# Contingency Tables

## Overview

1) **Types of Variables**
2) **Comparing (2) Categorical Variables**
   - Contingency (two-way) tables
   - $\chi^2$ Tests
3) **2 x 2 Tables**
   - Sampling designs
   - Testing for association
   - Estimation of effects
   - Paired binary data
4) **Stratified Tables**
   - Confounding
   - Effect Modification

## Factors and Contingency Tables

**Definition**: A **factor** is a categorical (discrete) variable taking a small number of values that represent the levels of the factor.

Examples

Gender with two levels: 1 = Male and 2 = Female

Disease status with three levels: 1 = Progression, 2 = Stable, 3 = Improved

AgeFactor with 4 levels: 1 = 20-29 yrs, 2 = 30-39, 3 = 40-49, 4 = 50-59

## Factors and Contingency Tables

**Data description**: Form one-way, two-way or multi-way tables of frequencies of factor levels and their combinations

- To assess whether two factors are related, we often construct an R x C table that cross-classifies the observations according to the 2 factors.
- Examining two-way tables of Factor A vs Factor B at each level of a third Factor C shows how the A/B association may be explained or modified by C (later).

**Data Summary:** Categorical data are often summarized by reporting the proportion or percent in each category. Alternatively, one sometimes sees a summary of the relative proportion (odds) in each category (relative to a "baseline" category).

**Testing:** We can test whether the factors are related using a $\chi^2$ test.

# Categorical Data

**Example**: From Doll and Hill (1952) - retrospective assessment of smoking frequency. The table displays the daily average number of cigarettes for lung cancer patients and control patients. Note there are equal numbers of cancer patients and controls.

| | Daily # cigarettes | | | | | | |
|---|---|---|---|---|---|---|---|
| | None | < 5 | 5-14 | 15-24 | 25-49 | 50+ | Total |
| Cancer | 7<br>0.5% | 55<br>4.1% | 489<br>36.0% | 475<br>35.0% | 293<br>21.6% | 38<br>2.8% | 1357 |
| Control | 61<br>4.5% | 129<br>9.5% | 570<br>42.0% | 431<br>31.8% | 154<br>11.3% | 12<br>0.9% | 1357 |
| Total | 68 | 184 | 1059 | 906 | 447 | 50 | 2714 |

# $\chi^2$ Test

We want to test whether the smoking frequency is the same for each of the populations sampled. We want to test whether the **groups** are **homogeneous** with respect to a characteristic.

$H_0$: smoking probability same in both groups

$H_A$: smoking probability not the same

**Q:** What does $H_0$ predict we would observe if all we knew were the marginal totals?

| | Daily # cigarettes | | | | | | |
|---|---|---|---|---|---|---|---|
| | None | < 5 | 5-14 | 15-24 | 25-49 | 50+ | Total |
| Cancer | | | | | | | 1357 |
| Control | | | | | | | 1357 |
| Total | 68 | 184 | 1059 | 906 | 447 | 50 | 2714 |

# $\chi^2$ Test

**A:** $H_0$ predicts the following **expectations**:

| | Daily # cigarettes | | | | | | |
|---|---|---|---|---|---|---|---|
| | None | < 5 | 5-14 | 15-24 | 25-49 | 50+ | Total |
| Cancer | 34 | 92 | 529.5 | 453 | 223.5 | 25 | 1357 |
| Control | 34 | 92 | 529.5 | 453 | 223.5 | 25 | 1357 |
| Total | 68 | 184 | 1059 | 906 | 447 | 50 | 2714 |

Each group has the same proportion in each cell as the overall **marginal proportion.** The "equal" expected number for each group is the result of the equal sample size in each group (what would change if there were half as many cases as controls?)

## $\chi^2$ Test

Summing the differences between the observed and expected counts provides an overall assessment of $H_0$.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1)\times(c-1))$$

$X^2$ is known as the **Pearson's Chi-square Statistic.**

- Large values of $X^2$ suggests the data are not consistent with $H_0$
- Small values of $X^2$ suggests the data are consistent with $H_0$

# $\chi^2$ Test

In example 3 the contributions to the $X^2$ statistic are:

| | Daily # cigarettes | | | | | | |
|---|---|---|---|---|---|---|---|
| | None | < 5 | 5-14 | 15-24 | 25-49 | 50+ | Total |
| Cancer | $\frac{(7-34)^2}{34}$ | $\frac{(55-92)^2}{92}$ | etc. | | | | |
| Control | $\frac{(61-34)^2}{34}$ | | | | | | |
| Total | | | | | | | |

| | Daily # cigarettes | | | | | | |
|---|---|---|---|---|---|---|---|
| | None | < 5 | 5-14 | 15-24 | 25-49 | 50+ | Total |
| Cancer | 21.44 | 14.88 | 3.10 | 1.07 | 21.61 | 6.76 | |
| Control | 21.44 | 14.88 | 3.10 | 1.07 | 21.61 | 6.76 | |
| Total | | | | | | | |

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 137.7$$

$p = P(X^2 > \chi^2(5) \mid H_0 \text{ true}) < 0.0001$

Conclusion?

# $\chi^2$ Test

| | Factor Levels | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | … | C | Total |
| 1 | $O_{11}$ | $O_{12}$ | … | $O_{1C}$ | $N_1$ |
| Group 2 | $O_{21}$ | | | | $N_2$ |
| 3 | $O_{31}$ | | | | $N_3$ |
| ⋮ | ⋮ | | | | |
| R | $O_{R1}$ | | | $O_{RC}$ | $N_R$ |
| Total | $M_1$ | $M_2$ | | $M_C$ | T |

1. Compute the expected cell counts under homogeneity assumption:

$$E_{ij} = N_i M_j / T$$

2. Compute the chi-square statistic:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3. Compare $X^2$ to $\chi^2(df)$ where

$$df = (R\text{-}1) \times (C\text{-}1)$$

4. Interpret acceptance/rejection or p-value.

# 2 x 2 Tables

**Example 1**: Pauling (1971)

Patients are randomized to either receive Vitamin C or placebo. Patients are followed-up to ascertain the development of a cold.

| | Cold - Y | Cold - N | Total |
|---|---|---|---|
| Vitamin C | 17 | 122 | 139 |
| Placebo | 31 | 109 | 140 |
| Total | 48 | 231 | 279 |

**Q:** Is treatment with Vitamin C associated with a reduced probability of getting a cold?

**Q:** If Vitamin C is associated with reducing colds, then what is the magnitude of the effect?

# 2 x 2 Tables

**Example 2**: Keller (AJPH, 1965)

Patients with (cases) and without (controls) oral cancer were surveyed regarding their smoking frequency (this table collapses over the smoking frequency categories).

| | Case | Control | Total |
|---|---|---|---|
| Smoker | 484 | 385 | 869 |
| Non-Smoker | 27 | 90 | 117 |
| Total | 511 | 475 | 986 |

**Q:** Is oral cancer associated with smoking?

**Q:** If smoking is associated with oral cancer, then what is the magnitude of the risk?

# 2 x 2 Tables

**Example 3**: Sex-linked traits

Suppose we collect a random sample of Drosophila and cross classify eye color and sex.

| | male | female | Total |
|---|---|---|---|
| red | 165 | 300 | 465 |
| white | 176 | 81 | 257 |
| Total | 341 | 381 | 722 |

**Q:** Is eye color associated with sex?

**Q:** If eye color is associated with sex, then what is the magnitude of the effect?

## 2 x 2 Tables

**Example 4**: Matched case control study

213 subjects with a history of acute myocardial infarction (AMI) were ***matched*** by age and sex with one of their siblings who did not have a history of AMI. The prevalence of a particular polymorphism was compared between the siblings

| | AMI | | |
|---|---|---|---|
| | carrier | noncarrier | Total |
| No AMI carrier | 73 | 14 | 87 |
| No AMI noncarrier | 23 | 103 | 126 |
| Total | 96 | 117 | 213 |

**Q:** Is there an association between the polymorphism and AMI?

**Q:** If there is an association then what is the magnitude of the effect?

# 2 x 2 Tables

Each of these tables (except for example 4) can be represented as follows:

Disease Status

Exposure Status

| | D | not D | Total |
|---|---|---|---|
| E | a | b | $(a + b) = n_1$ |
| not E | c | d | $(c + d) = n_2$ |
| Total | $(a + c) = m_1$ | $(b + d) = m_2$ | N |

The question of association can be addressed with **Pearson's** $X^2$ (except for example 4) We compute the **expected** cell counts as follows:

**Expected:**

| | D | not D | Total |
|---|---|---|---|
| E | $n_1m_1/N$ | $n_1m_2/N$ | $(a + b) = n_1$ |
| not E | $n_2m_1/N$ | $n_2m_2/N$ | $(c + d) = n_2$ |
| Total | $(a + c) = m_1$ | $(b + d) = m_2$ | N |

# 2 x 2 Tables

Pearson's chi-square is given by:

$$X^2 = \sum_{i=1}^{4} (O_i - E_i)^2 / E_i$$

$$= \left( a - \frac{n_1 m_1}{N} \right)^2 / \left( \frac{n_1 m_1}{N} \right) + \left( b - \frac{n_1 m_2}{N} \right)^2 / \left( \frac{n_1 m_2}{N} \right) +$$

$$\left( c - \frac{n_2 m_1}{N} \right)^2 / \left( \frac{n_2 m_1}{N} \right) + \left( d - \frac{n_2 m_2}{N} \right)^2 / \left( \frac{n_2 m_2}{N} \right) +$$

$$= \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2}$$

## 2 x 2 Tables

**Example 1**: Pauling (1971)

| | Cold - Y | Cold - N | Total |
|---|---|---|---|
| Vitamin C | 17 (12%) | 122 (88%) | 139 |
| Placebo | 31 (22%) | 109 (78%) | 140 |
| Total | 48 | 231 | 279 |

$H_0$ : probability of disease does not depend on treatment

$H_A$ : probability of disease does depend on treatment

$$X^2 = \frac{N(ad-bc)^2}{n_1 n_2 m_1 m_2}$$

$$= \frac{279(17\times 109 - 31\times 122)^2}{139\times 140\times 48\times 231}$$

$$= 4.81$$

For the p-value we compute $P(\chi^2(1) > 4.81) = 0.028$. Therefore, we reject the homogeneity of disease probability in the two treatment groups.

# 2 x 2 Tables
# Applications In Epidemiology

**Example 1** fixed the number of E and not E, then evaluated the disease status after a fixed period of time (same for everyone). This is a **prospective study**. Given this design we can estimate the **relative risk**:

$$RR = \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

The range of RR is [0, ∞). By taking the logarithm, we have (- ∞, +∞) as the range for ln(RR) and a better approximation to normality for the estimated $\ln(\hat{RR})$:

$$\ln\left(\hat{RR}\right) = \ln\left(\frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})}\right) = \ln\left(\frac{p_1}{p_2}\right)$$

$$= \ln\left(\frac{a/n_1}{c/n_2}\right)$$

$$\ln(\hat{RR}) \sim approx\ N\left(\ln(p_1/p_2), \frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}\right)$$

## Relative Risk

| | Cold - Y | Cold - N | Total |
|---|---|---|---|
| Vitamin C | 17 | 122 | 139 |
| Placebo | 31 | 109 | 140 |
| Total | 48 | 231 | 279 |

The estimated relative risk is:

$$\hat{RR} = \frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \bar{E})}$$

$$= \frac{17/139}{31/140} = 0.55$$

We can obtain a 95% confidence interval for the relative risk by first obtaining a confidence interval for the log-RR:

$$\ln\left(\hat{RR}\right) \pm 1.96 \times \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$$

and exponentiating the endpoints of the CI.

Note that disease status and exposure status are transposed here compared to previous tables.

```
. csi 17 31 122 109

                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        17          31  |        48
        Noncases |       122         109  |       231
-----------------+------------------------+----------
           Total |       139         140  |       279
                 |                        |
            Risk |  .1223022    .2214286  |   .172043
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |        -.0991264       | -.1868592   -.0113937
      Risk ratio |         .5523323       |  .3209178    .9506203
 Prev. frac. ex. |         .4476677       |  .0493797    .6790822
 Prev. frac. pop |         .2230316       |
                 +-----------------------------------------------
                            chi2(1) =     4.81  Pr>chi2 = 0.0283
```

# 2 x 2 Tables

**Example 2**: Keller (AJPH, 1965)

Patients with (cases) and without (controls) oral cancer were surveyed regarding their smoking frequency (this table collapses over the smoking frequency categories).

| | Case | Control | Total |
|---|---|---|---|
| Smoker | 484 | 385 | 869 |
| Non-Smoker | 27 | 90 | 117 |
| Total | 511 | 475 | 986 |

**Q:** Is oral cancer associated with smoking?

**Q:** If smoking is associated with oral cancer, then what is the magnitude of the risk?

## 2 x 2 Tables
## Applications In Epidemiology

In **Example 2** we fixed the number of **cases** and **controls** then ascertained exposure status. Such a design is known as **case- control study**. Based on this we are able to directly estimate:

$$P(E \mid D) \quad \text{and} \quad P(E \mid \overline{D})$$

However, we generally are interested in the relative risk of disease given exposure, which is not estimable from these data alone - we've fixed the number of diseased and diseased free subjects, and it can be shown that in general:

$$\mathrm{P(D \mid E)} \neq \mathrm{P(E \mid D)}$$

$$\frac{\mathrm{P}\left(\mathrm{D}|\mathrm{E}\right)}{\mathrm{P}\left(\mathrm{D}|\overline{\mathrm{E}}\right)} \neq \frac{\mathrm{P}\left(\mathrm{E}|\mathrm{D}\right)}{\mathrm{P}\left(\mathrm{E}|\overline{\mathrm{D}}\right)}$$

## Odds Ratio

Instead of the relative risk we can estimate the **exposure odds ratio** which (surprisingly) is equivalent to the **disease odds ratio**:

$$\frac{P(E \mid D)/(1-P(E \mid D))}{P(E \mid \overline{D})/(1-P(E \mid \overline{D}))} = \frac{P(D \mid E)/(1-P(D \mid E))}{P(D \mid \overline{E})/(1-P(D \mid \overline{E}))}$$

In other words, **the odds ratio can be estimated regardless of the sampling scheme.**

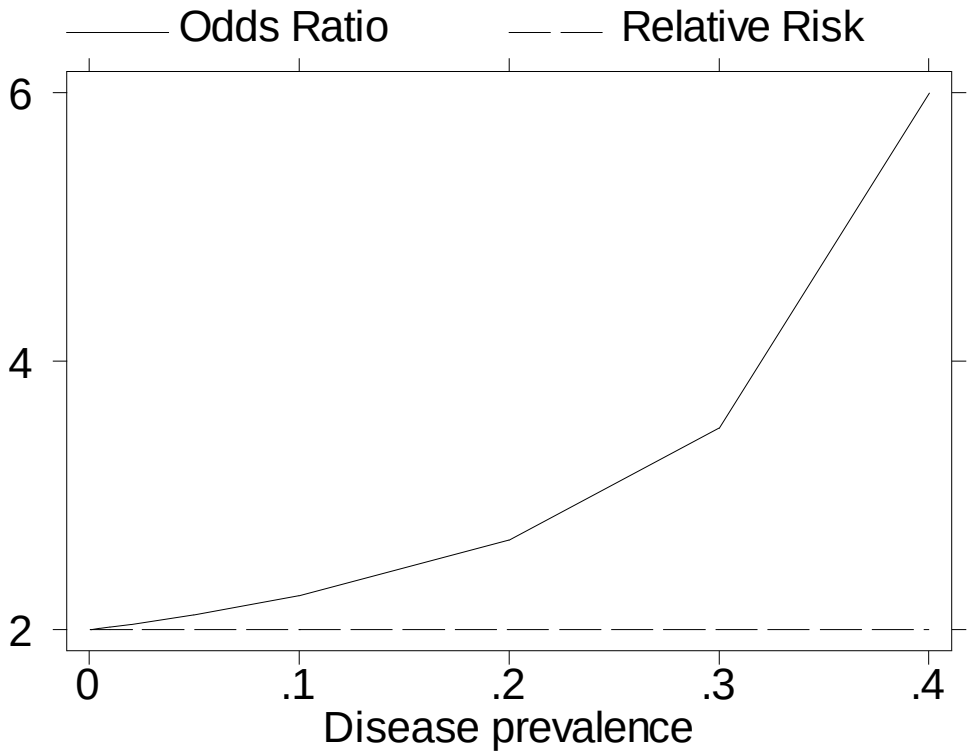
Furthermore, for rare diseases, P(D | E) ≈ 0 so that the disease odds ratio approximates the relative risk:

$$\frac{P(D \mid E)/(1-P(D \mid E))}{P(D \mid \overline{E})/(1-P(D \mid \overline{E}))} \approx \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

Since with case-control data we are able to effectively estimate the exposure odds ratio we are then able to equivalently estimate the disease odds ratio which for rare diseases approximates the relative risk.

**For rare diseases (e.g., prevalence <5%), the (sample) odds ratio estimates the (population) relative risk.**

# Odds Ratio



## Odds Ratio

Like the relative risk, the odds ratio has [0, ∞) as its range. The **log odds ratio** has (- ∞, +∞) as its range and the normal approximation is better as an approximation to the dist of the estimated log odds ratio.

$$OR = \frac{p_1 / 1 - p_1}{p_2 / 1 - p_2}$$

$$\hat{OR} = \frac{\hat{p}_1 / 1 - \hat{p}_1}{\hat{p}_2 / 1 - \hat{p}_2}$$

$$\hat{OR} = \frac{ad}{bc}$$

Confidence intervals are based upon:

$$\ln\left(\hat{OR}\right) \sim \mathrm{N}\left(\ln(\mathrm{OR}), \frac{1}{\mathrm{n}_1 p_1} + \frac{1}{\mathrm{n}_1 (1 - p_1)} + \frac{1}{\mathrm{n}_2 p_2} + \frac{1}{\mathrm{n}_2 (1 - p_2)}\right)$$

Therefore, a 95% confidence interval for the log odds ratio is given by:

$$\ln\left(\frac{ad}{bc}\right) \pm 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

# Odds Ratio

```
. cci 484 27 385 90

                                                        Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       484          27  |       511      0.9472
        Controls |       385          90  |       475      0.8105
-----------------+------------------------+----------------------
           Total |       869         117  |       986      0.8813
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         4.190476       |  2.633584    6.836229  (exact)
 Attr. frac. ex. |         .7613636       |  .6202893    .8537205  (exact)
 Attr. frac. pop |          .721135       |
                 +-----------------------------------------------
                            chi2(1) =    43.95  Pr>chi2 = 0.0000
```

## Interpreting Odds ratios

1. What is the outcome of interest? (i.e. disease)
2. What are the two groups being contrasted? (i.e. exposed and unexposed)

$$\text{OR} = \frac{\text{odds of OUTCOME in EXPOSED}}{\text{odds of OUTCOME in UNEXPOSED}}$$

- Similar to RR for rare diseases
- Meaningful for both cohort and case-control studies
- OR > 1 $\Rightarrow$ increased risk of OUTCOME with EXPOSURE
- OR < 1 $\Rightarrow$ decreased risk of OUTCOME with EXPOSURE

# 2 x 2 Tables

**Example 3**: Sex-linked traits

Suppose we collect a random sample of Drosophila and cross classify eye color and sex.

| | male | female | Total |
|---|---|---|---|
| red | 165 | 300 | 465 |
| white | 176 | 81 | 257 |
| Total | 341 | 381 | 722 |

**Q:** Is eye color associated with sex?

**Q:** If eye color is associated with sex, then what is the magnitude of the effect?

# 2 x 2 Tables
# Applications in Epidemiology

**Example 3** is an example of a **cross-sectional** study since only the total for the entire table is fixed in advance. The row totals or column totals are not fixed in advance.

| | male | female | Total |
|---|---|---|---|
| red | 165 (48%) | 300 (79%) | 465 |
| white | 176 | 81 | 257 |
| Total | 341 | 381 | 722 |

Cross-sectional studies

- Sample from the entire population, not by disease status or exposure status
- Use chi-square test to test for association
- Use RR or OR to summarize association
- Cases of disease are **prevalent** cases (compared to incident cases in a prospective or cohort study)

# 2 x 2 Tables
# Applications in Epidemiology

Case = red eye color
Noncase = white eye color

```
                     male        female
-----------------+------------------------+------------
           Cases |       165          300  |        465
        Noncases |       176           81  |        257
-----------------+------------------------+------------
           Total |       341          381  |        722
            Risk |   .483871    .7874016   |   .6440443
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
 Risk difference |        -.3035306        |   -.3706217   -.2364395
      Risk ratio |         .6145161        |     .544263    .6938375
 Prev. frac. ex. |         .3854839        |    .3061625     .455737
 Prev. frac. pop |         .1820637        |
      Odds ratio |          .253125        |    .1830613    .3500144
                 +-----------------------------------------------------
                              chi2(1) =    72.32  Pr>chi2 = 0.0000
```

## 2 x 2 Tables

**Example 4**: Matched case control study

213 subjects with a history of acute myocardial infarction (AMI) were ***matched*** by age and sex with one of their siblings who did not have a history of AMI. The prevalence of a particular polymorphism was compared between the siblings

| | AMI | | |
|---|---|---|---|
| | carrier | noncarrier | Total |
| No AMI carrier | 73 | 14 | 87 |
| No AMI noncarrier | 23 | 103 | 126 |
| Total | 96 | 117 | 213 |

**Q:** Is there an association between the polymorphism and AMI?

**Q:** If there is an association then what is the magnitude of the effect?

## Paired Binary Data

**Example 4** measures a binary response in sibs. This is an example of **paired binary data**. One way to display these data is the following:

| | Carrier | Noncarrier | Total |
|---|---|---|---|
| AMI | 96 | 117 | 213 |
| No AMI | 87 | 126 | 213 |
| Total | 183 | 243 | 426 |

**Q:** Can't we simply use $X^2$ Test of Homogeneity to assess whether this is evidence for an increase in knowledge?

**A:** NO!!! The $X^2$ tests assume that the rows are **independent** samples. In this design the 213 with AMI are genetically related to the 213 w/o AMI.

## Paired Binary Data

For paired binary data we display the results as follows:

| | AMI | |
|---|---|---|
| | 1 | 0 |
| No AMI 1 | $n_{11}$ | $n_{10}$ |
| 0 | $n_{01}$ | $n_{00}$ |

This analysis explicitly recognizes the heterogeneity of subjects. Thus, those that score (0,0) and (1,1) provide no information about the association between AMI and the polymorphism. These are known as the **concordant pairs**. The information regarding the association is in the **discordant pairs**, (0,1) and (1,0).

$$p_1 = P(\text{carrier} \mid \text{AMI})$$

$$p_0 = P(\text{carrier} \mid \text{No AMI})$$

$$H_0 : p_1 = p_0$$

$$H_A : p_1 \neq p_0$$

$$\hat{p}_1 - \hat{p}_0 = \frac{n_{11} + n_{01}}{N} - \frac{n_{11} + n_{10}}{N} = \frac{n_{01} - n_{10}}{N}$$

## Paired Binary Data
## McNemar's Test

Under the null hypothesis, $H_0: p_1 = p_0$, we expect equal numbers of 01's and 10's. ($E[n_{01}] = E[n_{10}]$). Specifically, under the null:

$$M = n_{01} + n_{10}$$

$$n_{10} \mid M \sim Bin\left(M, \frac{1}{2}\right)$$

$$Z = \frac{n_{10} - M\frac{1}{2}}{\sqrt{M\frac{1}{2}\left(1-\frac{1}{2}\right)}}$$

Under $H_0$, $Z^2 \sim \chi^2(1)$, and forms the basis for **McNemar's Test for Paired Binary Responses**.

The odds ratio comparing the odds of carrier in those with AMI to odds of carrier in those w/o AMI is estimated by:

$$\hat{OR} = \frac{n_{01}}{n_{10}}$$

Confidence intervals can be obtained as described in Breslow and Day (1981), section 5.2, or in Armitage and Berry (1987), chapter 16.

Example 4:

| | AMI | | |
|---|---|---|---|
| | carrier | noncarrier | Total |
| carrier | 73 | 14 | 87 |
| No AMI noncarrier | 23 | 103 | 126 |
| Total | 96 | 117 | 213 |

We can test $H_0$: $p_1 = p_2$ using **McNemar's Test:**

$$Z = \frac{n_{01} - M\frac{1}{2}}{\sqrt{M\,\frac{1}{2}\left(\frac{1}{2}\right)}}$$

$$= \frac{23 - \left(23+14\right)/2}{\sqrt{\left(23+14\right)/4}}$$

$$= 1.48$$

Comparing $1.48^2$ to a $\chi^2$ (1) we find that $p > 0.05$. Therefore, we do not reject the null hypothesis and find little evidence of association between gene and disease.

We estimate the odds ratio as $\hat{OR} = 23/14 = 1.64.$

# Matched case-control data

```
. mcci 73 23 14 103

                 | Controls                |
Cases            |   Exposed   Unexposed   |      Total
-----------------+-------------------------+-----------
         Exposed |        73          23   |         96
       Unexposed |        14         103   |        117
-----------------+-------------------------+-----------
           Total |        87         126   |        213

McNemar's chi2(1) =      2.19    Prob > chi2 = 0.1390
Exact McNemar significance probability       = 0.1877

Proportion with factor
        Cases       .4507042
        Controls    .4084507     [95% Conf. Interval]
                    ---------    --------------------
        difference  .0422535     -.0181247   .1026318
        ratio       1.103448      .9684942   1.257207
        rel. diff.  .0714286     -.0197486   .1626057

        odds ratio  1.642857      .8101776   3.452833   (exact)
```

## Two way tables - Review

- How were data collected?
  - Cohort design
  - Case-control design
  - Cross-sectional design
  - Matched pairs
- Is there an association?
  - R x C Tables
    - Chi-square tests of Homogeneity & Independence
  - 2 x 2 Tables
    - Chi-square test
    - Paired data and McNemar's
- What is the magnitude of the association?
  - Relative risk
  - Odds ratio (≈ relative risk for rare diseases)
  - Risk difference (attributable risk)

## SUMMARY
## Measures of Association for 2 x 2 Tables

**RD** = $p_1 - p_2$ = risk difference (null: RD = 0)

- also known as **attributable risk** or **excess risk**
- measures **absolute effect** – the proportion of cases among the exposed that can be attributed to exposure

**RR** = $p_1 / p_2$ = relative risk (null: RR = 1)

- measures **relative effect** of exposure
- bounded above by $1/p_2$

**OR** = $[p_1(1-p_2)]/[p_2(1-p_1)]$ = odds ratio (null: OR = 1)

- range is 0 to $\infty$
- approximates RR for rare events
- invariant of switching rows and cols
- good behavior of p-values and CI even for small to moderate sample size

## SUMMARY
## Models for 2 x 2 Tables

1. **Cohort** ("Prospective", "Followup")
   - Sample $n_1$ "exposed" and $n_2$ "unexposed"
   - Follow everyone for equal period of time
   - Observe incident disease – $r_1$ cases among exposed, $r_2$ cases among unexposed
   - Model: Two independent binomials

     $r_1 \sim \text{binom}(p_1, n_1)$

     $r_2 \sim \text{binom}(p_2, n_2)$

     $p_1 = P(D|E)$

     $p_2 = P(D|\overline{E})$
   - Useful measures of association – RR,OR,RD
   - Examples:

     $r_i$ = number of cases of HIV during 1 year followup of $n_i$ individuals in arm i of HIV prevention trial

     $r_i$ = number of low birthweight babies among $n_i$ live births

## SUMMARY
## Models for 2 x 2 Tables

**2. Case-Control**

- Sample $n_1$ "cases" and $n_2$ "controls"
- Observe exposure history – $r_1$ exposed among cases, $r_2$ exposed among controls
- Model: Two independent binomials

$$r_1 \sim \text{binom}(q_1, n_1)$$
$$r_2 \sim \text{binom}(q_2, n_2)$$
$$q_1 = P(E|D)$$
$$q_2 = P(E|\overline{D})$$

- Useful measures of association – OR
- Examples:
  - $r_i$ = consistent condom use (yes/no) among those with/without HPV infection
  - $r_i$ = number exposed to alcohol during pregnancy among $n_i$ low birthweight/normal birthweight babies

## SUMMARY
## Models for 2 x 2 Tables

**3. Cross-sectional**

- Sample n individuals from population
- Observe both "exposure" and (prevalent) "disease" status.
- No longitudinal followup
- Useful measures of association – RR,OR,RD
- Example:

  $n_{ij}$ = number of gay men with gonorrhea in random sample of STD clinic attendees