The statistical model of independence of income and elapsed time has the same intuition

Pr [income is level "i" and elapsed time is level "j")

= Prob (income is level "i") x Prob (elapsed time is level "j")

That is:

$$\pi_{ij \text{ HO: independence}} = [\pi_{i.}][\pi_{.j}]$$

Note the dots for overall row and overall column, respectively

Step 1: State Assumptions

1. The contingency table of count data is a ***random sample*** from some population
2. The cross-classification of each individual is independent of the cross-classifications of all other individuals.

Step 2: Null and Alternative Hypotheses

$H_O: \pi_{ij} = \pi_{i.}\pi_{.j}$ **This is the null hypothesis of independence again.**

$H_A: \pi_{ij} \neq \pi_{i.}\pi_{.j}$

Step 3: Estimate the π_{ij} under the Null Hypothesis of Independence

$$\hat{\pi}_{ij\ \text{NULL HYPOTHESIS}} = \hat{\pi}_{i.}\hat{\pi}_{.j} \quad \text{where}$$

$$\hat{\pi}_{i.} = \frac{n_{i.}}{n} = \frac{\text{row "i" total}}{\text{grand total}} \quad \textbf{and} \quad \hat{\pi}_{.j} = \frac{n_{.j}}{n} = \frac{\text{column "j" total}}{\text{grand total}}$$

Step 4: Obtain the null hypothesis model expected counts E_{ij}

$$E_{ij} = (\#\ \text{trials})[\hat{\pi}_{ij\ \text{NULL HYPOTHESIS}}] = (n)\hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{[\text{row "i" total}][\text{column "j" total}]}{n}$$

Step 5: The test statistic measure of "extremeness" of the data relative to an assumed null hypothesis compares "observed" to "null hypothesis" expected counts as follows

For each cell, we can obtain a sense for whether the assumption of the null hypothesis has led to (suspiciously) unusual data by **comparing observed counts versus the null hypothesis expected counts**. Large disparities are evidence against the null and in favor of the alternative. The statistical test is a {z-score}^{2} measure that involves observed and expected counts:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

The chi square test statistic of association is the sum of these over all the cells in the table:

$$\chi^2_{df = (R-1)(C-1)} = \sum_{i=1}^{R}\sum_{j=1}^{C}\left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right]$$

How to compute # degrees of freedom (df)

df = [total # cells] – [# constraints on data]

= [RC] – 1 Why *(-1): One df is lost because the grand total is fixed and cannot change*
- (R-1) Why *–(R-1): Each row total is fixed but then we subtract off the extra -1*
- (C-1) Why *–(C-1): Each col total is fixed but then we subtract off the extra -1*

= [RC] – 1 – R + 1 – C + 1

= [RC] – R – C + 1

= (R-1)(C-1) ✔ *Hopefully, this is easy to remember!*

Behavior of Chi Square Statistic, under each of the null and alternative hypotheses:

Null is true (no association)	**Alternative is true**
Each $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ is close to zero	Each $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ is >> 0
$\chi^2_{df=(R-1)(C-1)} = \sum_{i=1}^{R}\sum_{j=1}^{C}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$ is small and has expected value = (R-1)(C-1)	$\chi^2_{df=(R-1)(C-1)} = \sum_{i=1}^{R}\sum_{j=1}^{C}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$ is **large** and has expected value > > (R-1)(C-1)

Step 7: Decision Rule

Reject null hypothesis (H_O) when test statistic is large, as when

- achieved significance level is small
- test statistic value is greater than the critical value threshold, which is defined by the upper $(\alpha)100^{th}$ percentile of Chi square distribution.

Step 8: Computations

(1) For each cell (row=i, column = j), compute

$$E_{ij} = (\#\text{ trials})[\hat{\pi}_{ij\ \text{NULL HYPOTHESIS}}] = (n)\hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{[\text{row "i" total}][\text{column "j" total}]}{n}$$

(2) And then compute for each cell

$$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Observed Counts

Last Consulted Physician

Income	≤ 6 months	7-12 months	>12 months	Total
< $6000	O_{11} = 186	O_{12} =38	O_{13} =35	$O_{1.}$ =259
$6000-$9999	O_{21} =227	O_{22} =54	O_{23} =45	$O_{2.}$ =326
$10,000-$13,999	O_{31} =219	O_{32} =78	O_{33} =78	$O_{3.}$ =375
$14,000-$19,999	O_{41} =355	O_{42} =112	O_{43} =140	$O_{4.}$ =607
≥ $20,000	O_{51} =653	O_{52} =285	O_{53} =259	$O_{5.}$ =1197
Total	$O_{.1}$ =1640	$O_{.2}$ =567	$O_{.3}$ =557	$O_{..}$ =2764

Null Hypothesis Model Expected Counts

Income	Last Consulted Physician: ≤ 6 months	7-12 months	>12 months	Total
< $6000	$E_{11} = \frac{(259)(1640)}{2764} = 153.68$	E_{12} =53.13	E_{13} =52.19	$E_{1.}$ =259
$6000-$9999	E_{21} =193.43	E_{22} =66.87	E_{23} =65.70	$E_{2.}$ =326
$10,000-$13,999	E_{31} =222.50	E_{32} =76.93	E_{33} =75.57	$E_{3.}$ =375
$14,000-$19,999	E_{41} =360.16	E_{42} =124.52	E_{43} =122.32	$E_{4.}$ =607
≥ $20,000	E_{51} =710.23	E_{52} =245.55	$E_{53} = \frac{(1197)(557)}{2764} = 241.22$	$E_{5.}$ =1197
Total	$E_{.1}$ =1640	$E_{.2}$ =567	$E_{.3}$ =557	$E_{..}$ =2764

$$\chi^2_{(R-1)(C-1)} = \sum_{\text{all cells}} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] = \frac{(186-153.68)^2}{153.68} + \ldots + \frac{(259-241.22)^2}{241.22} = 47.90$$

with degrees of freedom = (R-1)(C-1) = (5-1)(3-1) = 8

Achieved significance level, p-value = Prob [Chi Square w df=8 ≥ 47.90] << .0001

Step 9: Statistical Conclusion

We **reject** the null hypothesis because its assumption and application to the observed data has produced a **very unlikely** result: "*under the null hypothesis model, the chances of obtaining an observed test statistic value as far away from small as 47.90 were less than 1 chance in 10,000*".

Special Case
The Chi Square Test of No Association in a 2 x 2 Table

Often (especially in epidemiology textbooks), the "a", "b", "c", and "c" notation is used to represent the cell counts in a 2x2 table as follows:

		2nd Classification Variable		
		1	2	
1st Classification	1	a	b	a + b
	2	c	d	c + d
		a + c	b + d	n

The calculation for the chi square test that you just learned, namely:

$$\chi^2_{1\,DF} = \sum_{\text{all cells}} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

has the following ***equivalent*** formula when using the "a" "b" "c" "d" "n" notation:

$$\chi^2_{1DF} = \frac{n\left(ad\text{-}bc\right)^2}{\left(a+c\right)\left(b+d\right)\left(c+d\right)\left(a+b\right)}$$

5. Rejection of Independence – The Chi Square Residual

Okay. You've rejected the null hypothesis of "no association." NOW WHAT? What have you actually learned? Not a lot, actually. You don't actually know where the observed counts deviate a lot from the null hypothesis expected counts. Ahhh … but we have a tool to help us that we've already seen. Z-score to the rescue!

Appendix A *gives us the following reasoning …*

IF	THEN	Comment
X has a distribution that is Binomial (n,p) exactly	X is approximately Normal ($n\mu$, $n\sigma^2$) with $\mu = p$ $\sigma^2 = p(1-p)$	This X is our "observed" count, O.
	$Z\text{-score} = \frac{X - E(X)}{SD(X)}$ $= \frac{X - n\mu}{\sqrt{n}\sigma}$ $= \frac{X - np}{\sqrt{np(1-p)}}$ is approx. Normal(0,1)	This E(X) is our expected count, E. Thus, the numerator of the Z-score is { O – E } The denominator of the Z-score is almost $\sqrt{E} = \sqrt{Np}$ (but not quite) $\text{Z-score} \approx \frac{\{ O - E \}}{\sqrt{E}}$
		This approximation to the z-score and similar formulae are called residuals.

There are at least two kinds of residuals that we can use to discover where the observed counts (O) are deviating significantly from the null hypothesis expected counts (E). They have the advantage of being interpretable as Z-scores (or at least reasonably so)!

Name	Calculation	Remark
Standardized Residuals	$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$	These are approximately Z-scores. Therefore, they are distributed Normal(0,1) approximately
Adjusted Standardized Residuals	$r_{ij}^{*} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}\left(1 - \frac{n_{i.}}{n}\right)\left(1 - \frac{n_{.j}}{n}\right)}}$	These are also reasonably approximated as distributed Normal(0,1)

How do these residuals behave under the Null (No association)? Under the alternative (Association)?

Null is True	Alternative is True
{O – E} will be near zero →	{O – E } will be appreciably different from zero when measured in SE units. →
Residual will be **small**.	Residual will be **large in absolute value**.
	How large is "significantly" large? We answer this using the Normal(0,1) distribution.

Example –
Investigation of Relationship between Income and Physician Visits (continued)

In the table below, I show only the adjusted standardized residuals, r^* that are bigger than 1.96 in magnitude (approximately), with negative ones shaded.

Income	≤ 6 months	7-12 months	>12 months
< $6000	**+4.3**	**-2.4**	**-2.8**
$6000-$9999	**+4.0**		**-3.0**
$10,000-$13,999			
$14,000-$19,999			**+2.0**
≥ $20,000	**-4.5**	**+3.8**	

It takes some practice deciphering the pattern of standardized residuals that have large magnitudes (irrespective of direction). In this example, it seems that:

1. Low-income individuals were ***relatively more likely*** to visit their physician within 6 months than were higher income individuals.

2. Low-income individuals were ***relatively less likely*** to delay seeing their physician beyond one year than were higher income individuals.

The Small Cell Frequency Problem

The problem.

Did you notice? We've been talking about counts which, on the face of it, suggests that we should be using discrete variable probability distribution models. And yet. We've been using a chi square probability distribution model which is for a continuous variable.

We've been using an approximation. Because the count data are actually discrete, the use of a continuous variable probability distribution model (the chi square) is actually an approximation.

So, when is it okay to use the chi square distribution approximation in the analysis of discrete outcomes? The short answer is: when the cell frequencies are sufficiently large.

How large is large? When is it okay to use the Chi Square Approximation?

The **Chi Square, Approximate, Test of General Association in a 2x2** table may be applied if:

All of the expected frequencies (E_{ij}) are ***greater than 5***

What do I do if the Chi Square approximation is *not* appropriate?

(1) Do a **Fisher's Exact test**. It's always valid. We have already learned the Fisher's Exact test for data in a single 2x2 table (BIOSTATS 640 Unit 2, *Discrete Distributions*). There exist similar procedures for larger tables (not shown here – but the idea is the same); ***or***

(2) **Combine** adjacent rows and/or columns to attain required minimum expected cell frequencies and try the Chi Square approximation again. The disadvantage to this approach is a loss of degrees of freedom. The resulting test statistic is less powerful.

6. Confidence Interval Estimation of Relative Risk (RR) and Odds Ratio (OR)

Unfortunately, the relative risk (RR) and odds ratio (OR) statistics do not have sampling distributions that can be modeled as approximately normal.

But **transformations of the relative risk (RR) and odds ratio (OR)** statistics can be modeled as **approximately normal**. The transformation we need is the natural logarithm, ln().

This suggests that we may follow the following **steps** to obtain confidence interval estimates of RR and OR:

- Step 1: Estimate RR (or OR)
- Step 2: Obtain the natural logarithms, ln[RR] (or ln[OR])
- Step 3: Obtain their associate variance estimates, var(ln[RR]) or var(ln[OR])
- Step 4: Use the z-score approach to obtain confidence intervals for ln[RR] or ln[OR]
- Step 5: Exponentiate these limits to obtain confidence intervals for RR or OR

Suppose, generically, we use the notation ***θ = function [statistic]*** to represent the transformation we need to obtain a new statistic whose sampling distribution is ***approximately normal***.

- For relative risk: define θ = ln [RR]
- For odds ratio: define θ = ln [OR].

The **"z-score" method** used to obtain confidence interval estimates of the relative risk (RR) or the odds ratio (OR) says the following:

If θ = parameter of interest

$\hat{\theta}$ = "best" guess based on a reasonably large sample size

$\widehat{\text{var}}(\hat{\theta})$ = "best" guess of the variance of $\hat{\theta}$

Then $\dfrac{\theta - \hat{\theta}}{\sqrt{\widehat{\text{var}}(\hat{\theta})}}$ "Z-score" has a distribution that is well approximated as **Normal (0,1)**

Thus, we can use this new "z-score" variable to obtain the confidence interval we're after. For a (1-α)100% confidence interval:

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta})}$$

Recall. We have two "notations" used for keeping track of counts in the 2x2 table:

Row var	Column Var: yes	Column Var: no	
yes	n_{11}	n_{12}	$n_{1.}$
no	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	

	Column Var: yes	Column Var: no	
yes	a	b	(a+b)
no	c	d	(c+d)
	(a+c)	(b+d)	

	Transformation θ = function (statistic) that is Normal, approximately	**Variance (transformation) both 2x2 table notations provided**
Relative Risk, RR	θ = ln (RR)	$\widehat{\text{var}}\left(\ln[RR]\right) \approx \left[\frac{1}{n_{11}}\right] - \left[\frac{1}{n_{1.}}\right] + \left[\frac{1}{n_{21}}\right] - \left[\frac{1}{n_{2.}}\right]$ $\widehat{\text{var}}\left(\ln[RR]\right) \approx \left[\frac{1}{a}\right] - \left[\frac{1}{a+b}\right] + \left[\frac{1}{c}\right] - \left[\frac{1}{c+d}\right]$
Odds Ratio, OR	θ = ln (OR)	$\widehat{\text{var}}\left(\ln[OR]\right) \approx \left[\frac{1}{n_{11}}\right] + \left[\frac{1}{n_{12}}\right] + \left[\frac{1}{n_{21}}\right] + \left[\frac{1}{n_{22}}\right]$ $\widehat{\text{var}}\left(\ln[OR]\right) \approx \left[\frac{1}{a}\right] + \left[\frac{1}{b}\right] + \left[\frac{1}{c}\right] + \left[\frac{1}{d}\right]$

Outline of Steps in Obtaining a Confidence Interval for RR or OR

Step 1: If you want CIs for RR or OR, as your first step, consider instead their natural logarithms ln (RR) and ln (OR). These can reasonably be assumed to be distributed Normal, even when RR and OR are not distributed Normal.

Step 2: Obtain CIs for ln (RR) or ln (OR)

Step 3: To obtain confidence interval estimates for RR and OR, exponentiate the confidence interval estimates for ln (RR) and ln (OR).

Confidence Interval Estimate of Relative Risk (RR)

Example -

	CHD	No CHD	
High Cholesterol	27	95	122
Not high cholesterol	44	443	487
	71	538	609

$$\hat{RR}=\frac{n_{11}/n_{1.}}{n_{21}/n_{2.}}=\frac{27/122}{44/487}=2.45$$

Step 1: Obtain the natural logarithm ln (RR).

$$\ln(\hat{RR})=\ln(2.45)=.896$$

$$\hat{Var}\left(\ln\left[RR\right]\right)\approx\left[\frac{1}{n_{11}}\right]-\left[\frac{1}{n_{1.}}\right]+\left[\frac{1}{n_{21}}\right]-\left[\frac{1}{n_{2.}}\right]$$

e.g., $$\hat{Var}\left(\ln\left[RR\right]\right)\approx\left[\frac{1}{27}\right]-\left[\frac{1}{122}\right]+\left[\frac{1}{44}\right]-\left[\frac{1}{487}\right]=0.0495$$

Step 2: Obtain confidence interval for ln(RR)

For a 95% confidence interval, $z_{1-\alpha/2}=z_{.975}=1.96$ so that with 95% confidence

$$.896-1.96\sqrt{0.0495}\le\ln(RR)\le.896+1.96\sqrt{0.0495}$$ or

$$.4599\le\ln(RR)\le1.332$$

Step 3: Exponentiate .4599 and 1.332 to obtain the confidence interval for RR With 95% confidence.

$$e^{.4599}\le RR\le e^{1.332}\ \rightarrow$$

so that with 95% confidence, we estimate that

$$1.584\le RR\le3.789$$

Confidence Interval Estimate of Odds Ratio (OR)

Example -

	Disease	No disease
Exposed	a=8	b=30
Not exposed	c=2	d=20

$$\hat{OR}=\frac{ad}{bc}=\frac{(8)(20)}{(2)(30)}=2.67$$

Step 1: Utilize the z-score approximation for the distribution of ln(OR).

$$\ln(\hat{OR})=\ln(2.67)=.981$$

$$\hat{var}(\ln[\hat{OR}])\approx\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$$

e.g, $\hat{var}(\ln[\hat{OR}])\approx\frac{1}{8}+\frac{1}{30}+\frac{1}{2}+\frac{1}{20}=.708$

Step 2: Obtain confidence interval for ln (OR)

For a 95% confidence interval, $z_{1-\alpha/2}=z_{.975}=1.96$ so that with 95% confidence

$$.981-1.96\sqrt{.708}\le\ln(OR)\le.981+1.96\sqrt{.708}$$ or

$$-.67\le\ln(OR)\le2.63$$

Step 3: Exponentiate -0.67 and 2.63 to obtain the confidence interval for OR

With 95% confidence.

$$e^{-.67}\le OR\le e^{2.63}\ \rightarrow$$

so that with 95% confidence, we estimate that

$$.512\ \le\ OR\ \le\ 13.87$$

7. Strategies for Controlling Confounding

We can control confounding at **study design**.

- Restriction
- Matching

Alternatively, we can also control confounding **analytically**.

- Stratification
- Standardization
- Matching

Restriction

Restriction is limiting the analysis to the sub-sample comprised solely of persons who are the same with respect to the confounder. Nice. But there's a catch. For example:

- A study of males only will not produce results that are confounded by gender effects. But nothing is learned about gender influences; and

- A study of non-smokers only will not produce results that are confounded by the effects of smoking. But nothing is learned about smoking's influence.

Thus, the advantage of restriction is a guarantee of control for confounding. However, there are also disadvantages. The sample size is limited and generalizability is reduced.

Matching in a Cohort Study

Matching in a cohort study involves the following.

- First, enroll exposed persons without restriction.
- Once you have your exposed persons, enroll unexposed only if they match exposed.

Matching in a Case-Control Study

In a case-control study the following occurs.

- First, enroll cases without restriction.
- Once you have your cases, enroll controls only if they match cases.

Be careful!! Matching is not necessarily a good idea

- In case-control studies, controls may be artificially similar to cases.
- Estimates of association may be spuriously low
- If matching is related to exposure only, not confounding, then spurious confounding may be introduced
- Sample size is reduced
- Identical matched pairs provide no information

When NOT to match

- Most case-control studies.
- On a variable that is intermediary.
- When many controls are available

When to consider matching

- In an experiment and some cohort studies
- On some variables (age, sex, site)

8. Multiple 2x2 Tables – Stratified Analysis of Rates

GOAL: To explore confounding and effect modification of an exposure-outcome relationship

Is coffee consumption associated with heart attacks (MI)? What about the potential role of smoking in all this? Is the coffee-MI association confounded by smoking? Is the coffee-MI association different (modified by), depending on smoking status? Recall:

- **Interaction (effect modification)** is present when a relationship of interest (exposure-outcome) is different, depending on the level of some third variable.

- **Confounding** of a relationship of interest (exposure-outcome) when the association is distorted when we fail to control for the level of the third variable.

Example -

Overall

		Myocardial Infarction (MI)	
		Yes	No
Coffee	Yes	a = 1394	b = 755
	No	c = 147	d = 200

$$OR_{crude} = \frac{ad}{bc} = \frac{(1394)(200)}{(147)(755)} = 2.51$$

Stratum: Smokers

		Myocardial Infarction (MI)	
		Yes	No
Coffee	Yes	a = 1011	b = 390
	No	c = 81	d = 77

$$OR_{among\ smokers} = \frac{ad}{bc} = \frac{(1011)(77)}{(81)(390)} = 2.46$$

Stratum: NON-smokers

		Myocardial Infarction (MI)	
		Yes	No
Coffee	Yes	a = 383	b = 365
	No	c = 66	d = 123

$$OR_{among\ non\text{-}smokers} = \frac{ad}{bc} = \frac{(383)(123)}{(66)(365)} = 1.96$$

Is coffee drinking associated with myocardial infarction (MI)?

Is an "apparent" association (OR_{crude} = 2.51) misleading or incorrect because we failed to control for **confounding** by smoking?

CONFOUNDING:
"It can happen that, when groups are combined, the overall OR is significantly different than the individual ORs across groups, even if these ORs are deemed 'homogeneous'" (Scott Evans)

We also want to know: is the physiologic effect of coffee and its relationship to MI different (**modified**), depending on smoking status? →

EFFECT MODIFICATION:
"The question that naturally arises is whether we should combine the information in those two tables and use all the available data to ascertain the effect of coffee on the risk of MI. However, if the association between coffee and MI were different in the group of smokers compared to the group of non-smokers (effect modification), then such an analysis would be inappropriate." (Scott Evans)

Exposure = Coffee drinker (yes or no)
Outcome = MI (yes or no)
3rd variable = Smoking (yes or no)

To address these two questions in a meaningful way, we proceed as follows:

Step 1.
Analyze the 2x2 tables separately to obtain:
1. $OR_{smokers}$
2. $OR_{non\text{-}smokers}$

↓

Step 2.
Test equality of OR's across the separate tables

NULL: "Equal OR's"
"NO effect modification"

Possible tests are:
- **Woolf Test, or**
- **Breslow-Day-Tarone**

↓

Step 3a. Equal ORs NOT rejected → No effect modification *Stratum specific OR are assumed equal*	**Step 3b.** Equal ORs Rejected → Effect Modification *Stratum specific OR are different*
Ok to combine the data and perform an overall (combined) analysis. Test the null hypothesis of no association (Null: common OR=1) Test to use: Mantel-Haenszel Test	*NOT* ok to combine the data! Report the separate, stratum-specific OR

Step 2.

- Test of Null: This null hypothesis says: "The OR, whatever it is, is the same in every stratum". If we provisionally entertain as true the null hypothesis of equal OR's, then we need to obtain an estimate of this quantity. It is called the **Mantel-Haenzel OR.** How the tests work:
  Woolf test: Compares the **separate** ORs to the **Mantel-Haenzel OR**.
  Breslow-Day-Tarone test: Compares observed cell "a" count with null hypothesis expected.
  Note – The null hypothesis expected counts make use of the Mantel-Haenszel OR.
  We'll let the computer obtain the null hypothesis expected counts for us "behind the scenes".

Nature _____ **Population/ Sample** _____ **Observation/ Data** _____ **Relationships/ Modeling** _____ **Analysis/ Synthesis**

Step 3a.

- **IF** we conclude that the stratum specific odds ratios (OR) can be assumed equal,
  THEN we do a Test of Null: This null hypothesis says
  "The common OR is actually = 1, which corresponds to no association and independence"
  So, when we can assume the stratum specific ORs are the same, we can ask the follow-up question about its magnitude and direction using the **Mantel-Haenszel test**.

Step 3b.

- **BUT… IF** we judge (in step 2) that the stratum specific odds ratios (OR) are different,
  THEN we report stratum specific OR as we have discovered effect modification

Example – continued

Step 1.
Analyze the 2x2 tables *separately* to obtain:
1. **$OR_{smokers}$**
2. **$OR_{non\text{-}smokers}$**

R Illustration

```
# Smokers: Data entered column by column
smokers <- data.frame(MI=c(1011, 81), Control=c(390,77))
smokers

##     MI Control
## 1 1011     390
## 2   81      77

orsmokers <- fisher.test(smokers)
round(orsmokers$estimate,3)

## odds ratio
##      2.463

# NON-Smokers: Data entered column by column
nonsmokers <- data.frame(MI=c(383, 66), Control=c(365,123))
nonsmokers

##    MI Control
## 1 383     365
## 2  66     123

ornonsmokers <- fisher.test(nonsmokers)
round(ornonsmokers$estimate,3)

## odds ratio
##      1.954
```

INTERPRETATION OF OUTPUT:

AMONG SMOKERS:
The odds of MI among coffee drinkers is 2.463 times greater than the odds of MI among NON-coffee drinkers; whereas

AMONG NON-SMOKERS:
The odds of MI among coffee drinkers is 1.954 times greater than the odds of MI among NON-coffee drinkers.

Does this sample provide statistically significant evidence that these odds ratios (2.463 versus 1.954) are different from each other (which would suggest effect modification)?

Step 2.
Test equality of OR's across the separate tables

NULL: "Equal OR's"
"NO effect modification"

Possible tests are:
- **Woolf Test, or**
- **Breslow-Day-Tarone**

A. Woolf Test of Homogeneity

H_O: $OR_{Stratum\ 1} = OR_{Stratum\ 2} = \ldots OR_{Stratum\ (K-1)} = OR_{Stratum\ K}$ **"no interaction"**
H_A: At least one differs from the others **"there is interaction/effect modification"**

Step 1. For each stratum "i", obtain the following **"observed"** calculations:

$$\ln[\,OR_i\,] = \ln\left[\frac{a_i d_i}{b_i c_i}\right] \quad \text{and} \quad \text{Weight, } w_i = \left[\frac{1}{a_i}+\frac{1}{b_i}+\frac{1}{c_i}+\frac{1}{d_i}\right]^{-1}$$

Step 2. Obtain a weighted average of the stratum specific ln[OR] which is used to obtain **null hypothesis expected**:

$$\ln\bar{OR} = \frac{\sum_{i=1}^{K\ strata}(w_i)(\ln[OR_i])}{\sum_{i=1}^{K\ strata} w_i}$$

Step 3. The **Woolf statistic**, under the null hypothesis of homogeneity of OR, is distributed chi square with degrees of freedom = (# strata – 1)

$$\chi^2_{strata-1} = \sum_{i=1}^{Kstrata}(w_i)\left[\ln(OR)_i - \ln\bar{OR}\right]^2$$

$$= \sum_{i=1}^{Kstrata}(w_i)[\ln(OR)_i]^2 - \frac{\left\{\sum_{i=1}^{Kstrata}(w_i)[\ln(OR)_i]\right\}^2}{\sum_{i=1}^{Kstrata} w_i}$$

Example – continued

You might not want to, but you could do this in Excel:

Stratum/counts	a	b	c	d	w	ln(or)	w*ln(or)	w*[lnor - lnorbar]^2
Smokers	1011	390	81	77	34.61894927	0.901904747	31.22299469	0.466865763
Non-smokers	383	365	66	123	34.92560501	0.670667249	23.42345943	0.462766562
Totals =					69.54455428		54.64645412	0.929632324

$$\ln\bar{OR} = \frac{\sum_{i=1}^{Kstrata} (w_i)\,(\ln[OR_i])}{\sum_{i=1}^{Kstrata} (w_i)} = \frac{54.6465}{69.5445} = 0.7858$$

$$\text{Woolf } \chi^2_{\#\,strata-1} = \sum_{i=1}^{k\ strata} (w_i)\left[\ln(OR)_i - \ln\bar{OR}\right]^2 = 0.9296$$

Step 4. Significance level calculation.

p-value = Probability [Chi square (df=1) $\geq$ 0.9296] = .335

Do not reject. Assumption of the null hypothesis model and its application to the data has not led to an unlikely result. The separate OR's relating coffee drinking to MI for smokers (OR=2.46) and non-smokers (OR=1.96) are not statistically significantly different in this sample (p-value = .335)

B. Breslow-Day-Tarone Test of Homogeneity

Dear class - Doing this test "by hand" requires solving a quartic equation! No worries!!! We'll let the computer do this for us.

H_O: $OR_{Stratum\ 1} = OR_{Stratum\ 2} = \ldots OR_{Stratum\ (K-1)} = OR_{Stratum\ K}$ "no interaction"
H_A: At least one differs from the others "there is interaction/effect modification"

The **Breslow-Day-Tarone** statistic is another choice of test of homogeneity of odds ratios across levels of some third variable (strata). Its derivation involves solving a quartic equation and so we'll let the computer do the work for us!

The idea is the following:

Under the null hypothesis of homogeneity of odds ratio, each of the stratum specific odds ratios has the same expected value and that value is the common odds ratio.

The estimate of the common odds ratio that is used in this test is the Mantel-Haenszel odds ratio OR_{MH}. **Note – Other choices are possible but these give rise to other statistics**.

As before, for each stratum specific 2x2 table, the row totals $\mathbf{m_1}$ and $\mathbf{m_0}$ are assumed fixed and the column totals $\mathbf{n_1}$ and $\mathbf{n_0}$ are assumed fixed. Thus, ***only one cell count can vary*** and the one selected is the cell count **a**. Thus, the 2x2 table layout on page 32 now looks like:

	Case	Control	
Exposed	**a**	$b = (m_1-a)$	m_1
Not exposed	$c = (n_1-a)$	$d = m_0-n_1+a$	m_0
	n_1	n_0	T

The Breslow-Day-Tarone test utilizes fitted values for each cell of each 2x2 table so that the each "fitted" 2x2 table has odds ratio equal to OR_{MH}. The observed counts **"a"** are then compared to the null hypothesis fitted counts represented as **"A"** in a chi square statistic that has degrees of freedom equal to (number of strata) -1 = K-1.

$$\chi^2_{\text{Breslow-Day-Tarone}} = \sum_{i=1}^{K\text{strata}} \frac{\left[a_i - A_i(\text{using } OR_{MH})\right]^2}{\text{Var}(a_i;\text{null})}$$

Step 1. Obtain OR_{MH}

$$OR_{MH} = \frac{\sum_{i=1}^{Kstrata} a_i d_i / T_i}{\sum_{i=1}^{Kstrata} b_i c_i / T_i}$$

Step 2. For each stratum **"i"**, use the fixed values of n_1, m_1 and m_0 in the following quartic expression in the null hypothesis fitted value **"A_i"** and solve for **"A_i"**. **Again, we'll let the computer do this for us!!**

$$OR_{MH} = \frac{(A_i) / (n_{i1} - A_i)}{(m_{i1} - A_i) / (m_{i0} - n_{i1} + A_i)}$$

Then obtain the remaining fitted values "B", "C", and "D" for each stratum specific 2x2 table:

$$B_i = m_{i1} - A_i$$
$$C_i = n_{i1} - A_i$$
$$D_i = m_{i0} - n_{i1} + A_i$$

Step 3. For each stratum "i", obtain the null hypothesis variance of the observed count "a_i"

$$Var(a_i;null) = \left(\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} \right)^{-1}$$

Step 4. The **Breslow-Day-Tarone statistic**, under the null hypothesis of homogeneity of OR, is distributed chi square. With degrees of freedom = (# strata – 1)

$$\chi^2_{Breslow\text{-}Day\text{-}Tarone} = \sum_{i=1}^{Kstrata} \frac{[a_i - A_i(using\ OR_{MH})]^2}{Var(a_i;null)}$$

R Illustration – Woolf Test only.

Packages needed (ONE time installation)

```
install.packages("epiDisplay")
install.packages("foreign")
install.packages("survival")
```

Note – in the solution that follows, only one library( ) command is needed and that is to attach the package epiDisplay; the others get attached for you automatically.

```
library(epiDisplay)

## Loading required package: foreign
## Loading required package: survival
## Loading required package: MASS

## Loading required package: nnet

# Enter K=2 2x2 tables using command array(), one 2x2 table per row and entered column by column
# array(c(first2x2,2nd2x2, etc, kth2x2))
# dim=c(#rows in each table ,#columns in each table, #strata)
tablek2x2 <- array(c(1011, 81, 390, 77,                    # a=1011,  c=81,   b=390      d=77
                     383, 66, 365, 123),                   # a=383,   c=66,   b=365      d=123
                   dim=c(2,2,2),
                   dimnames = list(
                   COFFEE = c("Drinker", "Non-drinker"),
                   MI = c("MI-yes", "MI-no"),
                   STRATUM = c("Smokers", "NON-smokers")))


# CHECK - Display the k 2x2 tables
tablek2x2

## STRATUM = Smokers
##
##              MI
## COFFEE        MI-yes MI-no
##   Drinker       1011   390
##   Non-drinker     81    77
##
## STRATUM = NON-smokers
##
##              MI
## COFFEE        MI-yes MI-no
##   Drinker        383   365
##   Non-drinker     66   123

#  Woolf Test of Homogeneity of ORs across K=2 tables
#  mhor(mhtable=NAMEOFDATAFRAME, decimal=2, graph=FALSE, design="case control")
mhor(mhtable=tablek2x2,decimal=2, graph=FALSE,design="case control")

##
## Stratified analysis by STRATUM
##                       OR lower lim. upper lim.  P value
## STRATUM Smokers      2.46       1.74       3.49 1.97e-07
## STRATUM NON-smokers  1.95       1.39       2.77 6.27e-05
## M-H combined         2.18       1.72       2.76 4.07e-11
##
## M-H Chi2(1) = 43.58, P value = 0
## Homogeneity test, chi-squared 1 d.f. = 0.93 , P value = 0.334
```

INTERPRETATION: These data do NOT provide statistically significant evidence of "effect modification", that the OR is different, depending on smoking status (p-value = .33). Do not reject the null.

C. How to Estimate the Mantel-Haenszel Odds Ratio, OR_{MH}

The Mantel-Haenszel OR estimate of a common odds ratio is used in the Breslow-Day-Tarone Test of Homogeneity and in the Mantel-Haenszel Test of Overall Association. So, we need to know how to calculate it.

- OR_{MH} = a weighted average of the stratum specific odds ratios, where
- the weights are a function of the variances of the stratum specific odds ratios

Step 1. For each stratum, obtain the following:

	Case	Control	
Exposed	a	b	M_1
UNexposed	c	d	M_0
	N_1	N_0	T

$$OR_{stratum} = \frac{ad}{bc} \qquad \text{Variance}[OR_{stratum}] = \frac{bc}{T}$$

Step 2. Calculate the OR_{MH} as a weighted average of stratum specific OR.

Mantel Haenszel Odds Ratio (OR_{MH})

$$OR_{MH} = \frac{\sum_{strata}(weight_{stratum})OR_{stratum}}{\sum_{strata}(weight_{stratum})} = \frac{\sum_{strata} ad/T}{\sum_{strata} bc/T}$$

Note –
T = overall total for the stratum specific 2x2 table. Depending on the notation you prefer, use"

T = (a + b + c + d)

T = $n_{..}$

D. Mantel Haenszel Test of No Association

$$H_O: OR_{COMMON} = 1$$
$$H_A: OR_{COMMON} \neq 1$$

This test assumes that, after "step 2", you have concluded that the stratum specific OR are equal. The next question is: Given equality of the OR, is exposure independent of outcome? To put it another way: are the stratum specific odds ratios all unity?

Step 1. For each stratum, the null hypothesis model of no association says that each count "a" has probability distribution that is central hypergeometric. **For details, see Unit 2. Discrete Distributions, pp 25-29.**

	Case	Control	
Exposed	a	b	M_1
UNexposed	c	d	M_0
	N_1	N_0	T

$$E[a] = \frac{N_1 M_1}{T} \qquad \text{var}[a] = \frac{N_1 N_0 M_1 M_0}{T^2 (T-1)}$$

Step 2. The test statistic will be the sum, over strata of the counts "a".

$$\chi^2_{df=1} = \left[\frac{(A - E[A])^2}{\text{var}[A]} \right] \quad \text{where}$$

$$A = \sum_{strata} a \qquad E[A] = \sum_{strata} \left[\frac{N_1 M_1}{T} \right] \qquad \text{var}[A] = \sum \left[\frac{N_1 N_0 M_1 M_0}{T^2 (T-1)} \right]$$

Example – continued

A	B	C	D	E	F	G	H	I	J	K
smokers	1011	390	81	77	1092	467	1401	158	1559	3786689398
NON-smokers	383	365	66	123	449	488	748	189	937	821778984
A=total=	1394									
E [A]	n1	m1	T	(n1m1)/T						
smokers	1092	1401	1559	981.3291						
NON-smokers	449	748	937	358.4333						
E[A] = total =				1339.762						
Var [A]	a	n1	m1	T	n1	n0	m1	m0	T^2(T-1)	(n1n0m1m0)/(T^2(T-1))
smokers	1011	1092	1401	1559	1092	467	1401	158	3.8E+09	29.81089792
NON-smokers	383	449	748	937	449	488	748	189	8.2E+08	37.69420035
Var[A] =total=										67.50509827
Chi Square (df=1)=	43.57778									
MH Odds Ratio	T	a	d	b	c	ad/T	bc/T			
smokers	1559	1011	77	390	81	49.93	20.3			
NON-smokers	937	383	123	365	66	50.28	25.7			
totals =						100.2	46			
MH Odds Ratio =	2.179779									

$$A = \sum_{strata} a = 1394 \qquad E[A] = \sum_{strata} \left[\frac{N_1 M_1}{T} \right] = 1339.762$$

$$Var[A] = \sum_{strata} \left[\frac{N_1 N_0 M_1 M_0}{T^2 (T-1)} \right] = 67.5051$$

$$\chi^2_{df=1} = \left[\frac{(A - E[A])^2}{Var[A]} \right] = \left[\frac{(1394 - 1339.762)^2}{67.5051} \right] = 43.58$$

Significance Level (P-value)

p-value = Prob [Chi square w df=1 $\geq$ 43.58] < < .0001

Reject. The estimated MH Odds ratio = 2.18. Assumption of the null hypothesis and its application to the data has led to a highly unlikely outcome. Reject the null hypothesis. Conclude that, overall, data ***do*** suggest an association of coffee consumption with MI.

R Illustration

Happily … we have this already. See again page 35. I'm just highlighting a different portion of the same output.

```
#  Woolf Test of Homogeneity of ORs across K=2 tables
#  mhor(mhtable=NAMEOFDATAFRAME, decimal=2, graph=FALSE, design="case control")
mhor(mhtable=tablek2x2,decimal=2, graph=FALSE,design="case control")

##
## Stratified analysis by STRATUM
##                       OR lower lim. upper lim.  P value
## STRATUM Smokers      2.46       1.74       3.49 1.97e-07
## STRATUM NON-smokers  1.95       1.39       2.77 6.27e-05
## M-H combined         2.18       1.72       2.76 4.07e-11
##
## M-H Chi2(1) = 43.58 , P value = 0
## Homogeneity test, chi-squared 1 d.f. = 0.93 , P value = 0.334
```

INTERPRETATION: These data do DO PROVIDE statistically significant evidence of an association of higher coffee consumption with a greater odds of MI ($p < .0001$). REJECT the null. The estimated relative odds of MI is $OR_{MH} = 2.18$

9. The R x C Table – Test of (Monotone) Trend
Cochran-Armitage Test

So far, we have done a global test of the null hypothesis of "no association" in which the alternative is H_A: "association". The chi square test of association is a general test. The null hypothesis is the hypothesis of independence of the two variables and is rejected for any evidence of association, irrespective of its nature. A general test of no association ignores any ordering of the exposures if such exists.

H_O: No association between exposure and disease
H_A: Any association between exposure and disease (unspecified)

Now we are interested in trend in the setting where both the row and column variables are ordinal. For example, do people who smoke more packs of cigarettes per day tend to drink more alcohol?

Tip - The RxC table test of trend is ONLY appropriate when both the row and column variables are ordinal. When there is an ordering in the values of the outcome (eg – 0=no disease, 1=disease OR 0=no disease 1=mild disease and 2= advanced disease) ***and*** there is an ordering of the values of the predictor (such as "dose"), the RxC table test of trend makes use of this additional information in the data. Specifically, it is sensitive to the existence of a trend in outcome:

H_O: No association between exposure (or row variable) and disease (or column)
H_A: Linear association between exposure and disease.

Example.
Source: Tuyns AJ, Pequignot G and Jenson OM (1977) Le cancer de l'oesophage en Ille-et-Villaine en function des niveaux de consummation d'alcool et de tabac. *Bull Cancer* 64: 45-60.

The following are excerpted data from a **case-control** study of the relationship between alcohol consumption at 4 increasing levels ("doses") and case-control status for the disease of esophageal cancer.

	Alcohol Consumption (g/day)				
	0-39	40-79	80-119	120+	Total
Cases	29	75	51	45	200
Controls	386	280	87	22	775
Total	415	355	138	67	975

Tip **-** Because this is a **case-control** study design, we are focusing on the **odds ratio** measure of association. We are specifically interested in how the relative odds of esophageal cancer changes with increasing alcohol consumption.

Thus, there are at least two research questions:

1. **Test of general association**
   H_A: Does the odds of esophageal cancer **differ** by level of alcohol consumption?

2. (**Test of trend**)
   If the odds of esophageal cancer differs by level of alcohol consumption, then
   H_A: does the odds of esophageal cancer **increase** with increasing level of alcohol consumption?

The test to address question #1 would be addressed using the general test of association described in Section 4. See pp 10-15. The following R x C table test of trend would be used to address question #2.

The R x C Test of Trend

See also:
(1) Everitt, BS *The Analysis of Contingency Tables*. 1977. pp 53-54
(2) Bland, M. *An Introduction to Medical Statistics*. 2000. pp 243-245

The solution for the chi square test of trend in a R x C table is actually a test of **monotone trend**. It is related to the ideas of simple linear regression that you learned previously and that you will see again in Unit 5 of BIOSTATS 640. There are two distinctions, however

(1) Here, we are only interested in testing the null hypothesis of zero slope. We are not interested in confidence interval estimation of a population slope parameter value.

(2) Because of the assumption of fixed row totals and fixed column totals, the total degrees of freedom is the total sample size n, not (n-1).

Note to class - The following steps outline the idea of the test of trend.

Step 1: Assign ordered scores to the row variable values and ordered scores to the column variable values.

Row variable: Define a random variable Y that has values determined by the *ordered* categories of the row variable. In this example, we might use:

Y = 0 for controls
Y = 1 for cases

Column variable: Define a random variable X that has values determined by the *ordered* categories of the column variable. In this example, we might use:

X = 1 for alcohol consumption "0-39 g/day"
X = 2 for alcohol consumption "40-79 g/day"
X = 3 for alcohol consumption "80-119 g/day"
X = 4 for alcohol consumption "120+ g/day"

Tip!

- It doesn't matter which you call "X" and which you call "Y"
- *Mostly*, it doesn't matter what scores you use as long as they are equally spaced.

Step 2: Fit a simple linear regression model to the data (more on this in Unit 5).

- The straight-line model relating Y to X is given by:

$$Y = \beta_0 + \beta_1 X$$

- The estimate of the slope is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- The estimated standard error of the estimated slope is a little different in the setting of a test for trend in a contingency table for reasons (not shown) having to do with the row and column totals being fixed. The Everitt and Bland sources give slightly different formulae (one has division by n, the other by [n-1]. We'll use the Everitt formula:

$$\hat{se}\left(\hat{\beta}_1\right) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Step 3: For sufficiently large sample size n, the test of trend statistic is a chi square statistic with degrees of freedom, df = 1.

$$\chi^2_{DF=1} = \frac{\hat{\beta}_1^2}{\left[\hat{se}\left(\hat{\beta}_1\right)\right]^2}$$

A Special Case is the 2xC table

In this special case, the outcome defines the rows. The columns defined increased levels of the exposure.

To make things simpler to follow, we'll use the somewhat simpler notation for row and column totals (R's and C's) as follows.

Increasing Exposure is Columns, "j"

		$j=1$	⋯	$j=C$	
		x=1		**X=C**	
Outcome	**Row 1 Y=1**	n_{11}	…	n_{1C}	R_1
	Row 2 Y=0	n_{R1}	…	n_{RC}	R_2
		C_1	C_C	n	

Example - continued

	Alcohol Consumption (g/day)				
	0-39	40-79	80-119	120+	Total
Cases	29	75	51	45	$R_1 = 200$
Controls	386	280	87	22	$R_2 = 775$
Total	$C_1=415$	$C_2=355$	$C_3=138$	$C_4=67$	n = 975

$$\chi^2_{DF=1} = \left[\frac{n}{R_1 R_2}\right] * \frac{\left\{ n \sum_{j=1}^{C} n_{1j} X_j \; - \; R_1 \sum_{j=1}^{C} C_j X_j \right\}^2}{n \sum_{j=1}^{C} C_j X_j^2 \; - \; \left(\sum_{j=1}^{C} C_j X_j \right)^2}$$

Reference: Zar JH. *Biostatistical Analysis*, 5th Edition. 2010, pp 559-560

Evidence of a linear association is reflected in a large test statistic value with accompanying small p-value:

	Behavior of Chi Square Statistic	p-value
Null Hypothesis True. H_O: No Association	Value near = 1	Large Do "not" reject Null
Alternative Hypothesis True. H_A: Monotone Association	Value >> 1	Small "reject" Null

R Illustration

Packages needed (ONE time installation)

```
install.packages("DescTools")
install.packages("gmodels")
```

IMPORTANT

In the R illustration that follows, the order of the rows for CASES and CONTROL must be flipped. This is because R assumes that the 2nd row defines the event of interest (here, "Case")

```
# Test of Trend 2xC Table
library(DescTools)
library(gmodels)
table2x4 <- as.table(rbind(c(386,280,87,22),
                           c(29,75,51,45)))
dimnames(table2x4) <-list(y=c("Control","Case"),
                          Alcohol=c("0-39","40-79","80-119","120+"))
table2x4
##          Alcohol
## y         0-39 40-79 80-119 120+
##   Control  386   280     87   22
##   Case      29    75     51   45

CrossTable(table2x4,prop.t=FALSE,prop.r=FALSE,prop.c=TRUE)

##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Col Total |   this entry is the column percent
## |-------------------------|
##
##
## Total Observations in Table:  975
##
##
##              | Alcohol
##            y |      0-39 |     40-79 |    80-119 |      120+ | Row Total |
## -------------|-----------|-----------|-----------|-----------|-----------|
##      Control |       386 |       280 |        87 |        22 |       775 |
##              |     9.550 |     0.017 |     4.694 |    18.345 |           |
##              |     0.930 |     0.789 |     0.630 |     0.328 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##         Case |        29 |        75 |        51 |        45 |       200 |
##              |    37.007 |     0.065 |    18.191 |    71.085 |           |
##              |     0.070 |     0.211 |     0.370 |     0.672 |           |  % CASE increases w increased alcohol
## -------------|-----------|-----------|-----------|-----------|-----------|
## Column Total |       415 |       355 |       138 |        67 |       975 |
##              |     0.426 |     0.364 |     0.142 |     0.069 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|
##
CochranArmitageTest(table2x4,alternative="increasing")
##
##  Cochran-Armitage test for trend
##
## data:  table2x4
## Z = -12.375, dim = 4, p-value < 2.2e-16                                  R produces a Z-statistic.
## alternative hypothesis: increasing
```

INTERPRETATION. These data provide statistically significant evidence of a positive trend in the odds of esophageal cancer with increasing levels of alcohol consumption (p-value < .0001). Reject the null.

Tips -

- It is possible for the test of trend to achieve statistical significance when the general test of association does NOT achieve statistical significance.

- In this setting, additional analyses might be needed to address the possible confounding effect of age and tobacco use!

Tip for R Users -

- If you use this approach, take care to order the rows of your table appropriately.

10. The Chi Square Goodness of Fit Test

So far you have been introduced to two chi square tests:

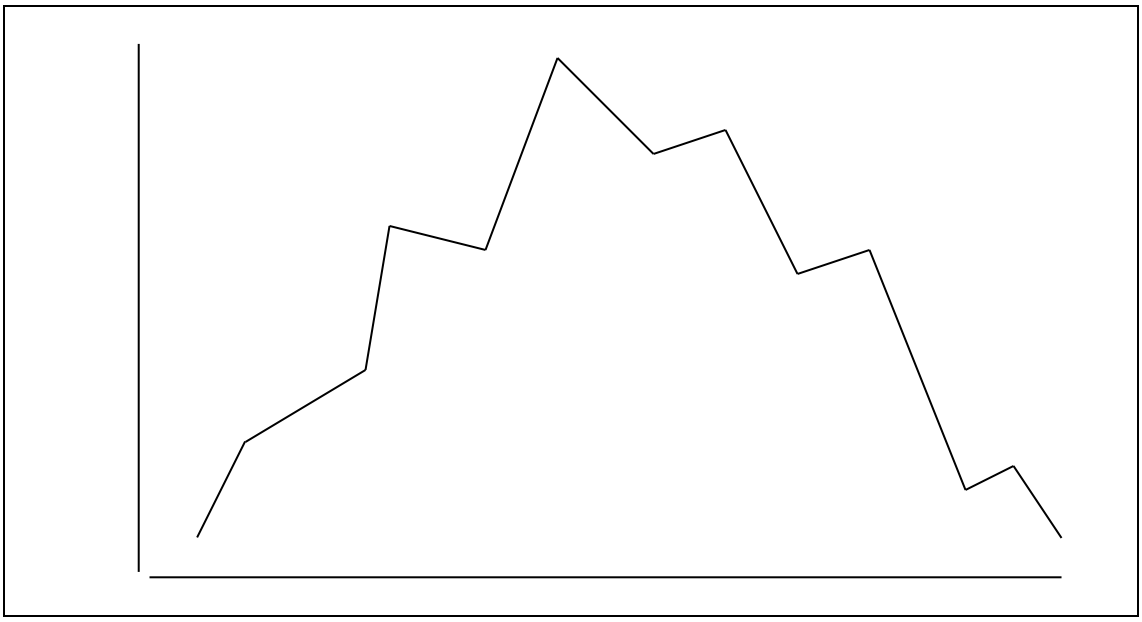
- Introductory Biostatistics:
  The one sample test of the variance σ^2 of a Normal distribution for modeling continuous data.

- BIOSTATS 640:
  The chi square test of no association for a contingency table of count data

Here is a 3rd chi square test!

- Next up: A chi square test (called the chi square "goodness-of-fit" test) is used to to assess whether two distributions can be reasonably assumed to be the same

Suppose that a histogram of the observed data looks like the following snarly-ness.



Suppose we want to know whether we can reasonably assume that the data represent a sample from a Normal distribution.

Many analyses make the assumption that the data are distributed normal.

Note – It doesn't have to be the Normal distribution. The chi square goodness of fit can also be used to assess the reasonableness of assuming that the data are distributed according to some other distribution (e.g., Binomial or Poisson).

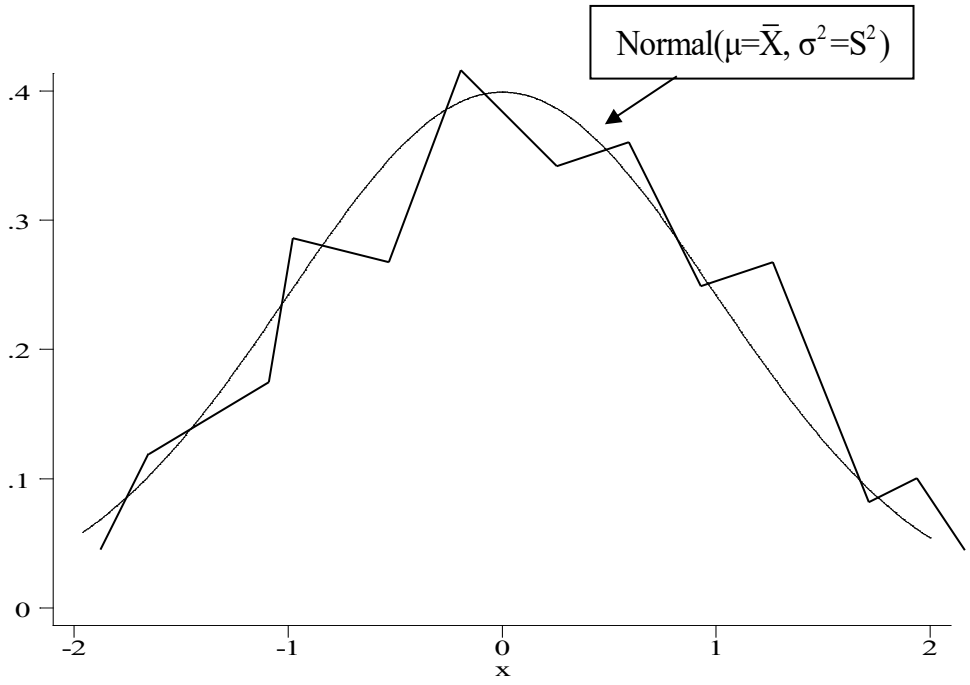
HOW TO: Construct a chi square goodness-of-fit test where interest is in goodness-of-fit to the Normal distribution.

Null Hypothesis. The null hypothesis is that our data are a simple random sample from a Normal distribution. But which Normal distribution? *Answer*: a good choice is the Normal distribution that is the "closest". But which Normal distribution is the "closest"? *Answer*: The "closest" Normal distribution has mean and variance parameter values that match our sample mean and variance values:

$$\mu_{NULL} = \text{sample mean} = \bar{X}$$
$$\sigma^2_{NULL} = \text{sample variance} = S^2$$

H_O: The data are a simple random sample from a Normal distribution
H_A: Not.

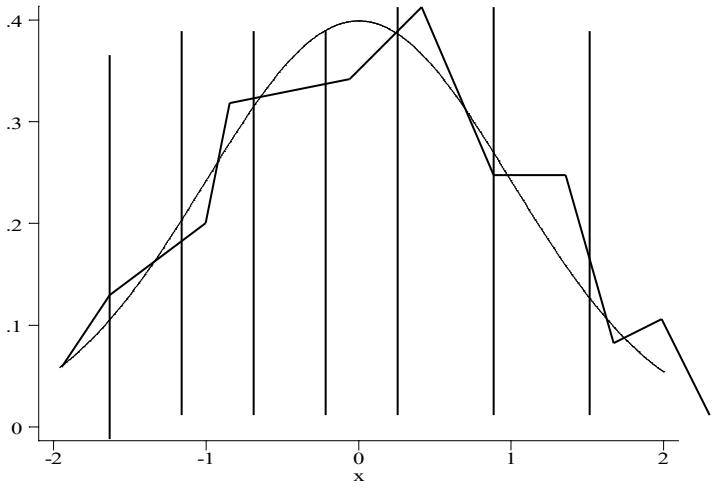
We overlay the null hypothesis ("closest") Normal distribution on top of our snarly looking histogram of observed data.



Steps in the Calculation of the Chi Square Goodness-of-Fit Test.

Step 1:
Divide up the range of the observed data into a total of K intervals. Index these intervals using i=1, 2, …, K.

Tip!
Make sure that the intervals span the entire real axis (-∞ to +∞)



Interval i=1 i=2 i=3 etc i=K

Step 2:
In each interval "i", obtain

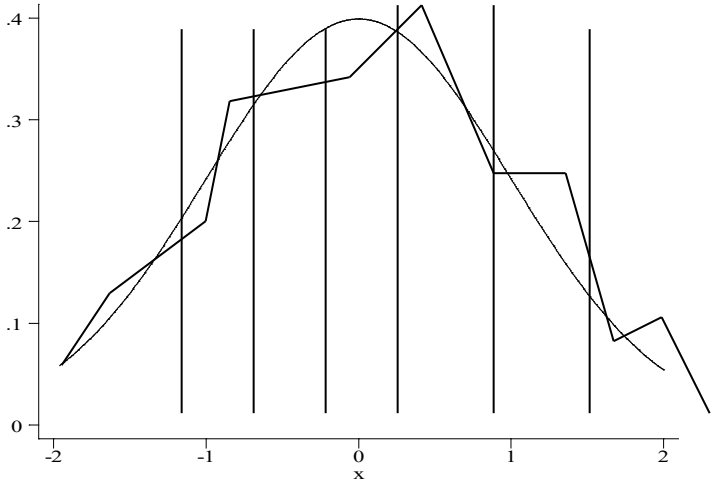
Observed count = O_i

Expected count = E_i

Also obtain for each interval "i" the following "component" chi square.

$$\frac{(O_i - E_i)^2}{E_i}$$

Notice: Each component chi square is a comparison of the observed and expected counts.



Observed O_1 O_2 etc O_K

Expected E_1 E_2 etc E_K

$$\frac{(O_1 - E_1)^2}{E_1} \quad \frac{(O_2 - E_2)^2}{E_2} \quad \text{etc} \quad \frac{(O_K - E_K)^2}{E_K}$$

Step 3:
Sum these to obtain the Chi square Goodness of Fit Test.

$$\chi^2_{gof} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

Behavior of the Chi Square Goodness-of-Fit Statistic

NOTE! This is a setting where, typically, we do NOT want to reject the null hypothesis.

The null hypothesis says that the "unknown true" (the distribution that gave rise to the data) is reasonably similar to the hypothesized (in this example, Normal).

Values of the chi square goodness of fit test will be small when the two distributions are reasonably similar. This is because the observed and expected counts are similar, giving rise to component chi square values that are small.

Degrees of freedom.
Degrees of Freedom $_{\text{CHI SQUARE GOF}}$ = (# intervals) - (1) - (# parameters estimated using data)

Example – (*Note - in practice, we let the computer do this for us)*
Source: Rosner, B. Fundamentals of Biostatistics, second edition. Boston: Duxbury, 1986 p. 352

Suppose you have n=14,736 blood pressure readings. Goal: Test the goodness of fit of the distribution of these observed data to the "closest" fitting Normal distribution. Suppose further that, magically, someone has already given you the values of the sample mean and variance: $\bar{X}=80.68$ and $S^2=12^2$. **(Note- it's not possible to do these calculations from the data in grouped form)**

Step #1. Construct a histogram of your observed data and obtain the observed counts in each interval.

i	Class Interval	Observed Count, O_i
1	<50	57
2	≥ 50 to < 60	330
3	≥ 60 to < 70	2132
4	≥ 70 to < 80	4584
5	≥ 80 to < 90	4604
6	≥ 90 to < 100	2119
7	≥ 100 to < 110	659
8	≥ 110	251
	TOTAL	14,736

Tip: Check that the sum of the observed counts MATCHES the total sample size.

Step #2. Obtain the μ and σ^2 of the "closest fitting" normal distribution by matching them to your sample mean and sample variance. For these data, we have

$$\bar{X} = 80.68$$

$$S^2 = 12^2$$

So, we'll compare the data to the normal distribution with

$$\mu = 80.68$$

$$\sigma^2 = 12^2$$

Step #3. Calculate the likelihood of a value in each interval using the z-score method.

For interval i=1:

$$\Pr[X<50]=\Pr[Z<\frac{50-80.68}{12}] = \Pr[Z<-2.556]=0.00529$$

For interval i=2:

$$\Pr[50<X<60]=\Pr[\frac{50-80.68}{12} < Z<\frac{60-80.68}{12}] = \Pr[-2.556<Z<-1.7233]=.04242-.00529=.0371$$

etc.

For interval i=K=8:

$$\Pr[X>110]=\Pr[Z>\frac{110-80.68}{12}] = \Pr[Z>+2.4433]=0.00728$$

Step #4 Calculate the expected count of observations in each interval using

Expected count = (sample size) x (probability of interval)

Important – Do ***not*** round these expected values to the nearest whole integer (rounding errors have a nasty way of accumulating and eventually "adding up").

For interval i=1:

$E_1 = (14{,}736)\ [0.00529] = 77.95$

For interval i=2:

$E_2 = (14{,}736)\ [0.0371] = 546.71$

Etc.

For interval i=K=8:

$E_8 = (14{,}736)\ [0.00728] = 107.28$

Step #5 Retrieve the "observed" counts from step #1 and put these into your table:

i	Class Interval	Observed Count, O_i	Expected Count, E_i	Component $\frac{(O_i - E_i)^2}{E_i}$
1	<50	57	77.86	5.5912
2	≥ 50 to < 60	330	547.15	86.1808
3	≥ 60 to < 70	2132	2126.68	0.0133
4	≥ 70 to < 80	4584	4283.35	21.1029
5	≥ 8 to < 90	4604	4478.52	3.5157
6	≥ 90 to < 100	2119	2431.13	40.0734
7	≥ 100 to < 110	659	684.09	0.9199
8	≥ 110	251	107.22	191.8006
	TOTAL	**14,736** ⟷	**14,736**	**350.2**

Tip: Check that sum of observed = sum of expected = sample size!

Step #6. Determine degrees of freedom, df

df = [K] - [1] - [# parameters estimated]
= [8] – [1] - [1 for μ] – [1 for σ]
= 5

Step #7 Assess statistical significance

$\chi^2_{\text{goodness of fit; df=5}} = 350.2$

p-value = Prob [Chi square w df=5 $\geq$ 350.2] << 0.0001

A p-value << .0001 suggests that the data cannot reasonably be assumed to follow a normal distribution. The null hypothesis of "goodness-of-fit" is rejected. Examination of the component chi-squares suggests that the normal distribution fit is reasonable for blood pressures between 60 and 110 mm Hg but is poor for readings below 60 mm Hg or above 110 mm Hg.

R Illustration (not the only solution; let me know if you find something better please!)

```
options(scipen=1000)                                                    # Turn off scientific notation
i <- c(1,2,3,4,5,6,7,8)                                                 # Interval id
a <- c(-999999999,50,60,70,80,90,100,110)                               # a = lower interval bound
b <- c(50,60,70,80,90,100,110,999999999)                                # b = upper interval bound
obs <- c(57,330,2132,4584,4604,2119,659,251)                            # observed counts
temp <- data.frame(i,a,b,obs)                                           # combine to make data frame
xbar <- 80.68                                                           # Null hypothesis mean of normal = sample mean
s <- 12                                                                 # Null hypothesis sd = sample sd
ntrials <- 14736                                                        # total sample size


temp$prob <- pnorm(temp$b,mean=xbar,sd=s)-pnorm(temp$a,mean=xbar,sd=s)  # Null hypothesis interval probabilities
temp$exp <- ntrials*temp$prob                                           # Null hypothesis expected counts
temp$ichisq <- ((temp$obs-temp$exp)^2)/(temp$exp)                       # Component chi square

temp[ , c("i","obs","prob","exp","ichisq")]                             # Show every row and selected columns
##   i  obs        prob        exp       ichisq
## 1 1   57 0.005284022   77.86534   5.59122309
## 2 2  330 0.037130110  547.14930  86.18089937
## 3 3 2132 0.144318811 2126.68201   0.01329821
## 4 4 4584 0.290672421 4283.34879  21.10291587
## 5 5 4604 0.303916907 4478.51955   3.51574760
## 6 6 2119 0.164978801 2431.12761  40.07343909
## 7 7  659 0.046422783  684.08614   0.91993423
## 8 8  251 0.007276145  107.22127 192.80058836

paste("Chi Square Goodness of Fit (Null: Distribution is Normal)")
## [1] "Chi Square Goodness of Fit (Null: Distribution is Normal)"

chisqvalue <- sum(temp$ichisq)
dfvalue <- length(temp$ichisq) - 3
pvalue <- pchisq(chisqvalue,df=dfvalue,lower.tail=FALSE)
pvalue <- round(pvalue,8)
paste("Chi Square statistic =  ", chisqvalue)
## [1] "Chi Square statistic =   350.198045828432"
paste("Degrees of freedom = ", dfvalue)
## [1] "Degrees of freedom =  5"
paste("p-value =", pvalue)
## [1] "p-value = 0"
```

Example –

Source: Zar, JH. Biostatistical Analysis, third edition. Upper Saddle River: Prentice Hall, 1996 p. 461

A plant geneticist wishes to know if a sample of n=250 seedlings come from a population having a 9:3:3:1 ratio of yellow smooth: yellow wrinkled: green smooth: green wrinkled seeds.

In this example, expected counts are computed using the hypothesized phenotype ratios.

i	Phenotype	O_i	Expected Count, E_i	Component $\frac{(O_i - E_i)^2}{E_i}$
1	Yellow smooth	152	$(n)[\Pr(\text{phenotype})]=(n)\left[\frac{9}{9+3+3+1}\right]=(250)[.5625]=140.625$	0.9201
2	Yellow wrinkled	39	$(n)[\Pr(\text{phenotype})]=(n)\left[\frac{3}{9+3+3+1}\right]=(250)[.1875]=46.875$	1.3230
3	Green smooth	53	$(n)[\Pr(\text{phenotype})]=(n)\left[\frac{3}{9+3+3+1}\right]=(250)[.1875]=46.875$	0.8003
4	Green wrinkled	6	$(n)[\Pr(\text{phenotype})]=(n)\left[\frac{1}{9+3+3+1}\right]=(250)[.0625]=15.625$	5.9290
	TOTAL	**250**	**250**	**8.972**

DF = [K] - [1] - [# parameters estimated]
= [4] – [1] - [0, because we didn't have to estimate any!]
= 3

$\chi^2_{\text{goodness of fit; df=3}} = 8.972$

p-value = Prob [Chi square w df = 3 $\geq$ 8.972] = 0.02967

Reject the null. This suggests that the data do NOT come from a population having a 9:3:3:1 ratio of the four seedling types.

1. Appendix A
The Chi Square Distribution

In your introductory biostatistics course, you might have been introduced to the chi square distribution in the setting of tests of the variance parameter of a Normal distibution, which is a model for continuous outcomes. You might have also been introduced to the chi square distribution for tests of counts in a contingency table.

This appendix explains the appropriateness of using the chi square distribution (a model for a continuous random variable) for the analysis of discrete data.

The chi square distribution is related to the normal distribution:

If	Then	has a Chi Square Distribution with df =
Z has a distribution that is Normal (0,1)	Z^2	1
X has a distribution that is Normal (μ, σ^2), so that $Z\text{-score} = \frac{X-\mu}{\sigma}$	$\{ Z\text{-score} \}^2$	1
$X_1, X_2, \ldots, X_n$ are each distributed Normal (μ, σ^2) and are independent, so that $\overline{X}$ is Normal (μ, σ^2/n) and $Z\text{-score} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$\{ Z\text{-score} \}^2$	1
$X_1, X_2, \ldots, X_n$ are each distributed Normal (μ, σ^2) and are independent and we calculate $S^2 = \frac{\sum_{i=1}^{n}(X-\overline{X})^2}{n-1}$	$\frac{(n-1)S^2}{\sigma^2}$	(n-1)

The chi square distribution can be used in the analysis of categorical (count) data for reasons related to the normal distribution and the central limit theorem:

$Z_1, Z_2, \ldots, Z_n$ are each Bernoulli with probability of event = p.

$$E[Z_i] = \mu = p$$
$$Var[Z_i] = \sigma^2 = p(1-p)$$

↓

1. The net number of events $X = \sum_{i=1}^{n} Z_i$ is Binomial (N,p)

2. Tthe distribution of the average of the Z_i is well modeled using the Normal(μ, σ^2/n).

Apply this notion here: By convention,

$$\overline{Z} = \frac{\sum_{i=1}^{n} Z_i}{n} = \frac{X}{n} = \overline{X}$$

↓

3. So perhaps the distribution of the sum is also well modeled using the Normal. At least approximately

If $\overline{X}$ is modeled well as Normal (μ, σ^2/n)
Then $X = n\overline{X}$ is modeled well as Normal ($n\mu$, $n\sigma^2$)

↓

Exactly: X is distributed Binomial(n,p)
Approximately: X is distributed Normal ($n\mu$, $n\sigma^2$)

Where: $\mu = p$ and $\sigma^2 = p(1-p)$

Putting it all together …

If	Then	Comment
X has a distribution that is Binomial (n,p) exactly	X has a distribution that is Normal ($n\mu$, $n\sigma^2$) approximately, where $\mu = p$ $\sigma^2 = p(1-p)$ ↓	
	$Z\text{-score} = \frac{X - E(X)}{SD(X)}$ $= \frac{X - n\mu}{\sqrt{n}\sigma}$ $= \frac{X - np}{\sqrt{np(1-p)}}$ is approx. Normal (0,1) ↓	
	{ Z-score }^{2} has distribution that is modeled well as a Chi Square random variable.	***We arrive at a continuous distribution model for (discrete) count data!!***

A Feel for things … continued –

You will come to think of the chi square distribution as this when analyzing count data:

For one cell

$$\frac{\left[\begin{array}{l}\text{Observed} \\ \text{Count}\end{array} - \begin{array}{l}\text{Expected} \\ \text{Count}\end{array}\right]^2}{\text{Expected Count}}$$ is distributed Chi Square (df = 1) approximately.

For the sum of all RC cells in a R x C table

$$\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{\left[\begin{array}{l}\text{Observed} \\ \text{Count (i,j)}\end{array} - \begin{array}{l}\text{Expected} \\ \text{Count (i,j)}\end{array}\right]^2}{\text{Expected Count (i,j)}}$$ is distributed Chi Square, df = (R-1) x (C-1), approx.

Appendix B
Selected Models for Categorical Data

Various study designs (e.g. – case control, cohort, surveillance) give rise to categorical data, utilizing some of the probability distributions that have been introduced in Unit 2 (eg – Binomial, Poisson, Product Binomial, and Product Poisson).

#1. Cohort: The random variables are 2 counts of events of <u>disease</u>

	Disease	Not	
Exposed	a	b	FIXED
Unexposed	c	d	FIXED

→

The count "a" is distributed Binomial { # trials = a+b, $Prob_{exposed}$ [disease] }
The count "c" is distributed Binomial { # trials = c+d, $Prob_{unexposed}$ [disease] }

#2. Case-Control: The random variables are 2 counts of events of <u>exposure</u>

	Case	Control
Exposed	a	b
Not	c	d
	FIXED	FIXED

↓

The count "a" is distributed Binomial { # trials = a+c, $Prob_{case}$ [exposed] })

The count "b" is distributed Binomial { # trials = b+d, $Prob_{control}$ [exposed] }

#3. 2x2 Table: The random variable is one count, the count of the events of joint occurrence of <u>exposure and disease</u>

	Disease	Not	
Exposed	a	b	FIXED
Not	c	d	FIXED
	FIXED	FIXED	

The count "a" is distributed Hypergeometric

#4. 2x2 Table: The random variables are 4 separate, independent, counts of events, each with its own probability distribution model

	Disease	Not
Exposed	a	b
Not	c	d

The count "a" is distributed Poisson (μ_a)
The count "b" is distributed Poisson (μ_b)
The count "c" is distributed Poisson (μ_c)
The count "d" is distributed Poisson (μ_d)

#5. RxC Table – General

	Mild	Moderate	Severe	
Exposed	a	b	c	FIXED
Not	d	e	f	FIXED

The triplet of counts (a,b,c) is distributed Multinomial
The triplet of counts (d,e,f) is distributed Multinomial

Note – The Multinomial distribution is not discussed in this course. It is an extension of the Binomial distribution to the setting of more than two outcomes.

Appendix C
Concepts of Observed versus Expected

In categorical data analysis methodology, we compare observed counts of events with null hypothesis expected counts of events. *(Emphasis on "counts")*

Consider an investigation of a possible association between electronic fetal monitoring (EFM) and delivery by caesarian section:

		Caesarian Section: Yes	Caesarian Section: No	
EFM Exposure	Yes	5	1	6
	No	2	7	9
		7	8	15

The observed counts are:

with EFM exposure=yes AND Caesarian section=yes: 5
with EFM exposure=yes AND Caesarian section=no: 1
with EFM exposure=no AND Caesarian section=yes: 2
with EFM exposure=no AND Caesarian section=no: 7

The expected counts depend on what we believe.
Suppose, for starters, that ***we have no hypothesis one way or the other***.

Cohort Study: Suppose we allow for possibility of different probabilities of caesarian section for EFM exposed women versus non-EFM exposed women.

Best guess of pr [caesarian section] for EFM exposed women = 5/6
Best guess of pr [caesarian section] for non-EFM exposed women = 2/9

Case-Control Study: Suppose we allow for possibility of different probabilities of history EFM exposure caesarian section women versus non caesarian section women.

Best guess of pr [EFM history] for C-section women = 5/7
Best guess of pr [EFM history] for non C-section women = 1/8

Now suppose we consider the ***null hypothesis of "Independence", "No Association", "Homogeneity"***

$$\text{Expected}_{\text{HO true}} = \frac{(\text{row total})(\text{column total})}{(\text{grand total})}$$

Example – Null Hypothesis Expected Count in a Cohort Study

Viewed as a *cohort study,* the outcome is "caesarian section". The null hypothesis of "independence", "no association", "homogeneity of proportions" suggests that

Best Guess of pr [caesarian section]: Overall proportion of c-section =

$$\hat{p}_{\text{c-section}} = \frac{7}{15} = \frac{\text{column "yes" total}}{\text{grand total}}$$

Best Guess of pr [NO caesarian section]: Overall proportion of NON-c-section =

$$\hat{p}_{\text{NON c-section}} = \frac{8}{15} = \frac{\text{column "no" total}}{\text{grand total}}$$

Caesarian Section

efm	Yes	No
Yes	$(n_{\text{efm=yes}})\hat{p}_{\text{c-section}} = (6)\left[\frac{7}{15}\right]$ $= (\text{row "yes" total})[\frac{\text{column "yes" total}}{\text{grand total}}]$ $= \frac{(\text{row "yes" total})(\text{column "yes" total})}{\text{grand total}}$	$(n_{\text{efm=yes}})\hat{p}_{\text{NO c-section}} = (6)\left[\frac{8}{15}\right]$ $= (\text{row "yes" total})[\frac{\text{column "no" total}}{\text{grand total}}]$ $= \frac{(\text{row "yes" total})(\text{column "no" total})}{\text{grand total}}$
No	$(n_{\text{efm=no}})\hat{p}_{\text{c-section}} = (9)\left[\frac{7}{15}\right]$ $= (\text{row "no" total})[\frac{\text{column "yes" total}}{\text{grand total}}]$ $= \frac{(\text{row "no" total})(\text{column "yes" total})}{\text{grand total}}$	$(n_{\text{efm=no}})\hat{p}_{\text{NO c-section}} = (9)\left[\frac{8}{15}\right]$ $= (\text{row "no" total})[\frac{\text{column "no" total}}{\text{grand total}}]$ $= \frac{(\text{row "no" total})(\text{column "no" total})}{\text{grand total}}$

Expected Counts Under "Independence", "No Association", "Homogeneity"

$$= \frac{(\text{row total})(\text{column total})}{(\text{grand total})}$$

Example – Null Hypothesis Expected Count in a Case-Control Study

Viewed as a *case-control study,* the outcome is "history EFM exposure". The null hypothesis of "independence", "no association", "homogeneity of proportions" suggests that

Best Guess of pr[hx EFM]: Overall proportion of EFM exposure =

$$\hat{p}_{\text{hx EFM}} = \frac{6}{15} = \frac{\text{row "yes" total}}{\text{grand total}}$$

Best Guess of pr[hx NO EFM]: Overall proportion of NO EFM exposure =

$$\hat{p}_{\text{NO EFM}} = \frac{9}{15} = \frac{\text{row "no" total}}{\text{grand total}}$$

		Caesarian Section	
		Yes	No
efm	**Yes**	$(n_{\text{c-section=yes}})\hat{p}_{\text{hx EFM}} = (7)\left[\frac{6}{15}\right]$ $= (\text{column "yes" total})[\frac{\text{row "yes" total}}{\text{grand total}}]$ $= \frac{(\text{column "yes" total})(\text{row "yes" total})}{\text{grand total}}$	$(n_{\text{c-section=no}})\hat{p}_{\text{hx EFM}} = (8)\left[\frac{6}{15}\right]$ $= (\text{column "no" total})[\frac{\text{row "yes" total}}{\text{grand total}}]$ $= \frac{(\text{column "no" total})(\text{row "yes" total})}{\text{grand total}}$
	No	$(n_{\text{c-section=yes}})\hat{p}_{\text{NO hx EFM}} = (7)\left[\frac{9}{15}\right]$ $= (\text{column "yes" total})[\frac{\text{row "no" total}}{\text{grand total}}]$ $= \frac{(\text{column "yes" total})(\text{row "no" total})}{\text{grand total}}$	$(n_{\text{c-section}})\hat{p}_{\text{NO hx EFM}} = (8)\left[\frac{9}{15}\right]$ $= (\text{column "no" total})[\frac{\text{row "no" total}}{\text{grand total}}]$ $= \frac{(\text{column "no" total})(\text{row "no" total})}{\text{grand total}}$

Observed and Expected Counts
General R x C Table

A useful notation is "O" for observed and "E" for expected and the following subscripts:

O_{ij} = Observed count in row "i" and column "j"
E_{ij} = Expected count in row "i" and column "j"
$O_{i.} = E_{i.} = n_{i.}$ = Observed and Expected row total for row "i"
$O_{.j} = E_{.j} = n_{.j}$ = Observed and Expected column total for column "j"

Yes, it's true … Under the null hypothesis, the expected and observed totals (row totals, column totals, grand total) match!

Observed Counts

		Columns, "j"			
		$j=1$	…	$j=C$	
Rows, "i"	$i=1$	O_{11}	…	O_{1C}	$N_{1.} = O_{1.}$
	…	…			
	$i=R$	O_{R1}	…	O_{RC}	$N_{R.} = O_{R.}$
		$N_{.1} = O_{.1}$	…	$N_{.C} = O_{.C}$	$N=O_{..}$

Expected Counts under Null: "Independence, No Association, Homogeneity"

		Columns, "j"			
		$j=1$	…	$j=C$	
Rows, "i"	$i=1$	$E_{11} = \frac{n_{1.}n_{.1}}{n_{..}}$	…	$E_{1C} = \frac{n_{1.}n_{.C}}{n_{..}}$	$N_{1.} = O_{1.}$
	…	…			
	$i=R$	$E_{R1} = \frac{n_{R.}n_{.1}}{n_{..}}$	…	$E_{RC} = \frac{n_{R.}n_{.C}}{n_{..}}$	$N_{R.} = O_{R.}$
		$N_{.1} = O_{.1}$	…	$N_{.C} =$	$N=O_{..}$

Appendix D
Review: Measures of Association

Recall that various epidemiological studies (prevalence, cohort, case-control) give rise to a 2x2 table of count data. The two rows index exposure (exposed versus not exposed) and the two columns index case status (disease versus healthy)

Recall also the convention of representing the counts with the notation "a", "b", "c", and "d":

	Disease	Healthy	
Exposed	a	b	a+ b
Not Exposed	c	d	c + c
	a+ c	b + d	

Let's consider some actual counts:

	Disease	Healthy	
Exposed	a =2	b = 8	10
Not Exposed	c = 10	d = 290	300
	12	298	310

We might have more than one 2x2 table if the population of interest is partitioned into subgroups or strata.

Example: Stratification by gender would yield a separate 2x2 table for men and women.

A good measure of an exposure-outcome association is a single measure that is stable over the different levels (strata) of other characteristics of the population (unless there is effect modification which we want to discover!).

Excess Risk

Suppose that the cumulative incidence of disease among exposed = π_1
and that the cumulative incidence of disease among non-exposed = π_0

Excess Risk b: The difference between the cumulative incidence rates

$$b = (\pi_1 - \pi_0)$$

Example: In our 2x2 table, we have $\pi_1 = 2/10 = .20$, $\pi_0 = 10/300 = .0333$
Thus, $b = (.20 - .0333) = .1667$

- The effect of exposure is said to be additive because we can write

$$\pi_1 = \pi_0 + b$$

- Hypothesis testing focuses on H_O: $b = 0$

- For a population that has been stratified with strata k = 1 ··· K, the additive model says that

$$\pi_{k1} = \pi_{k0} + b$$

Note: The absence of a subscript "k" on the excess risk = b says that we are assuming that the excess risk is constant in every stratum (e.g., among men and women).

- Biological mechanisms which relate exposure to disease in an additive model often do not operate in the same way across strata.

- If so, the additive risk model does not satisfy our criterion of being stable.

Relative Risk (RR)

The relative risk is the ratio of the cumulative incidence rate of disease among the exposed, π_1, to the cumulative incidence rate of disease among the non-exposed, π_0.

Relative Risk, RR: The ratio of the cumulative incidence rates

$$RR = \pi_1 / \pi_0$$

Example: In our 2x2 table, we have $\pi_1 = 2/10 = .20$, $\pi_0 = 10/300 = .0333$
Thus, $RR = .20 / .0333 = 6.006$

- The effect of exposure is said to be multiplicative because we can write

$$\pi_1 = [\pi_0] RR$$

- Hypothesis testing focuses on H_O: $RR = 1$

- This model is also said to be additive on the log scale.
  It is also said to be an example of a log-linear model.

  To see this:
  $$\pi_1 = \pi_0 RR \Rightarrow$$
  $$\ln[\pi_1] = \ln[\pi_0] + \ln[RR] \Rightarrow$$
  $$\ln[\pi_1] = \ln[\pi_0] + \beta \quad \text{where} \quad \beta = \ln[RR].$$

- Amazing. It has been found empirically that many exposure-disease relationships vary with age in such a way that a log linear model is a good fit. Specifically, the change with age in the relative risk of disease with exposure is reasonably stable. In such instances, the model is preferable to the additive risk model.

Attributable Risk

The attributable risk is proportion of the incidence of disease among exposed persons that is in excess of the incidence of cases of disease among non-exposed persons. Often, it is expressed as a percent.

Attributable Risk

$$AR = \frac{\pi_1 - \pi_0}{\pi_1}$$ when expressed as a percent.

Recalling that $RR = \pi_1 / \pi_0$ reveals that

$$AR = \frac{RR - 1}{RR}$$

Example: In our 2x2 table, a RR = 6.006 yields an attributable risk value of

$$AR = (6.006 - 1) / 6.006 = .8335 = 83.35\%$$

Odds Ratio

Recall that the odds ratio measure of association has some wonderful advantages, both biological and analytical. Recall first the meaning of an "odds":

* Probability[event] = π
* Odds[Event] = $\pi/(1-\pi)$

Let's look at the odds that are possible in our 2x2 table:

	Disease	Healthy	
Exposed	a	b	a+ b
Not Exposed	c	d	c + d
	a+ c	b + d	

Cohort study design:

Estimated Odds of disease among exposed = $\left[\frac{a/(a+b)}{b/(a+b)}\right] = \frac{a}{b} = \frac{2}{8} = .25$

Estimated Odds of disease among non-exposed = $\left[\frac{c/(c+d)}{d/(c+d)}\right] = \frac{c}{d} = \frac{10}{290} = .0345$

Case-control study design

Estimated Odds of exposure among diseased = $\left[\frac{a/(a+c)}{c/(a+c)}\right] = \frac{a}{c} = \frac{2}{10} = .20$

Estimated Odds of exposure among healthy = $\left[\frac{b/(b+d)}{d/(b+d)}\right] = \frac{b}{d} = \frac{8}{290} = .0276$

Odds ratio

Cohort study design

$$OR = \frac{\textbf{Odds disease among exposed}}{\textbf{Odds disease among non-exposed}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Case-control study design

$$OR = \frac{\textbf{Odds exposure among disease}}{\textbf{Odds exposure among healthy}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Terrific!

The OR is the same = (ad)/(bc), regardless of the study design, cohort (prospective) or case-control (retrospective)

Example: In our 2x2 table, a =2, b=8, c=10, and d=290 so the OR = 7.25. This is slightly larger than the value of the RR = 6.006.

Thus, there are advantages of the Odds Ratio, OR.

1. Many exposure disease relationships are described better using ratio measures of association rather than difference measures of association

2. $OR_{\text{cohort study}} = OR_{\text{case-control study}}$

3. The OR is the appropriate measure of association in a case-control study.

    - Note that it is not possible to estimate an incidence of disease in a retrospective study. This is because we select our study persons based on their disease status.

4. When the disease is rare $OR_{\text{case-control}} = RR$.

Appendix E.
Review: Confounding of Rates

Is our estimate of a disease-exposure relationship measuring what we think it is? Or is there some other influence that plays a role? The presence of other influences might be as confounders or effect modifiers. A confounded association does not tell us about the association of interest.

A confounded relationship is biased because of an extraneous variable. An effect modified relationship changes with variations in the extraneous variable. Several examples illustrate these ideas.

Example: Among 600 women, it appears that nulliparity is protective against breast cancer:

		Case-Control Status: Breast Cancer	Control	
Exposure Status	Null	120 (40%=120/300)	180 (60%=180/300)	300
	Parous	180	120	300
		300	300	600

Odds Ratio = 0.44

However, when we take into account exposure to radiation, a different story emerges.

No radiation

	Cancer	Control	
Null	30	170	200
Parous	10	90	100
	40	260	300

Odds Ratio = 1.6

Radiation

	Cancer	Control	
Null	90	10	100
Parous	170	30	200
	260	40	300

Odds Ratio = 1.6

The unadjusted odds ratio of 0.44 is reversed. It now appears that nulliparity is a risk factor for breast cancer; this is reflected in the odds ratio that is greater than 1.

How did this apparent contradiction occur?

- In the nulliparous group, there are disproportionately fewer women exposed to radiation.
- Women exposed to radiation are more likely to have breast cancer.
- Women exposed to radiation were less likely to be nulliparous with the result that
- OR = 0.44 is biased due to the confounding effect of exposure to radiation.

The calculation of an association (for example an RR or an OR) for a 2x2 table of counts may be misleading because of one or more extraneous influences. An extraneous influence can be

- Confounder
- Effect modifier
- Both
- Neither

A **confounded association** is biased and does not tell us about the association of interest. An **effect modified** relationship changes with variations in the extraneous variable.

Intuitively, confounding is the

- Distortion of a predictor-outcome relationship due to a third variable that is related to both
  - predictor, and
  - outcome
- The bias from confounding can be a spurious
  - strengthening
  - weakening
  - elimination
  - reversal
- A reversal is said to be an example of **Simpson's Paradox**

Apparent, but not true, confounding can occur in the absence of a relationship between exposure and disease.

Example. Are breath mints associated with cancer?

Case-control Status

Exposure Status		Cancer	Control	
Breath Mints		200 (77%=200/260)	1646 (18%=1646/8935)	1846
None		67	7289	7356
		260	8935	9202

Odds Ratio = 13.22

It looks like we should not be eating breath mints. What happens if we control for smoking?

Smokers

	Cancer	Control	
Breath Mints	194	706	900
None	21	79	100
	215	785	1000

Odds Ratio = 1.03

Non-Smokers

	Cancer	Control	
Breath Mints	6	940	946
None	46	7210	7256
	52	8150	8202

Odds Ratio = 1.00

Controlling for smoking, eating breath mints is no longer associated with cancer.

If the extraneous variable has no effect on disease, then it will not cause confounding.

Example. Hot tea is suspected of being associated with esophageal cancer.

Case-Control Status

Exposure Drink	Cancer	Control	
Tea	1420 (94%=1420/1504)	3650 (81%=3650/4499)	5070
Water	84	849	933
	1504	4499	6003

Odds Ratio = 3.93

Notice that the tea drinkers have disproportionately fewer smokers.

Exposure Drink	Smoker	NON-Smoker	
Tea	70 (1.4%=70/5070)	5000	5070
Water	833 (89%=833/933)	100	933
	903	5100	6003

Interestingly, smoking status does not distort the association of tea with cancer.

SMOKERS

	Cancer	Control	
Tea	20	50	70
Water	75	758	833
	95	808	903

Odds Ratio = 4.04

NON-SMOKERS

	Cancer	Control	
Tea	1400	3600	5000
Water	9	91	100
	1409	3691	5100

Odds Ratio = 3.93

This is because smoking itself is ***not*** associated with esophageal cancer:

WATER

	Cancer	Control	
SMOKER	75	758	833
NOT	9	91	100
	84	849	933

Odds Ratio = 1.00

TEA

	Cancer	Control	
SMOKER	20	50	70
NOT	1400	3600	5000
	1420	3650	5070

Odds Ratio = 1.03

Thus,

- It is possible to observe a strong relationship between the extraneous variable (smoking) and exposure (tea).
- with no confounding of the exposure-disease relationship of interest.
- This will occur when the extraneous variable is unrelated to the disease outcome.

If the extraneous variable has no relationship to exposure, then it will not cause confounding.

Example: A crude analysis suggests that use of sugar substitutes is associated with bladder cancer.

Case Control Status

Exposure Status	Cancer	Healthy	
Substitute	106=75%	738=13%	844
Sugar	35	5149	5184
	141	5887	6028

Odds Ratio = 21.13

However, we have learned that smoking is associated with bladder cancer.

	Cancer	Healthy	
Smoker	127=90%	3051=52%	3178
NON-Smoker	14	2836	2850
	141	5887	6028

Odds Ratio = 8.4

However, smoking is not related to the use of sugar substitutes.

	Substitute	Sugar	
Smoker	445=14%	2733	3178
NON-Smoker	399=14%	2451	2850
	844	5184	6028

Odds Ratio = 1.0

The independence of smoking and sugar substitute use means that the stratum specific odds ratios will be close to the unadjusted odds ratio.

Stratum: Smokers

	Cancer	Control	
Substitute	95	350	445
Sugar	32	2701	2733
	127	3051	3178

Odds Ratio = 22.91

Stratum: NON-Smokers

	Cancer	Control	
Substitute	11	388	399
Sugar	3	2488	2491
	14	2876	2890

Odds Ratio = 23.51

Thus, an extraneous variable unrelated to exposure does not cause confounding.

We have what we need to define confounding.

Definition Confounding

A variable is confounding if

1. It is extraneous, not intermediary; and

2. It is related to disease, BOTH
   - among the exposed AND
   - among the unexposed; and

3. It is related to exposure.

Recall that an intermediary variable is an intermediate in a causal pathway.

- Example: Coal dust → Asthma → Lesions on Lung
  Asthma is the intermediary variable.

- Stratification on an intermediary variable eliminates the exposure disease relationship

When we discuss the logistic regression model, we'll learn about effect modification