UCLA STAT 13

Introduction to Statistical Methods for the Life and Health Sciences

Instructor: **Ivo Dinov**,

Asst. Prof. of Statistics and Neurology

Teaching Assistants:

Fred Phoa, Kirsten Johnson, Ming Zheng & Matilda Hsieh

University of California, Los Angeles, Fall 2005

http://www.stat.ucla.edu/~dinov/courses_students.html

Chapter 10

Chi-Square Test

Relative Risk/Odds Ratios

Further Considerations in Paired Experiments

- Many studies compare measurements before and after treatment
  - There can be difficulty because the effect of treatment could be confounded with other changes over time or outside variability
    - for example suppose we want to study a cholesterol lowering medication. Some patients may have a response because they are under study, not because of the medication.
  - We can protect against this by using randomized concurrent controls

Further Considerations in Paired Experiments

Example: A researcher conducts a study to examine the effect of a new **anti-smoking pill on smoking behavior**. Suppose he has collected data on 25 randomly selected smokers, 12 will receive treatment (a treatment pill once a day for three months) and 13 will receive a placebo (a mock pill once a day for three months). The researcher measures the number of cigarettes smoked per week before and after treatment, regardless of treatment group. Assume normality. The summary statistics are:

# cigs / week	n	$\bar{y}_{before}$	$\bar{y}_{after}$	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	163.92	152.50	11.42	1.10
Placebo	13	163.08	160.23	2.85	1.29

Further Considerations in Paired Experiments

Test to see if there is a difference in number of cigs smoked per week ***before and after the new treatment***, using $\alpha = 0.05$

H_o: $\mu_d = 0$

H_a: $\mu_d \neq 0$ ⇨

# cigs / week	n	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	11.42	1.10
Placebo	13	2.85	1.29

$$t_s = \frac{11.42}{1.10} = 10.38$$

df = $n_d - 1 = 12 - 1 = 11$

p < (0.0005)2 = 0.001, reject Ho.

These data show that there is a statistically significant difference in the true mean number of cigs/week before and after treatment with the new drug

Further Considerations in Paired Experiments

- This result does not necessarily demonstrate the effectiveness of the new medication
  - Smoking less per week could be due to the fact that patients know they are being studied (ie. difference statistically significantly different from zero)
  - All we can say is that he new medication appears to have a significant effect on smoking behavior

Further Considerations in Paired Experiments

Test to see if there is a difference in number of cigs smoked per week ***before and after in the placebo group***, using α = 0.05

H_o: $\mu_d = 0$

H_a: μ_d != 0

$t_s = \frac{2.85}{1.29} = 2.21$

\# cigs / week

	n	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	11.42	1.10
Placebo	13	2.85	1.29

df = $n_d - 1 = 13 - 1 = 12$

(0.02)2 < p < (0.025)2

0.04 < p < 0.05 , reject H_o

These data show that there is a statistically significant difference in the true mean number of cigs/week before and after treatment with the a placebo

Further Considerations in Paired Experiments

- Patients who did not receive the new drug also experienced a statistically significant drop in the number of cigs smoked per week
  - This doesn't necessarily mean that the treatment was a failure because both groups had a significant decrease
  - We need to isolate the effect of therapy on the treatment group
  - Now the question becomes: was the drop in # of cigs/week significantly different between the medication and placebo groups?
  - How can we verify this?

Further Considerations in Paired Experiments

Test to see if there is the *difference in number of cigs smoked per week before and after treatment* was significant ***between the treatment and placebo groups***, using α = 0.05

$$SE_{\bar{y}_1-\bar{y}_2} = \sqrt{(1.10)^2 + (1.29)^2} = 1.695$$

Ho: $\mu_d = 0$

H_a: μ_d != 0

$$t_s = \frac{11.42 - 2.85}{1.695} = 5.06$$

\# cigs / week

	n	$\bar{d}$	$SE_{\bar{d}}$
Treatment	12	11.42	1.10
Placebo	13	2.85	1.29

df = 22

p < (0.0005)2 = 0.001, reject H_o

These hypothesis tests provide strong evidence that the new anti-smoking medication is effective. If the experimental design had not included the placebo group, the last comparison could not have been made and we could not support the efficacy of the drug.

Reporting of Paired Data

- Common in publications to report the mean and standard deviation of the two groups being compared
  - In a paired situation it is important to report the mean of the differences as well as the standard deviation of the differences
  - Why?

Limitations of $\bar{d}$

- There are two majors limitations of $\bar{d}$
  1. we are restricted to questions concerning $\bar{d}$
     - ❑ When some of the differences are positive and some are negative, the magnitude of does not reflect the "typical" magnitude of the differences.
     - ❑ Suppose we had the following differences: +40, -35, +20, -42, +61, -31.

```
Descriptive Statistics: data

Variable  N  N*  Mean  SE Mean  StDev  Minimum     Q1  Median    Q3   Max
data      6   0  2.17     17.9   43.9    -42.0  -36.8   -5.50  45.3  61.0
```

- ❑ What is the problem with this? Small average, but differences are large.
- ❑ What other statistic would help the reader recognize this issue?

Limitations of $\bar{d}$

2. limited to questions about aggregate differences
   - ❑ If treatment A is given to one group of subjects and treatment B is given to a second group of subjects, it is impossible to know how a person in group A would have responded to treatment B.

- Need to beware of these viewpoints and take time to look at the data, not just the summaries
- To verify accuracy we need to look at the individual measurements.
  - Accuracy implies that the d's are small

Inference for Proportions

- We have discussed two major methods of data analysis:
  - Confidence intervals: quantitative and categorical data
  - Hypothesis Testing: quantitative data
- In chapter 10, we will be discussing hypothesis tests for categorical variables
- RECALL: Categorical data
  - Gender (M or F)
  - Type of car (compact, mid-size, luxury, SUV, Truck)
- We typically summarize this type of data with proportions, or probabilities of the various categories

The χ^2 Goodness of Fit Test

- Let's start by considering analysis of a single sample of categorical data
- This is a hypothesis test, so we will be considering the four major parts
- #1 The general for of the hypotheses:
  - Ho: probabilities are equal to some specified values
  - Ha: probabilities are not equal to some specified values
- #2 The Chi-Square test statistic (p.393)

O – Observed frequency

E – Expected frequency (according to H_o)

For the goodness of fit test

df = # of categories – 1

$$\chi_s^2 = \sum \frac{(O-E)^2}{E}$$

The χ^2 Goodness of Fit Test

- Like other test statistics a smaller value for indicates that the data agree with H_o
- If there is disagreement from Ho, the test stat will be large because the difference between the observed and expected values is large
- #3 P-value:
  - Table 9, p.686
  - Uses df (similar idea to the t table)
    - After first n-1 categories have been specified, the last can be determined because the proportions must add to 1
  - One tailed distribution, not symmetric (different from t table)
- #4 Conclusion similar to other conclusions (TBD)

The χ^2 Goodness of Fit Test

Example: Mendel's pea experiment. Suppose a tall offspring is the event of interest and that the true proportion of tall peas (based on a 3:1 phenotypic ratio) is 3/4 or p = 0.75. He would like to show that his data follow this 3:1 phenotypic ratio.

The hypotheses (#1):

H_o:P(tall) = 0.75 (No effect, follows a 3:1phenotypic ratio)

P(dwarf) = 0.25

H_a: P(tall) $\neq$ 0.75

P(dwarf) $\neq$ 0.25

The χ^2 Goodness of Fit Test

Suppose the data were:

N = 1064

Tall = 787 These are the O's (observed values)

Dwarf = 277

To calculate the E's (expected values), we will take the hypothesized proportions under Ho and multiply them by the total sample size

Tall = (0.75)(1064) = 798

Dwarf = (0.25)(1064) = 266

Quick check to see if total = 1064

The χ^2 Goodness of Fit Test

Next calculate the test statistic (#2)

$$\chi_s^2 = \frac{(787-798)^2}{798} + \frac{(277-266)^2}{266} = 0.152 + 0.455 = 0.607$$

The p-value (#3):

df = 2 - 1 = 1

P > 0.20, fail to reject Ho

CONCLUSION: These data provide evidence that the true proportions of tall and dwarf offspring are not statistically significantly different from their hypothesized values of 0.75 and 0.25, respectively. In other words, these data are reasonably consistent with the Mendelian 3:1 phenotypic ratio.

The χ^2 Goodness of Fit Test

- Tips for calculating χ^2 (p.393):
  - Use the SOCR Resource (**www.socr.ucla.edu**)
  - The table of observed frequencies must include ALL categories, so that the sum of the Observed's is equal to the total number of observations
  - The O's must be absolute, rather than relative frequencies (ie. counts not percentages)
  - Can round each part to a minimum of 2 decimal places, if you aren't using your calculator's memory

Compound Hypotheses

- The hypotheses for the t-test contained one assertion: that the means were equal or not.
- The goodness of fit test can contain more than one assertion
  - this is called a compound hypothesis
  - The alternative hypothesis is non-directional, it measures deviations in all directions (*at least one* probability differs from its hypothesized value)

Directionality

- RECALL: dichotomous – having two categories
- If the categorical variable is dichotomous, Ho is not compound, so we can specify a directional alternative
  - when one category goes up the other must go down
  - RULE OF THUMB: when df = 1, the alternative can be specified as directional

Directionality

Example: A hotspot is defined as a 10 km^2 area that is species rich (heavily populated by the species of interest). Suppose in a study of butterfly hotspots in a particular region, the number of butterfly hotspots in a sample of 2,588, 10 km^2 areas is 165. In theory, 5% of the areas should be butterfly hotspots. Do the data provide evidence to suggest that the number of butterfly hotspots is increasing from the theoretical standards? Test using α = 0.01.

Directionality

H_o: p(hotspot) = 0.05

p(other spot) = 0.95

H_a: p(hotspot) > 0.05

p(other spot) < 0.95

	Hotspot	Other spot	Total
Observed	165	2423	2588
Expected	(0.05)(2588) = 129.4	(0.95)(2588) = 2458.6	2588

$$\chi^2_s = \frac{(165-129.4)^2}{129.4} + \frac{(2423-2458.6)^2}{2458.6}$$
$$= 9.79 + 0.52 = 10.31$$

Directionality

df = 2 - 1 = 1

0.001 < p < 0.01, however because of directional alternative the p-value needs to be divided by 2 (* see note at top of table 9)

Therefore, 0.0005 < p < 0.005; Reject H_o

CONCLUSION: These data provide evidence that in this region the number of butterfly hotspots is increasing from theoretical standards (ie. greater than 5%).

Goodness of Fit Test, in general

- The expected cell counts can be determined by:
  - Pre-specified proportions set-up in the experiment
    - For example: 5% hot spots, 95% other spots
  - Implied
    - For example: Of 250 births at a local hospital is there evidence that there is a gender difference in the proportion of males and females? Without further information this implies that we are looking for P(males) = 0.50 and P(females) = 0.50.

Goodness of Fit Test, in general

- Goodness of fit tests can be compound (i.e., Have more than 2 categories):
  - For example: Of 250 randomly selected CP college students is there evidence to show that there is a difference in area of home residence, defined as: Northern California (North of SLO); Southern California (In SLO or South of SLO); or Out of State? Without further information this implies that we are looking for P(N.Cal) = 0.33, P(S.Cal) = 0.33, and P(Out of State) = 0.33.

The χ^2 Test for the 2 X 2 Contingency Table

- We will now consider analysis of two samples of categorical data
- This type of analysis utilizes tables, called contingency tables
  - These tables focus on the dependency or association between column and row variables

The χ^2 Test for the 2 X 2 Contingency Table

Example: Suppose 200 randomly selected cancer patients were asked if their primary diagnosis was Brain cancer and if they owned a cell phone before their diagnosis. The results are presented in the table below:

		Brain cancer		
		Yes	No	Total
Cell Phone	Yes	18	80	98
	No	7	95	102
	Total	25	175	200

The χ^2 Test for the 2 X 2 Contingency Table

- Does it seem like there is an association between brain cancer and cell phone use?
  - How could we tell quickly?

Of the brain cancer patients 18/25 = 0.72, owned a cell phone before their diagnosis.

$\hat{P}$ (CP|BC) = 0.72, estimated probability of owning a cell phone given that the patient has brain cancer.

Of the other cancer patients, 80/175 = 0.46, owned a cell phone before their diagnosis.

$\hat{P}$ (CP|NBC) = 0.46, estimated probability of owning a cell phone given that the patient has another cancer.

		Brain cancer		
		Yes	No	Total
Cell Phone	Yes	18	80	98
	No	7	95	102
	Total	25	175	200

The χ^2 Test for the 2 X 2 Contingency Table

- The goal: We want to analyze the association, if any, between brain cancer and cell phone use
- This is a 2 X 2 table because there are two possible outcomes for each variable (each variable is dichotomous)
- Consider the following population parameters:

P(CP|BC) = true probability of owning a cell phone (CP) given that the patient had brain cancer (BC) is estimated by

$\hat{P}$ = (CP|BC) = 0.72

P(CP|NBC) = true probability of owning a cell phone given that the patient had another cancer, is estimated by

$\hat{P}$ = (CP|NBC) = 0.46

The χ^2 Test for the 2 X 2 Contingency Table

- The general form of a hypothesis test for a contingency table:
  - #1 The hypotheses:

    H_o: there is no association between variable 1 and variable 2 (independence)

    H_a: there is an association between variable 1 and variable 2 (dependence)

    NOTE: Using symbols can be tricky, be careful and read section 10.3

The χ^2 Test for the 2 X 2 Contingency Table

- #2 The test statistic:

  Expected cell counts can be calculated by

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

$$\chi_s^2 = \sum \frac{(O-E)^2}{E}$$

with df = (# rows – 1)(# col – 1)

  - #3 p-value and #4 conclusion are similar to the goodness of fit test.

The χ^2 Test for the 2 X 2 Contingency Table

Example: Brain cancer (cont')

Test to see if there is an association between brain cancer and cell phone use using α = 0.05

H_o: there is no association between brain cancer and cell phone (using notation P(CP|BC) = P(CP|NBC))

H_a: there is an association between brain cancer and cell phone (using notation P(CP|BC) $\neq$ P(CP|NBC))

(98)(25)/200

		Brain cancer		
		Yes	No	Total
Cell Phone	Yes	18 (12.25)	80 (85.75)	98
	No	7 (12.75)	95 (89.25)	102
	Total	25	175	200

The χ^2 Test for the 2 X 2 Contingency Table

$$\chi_s^2 = \frac{(18-12.25)^2}{12.25} + \frac{(7-12.75)^2}{12.75} + \frac{(80-85.75)^2}{85.75} + \frac{(95-89.25)^2}{89.25}$$
$$= 2.699 + 2.539 + 0.386 + 0.370 = 6.048$$

df = (2-1)(2-1) = 1

$0.01 < p < 0.02$, reject H_o.

CONCLUSION: These data show that there is a statistically significant association between brain cancer and cell phone use in patients that have been previously diagnosed with cancer.

The χ^2 Test for the 2 X 2 Contingency Table

- Output:

```
Chi-Square Test: C1, C2

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected
counts
            C1      C2  Total
    1       18      80     98
         12.25   85.75
         2.699   0.386

    2        7      95    102
         12.75   89.25
         2.593   0.370
Total       25     175    200
Chi-Sq = 6.048, DF = 1, P-Value = 0.014
```


The χ^2 Test for the 2 X 2 Contingency Table

- NOTE: df = 1, we could have carried this out as a one-tailed test
  - The probability that a patient with brain cancer owned a cell phone is greater than the probability that another cancer patient owned a cell phone
    - H_a: P(CP|BC) > P(CP|NBC)
  - Why didn't we carry this out as a one tailed test?
- CAUTION: *Association does not imply Causality!*

Computational Notes

1. Contingency table is useful for calculations, but not nice for presentation in reports.

2. When calculating observed values should be absolute frequencies, not relative frequencies. Also sum of observed values should equal grand total.

- To eyeball a contingency table for differences, check for proportionality of columns:
  - If the columns are nearly proportional then the data seem to agree with H_o
  - If the columns are not proportional then the data seem to disagree with H_o

Independence and Association in the 2x2 Contingency Table

- There are two main contexts for contingency tables:
  - Two independent samples with a dichotomous observed variable
  - One sample with two dichotomous observed variables

  NOTE: The χ^2 test procedure is the same for both situations

Example: Vitamin E. Subjects treated with either vitamin E or placebo for two years, then evaluated for a reduction in plaque from their baseline (Yes or No).

- Any study involving a dichotomous observed variable and completely randomized allocation to two treatments can be viewed this way

Example: Brain cancer and cell phone use. One sample, cancer patients, two observed variables: brain cancer (yes or no) and cell phone use (yes or no)

Independence and Association in the 2x2 Contingency Table

- When a dataset is viewed as a single sample with two observed variables, the relationship between the variables is thought of as independence or association.
  - Ho: independence (no association) between the variables
  - Ha: dependence (association) between the variables
- χ^2 is often called a test of independence or a test of association.

  NOTE: If columns and rows are interchanged test statistic will be the same

The r X k Contingency Table

- We now consider tables that are larger than a 2x2 (more than 2 groups or more than 2 categories), called rxk contingency tables
- Testing procedure is the same as the 2x2 contingency table, just more work and no possibility for a directional alternative
  - The goal of an rxk contingency table is to investigate the relationship between the row and column variables
- NOTE: Ho is a compound hypothesis because it contains more than one independent assertion
  - This will be true for all rxk tables larger than 2x2
  - In other words, the alternative hypothesis for rxk tables larger than 2x2, will always be nondirectional.

The r X k Contingency Table

Example: Many factors are considered when purchasing earthquake insurance. One factor of interest may be location with respect to a major earthquake fault. Suppose a survey was mailed to California residents in four counties (data shown below). Is there a statistically significant association between county of residence and purchase of earthquake insurance? Test using $\alpha = 0.05$.

		County				
		Contra Costa CC	Santa Clara SC	Los Angeles LA	San Bernardino SB	Total
Earthquake Insurance	Yes	117	222	133	109	581
	No	404	334	204	263	1205
	Total	521	556	337	372	1786

The r X k Contingency Table

H_o: There is no association between Earthquake insurance and county of residence in California.

$$\begin{cases} P(Y|CC) = P(Y|SC) = P(Y|LA) = P(Y|SB) \\ P(N|CC) = P(N|SC) = P(N|LA) = P(N|SB) \end{cases}$$

H_a: There is an association between Earthquake insurance and county of residence in California.

The probability of having earthquake insurance is not the same in each county.

The r X k Contingency Table

Chi-Square Test: C1, C2, C3, C4

http://socr.stat.ucla.edu/Applets.dir/ChiSquareTable.html

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts
           C1      C2      C3      C4  Total
    1     117     222     133     109    581
       169.49  180.87  109.63  121.01
       16.253   9.352   4.982   1.193

    2     404     334     204     263   1205
       351.51  375.13  227.37  250.99
        7.837   4.509   2.402   0.575

Total     521     556     337     372   1786
Chi-Sq = 47.105, DF = 3, P-Value = 0.000
```


The r X k Contingency Table

$p = 0.000 < 0.05$, reject H_o.

CONCLUSION: These data show that there is a statistically significant association between purchase of earthquake insurance and county of residence in California.

Applicability of Methods

- Conditions for validity of the χ^2 test:
  1. Design conditions
  - for a goodness of fit, it must be reasonable to regard the data as a random sample of categorical observations from a large population.
  - for a contingency table, it must be appropriate to view the data in one of the following ways:
    - as two or more independent random samples, observed with respect to a categorical variable
    - as one random sample, observed with respect to two categorical variables

* for either type of test, the observations within a sample must be independent of one another.

Applicability of Methods

- Conditions for validity of the χ^2 test (cont'):
  2. Sample conditions
  - critical values for table 9 only work if each expected value ≥ 5
  3. Form of H_o
  - for goodness of fit, H_o specifies values
  - for contingency table, H_o: row and column are not associated or use notation

Verification of Conditions

- Data consisting of several samples need to be independent sample.
  - If the design contains blocking or pairing the samples are not independent
- Try to reduce bias
- Only simple random sampling
  - No pairing for the version we are learning, although there is a paired Chi-Square test (section 10.8)
- No hierarchical structure
- **Check expected cell counts**

CI for the difference between probabilities

- Chi-Square tests for contingency tables tell us if there is an association or not between categories.
  - They tell us that there is a difference, but is it an important difference?
  - They do not give us any information as to the magnitude of any differences between probabilities
- For this we will calculate a confidence interval for the difference between probabilities

CI for the difference between probabilities

- A 95% confidence interval for $p_1 - p_2$

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{0.025}\left(SE_{\tilde{p}_1 - \tilde{p}_2}\right)$$

Where

Sample	
1	2
y_1	y_2
$n_1 - y_1$	$n_2 - y_2$
n_1	n_2

$$\tilde{p}_1 = \frac{y_1 + 1}{n_1 + 2} \qquad \tilde{p}_2 = \frac{y_2 + 1}{n_2 + 2}$$

$$SE_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 + 2}}$$

Slide 49 Stat 13, UCLA, Ivo Dinov

CI for the difference between probabilities

Example: Brain cancer continued

Cell Phone	Brain cancer: Yes	No	Total
Yes	18	80	98
No	7	95	102
Total	25	175	200

Calculate a 95% confidence interval for the difference in cell phone use between brain cancer and other cancer patients

$$\tilde{p}_1 = \frac{18+1}{25+2} = 0.704 \qquad \tilde{p}_2 = \frac{80+1}{175+2} = .458$$

Slide 50 Stat 13, UCLA, Ivo Dinov

CI for the difference between probabilities

95% CI continued…

$$SE_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{0.704(0.296)}{25+2} + \frac{0.457(0.543)}{175+2}} = \sqrt{0.009} = 0.095$$

$$(0.704 - 0.458) \pm 1.96(0.095)$$
$$= 0.246 \pm 0.186 = (0.06, 0.432)$$

We are 95% confident that the difference in the proportion of cell phone ownership between patients with brain cancer and those without brain cancer, is between 6% and 43%.

Slide 51 Stat 13, UCLA, Ivo Dinov

CI for the difference between probabilities

What does this mean?

Does this seem like a significant difference?

Can we say that based on this data it appears that owning a cell phone increases the probability of brain cancer?

Slide 52 Stat 13, UCLA, Ivo Dinov

Relative Risk

- The chi-square test is often referred to as a test of independence
- Another measure of dependence is relative risk
  - Allows researchers to compare probabilities in terms of their ratio (p_1 / p_2) rather than their difference ($p_1 - p_2$)
  - widely used in studies of public health
- In general a relative risk of 1 indicates that the probabilities of two events are the same.
  - A relative risk > 1 implies that there is increased risk
  - A relative risk < 1 implies that there is decreased risk

Slide 53 Stat 13, UCLA, Ivo Dinov

Relative Risk

Example: Brain Cancer and cell phone use (continued)

Cell Phone	Brain cancer: Yes	No	Total
Yes	18	80	98
No	7	95	102
Total	25	175	200

Thinking in terms of conditional probability again, but switching the conditional probability around…

$\hat{P}$ = (BC|CP) = 18/98 = 0.184

$\hat{P}$ = (BC|NCP) = 7/102 = 0.069

So the relative risk is 0.184 / 0.069 = 2.67

The risk of having brain cancer is more than 2.5 times greater for cell phone owners when compared to non-cell phone owners.

Slide 54 Stat 13, UCLA, Ivo Dinov

Odds Ratio

- Another way to compare two probabilities is in terms of odds
- If an event takes place with probability p, then the odds in favor of the event are p / (1 - p)
  - If event A|B has p = ½, then the odds are (1/2) / (1/2) =1 or
    1 to 1 (the probability that event A|B occurs is equal to the probability that it does not occur)
  - If event A|C has p = ¾, then the odds are (3/4) / (1/4) = 3 or 3 to 1 (the probability that event A|C occurs is three times as large as the probability that it does not occur)

Odds Ratio: OR

- The **odds ratio** is the ratio of odds for two probabilities

$$\hat{\theta} = \frac{\dfrac{\hat{P}(A|B)}{1-\hat{P}(A|B)}}{\dfrac{\hat{P}(A|C)}{1-\hat{P}(A|C)}}$$

		Brain cancer		
		Yes: A	**No**	**Total**
Cell Phone	**Yes**	18	80	98 **B**
	No	7	95	102 **C**
	Total	25	175	200

- In general an OR and it's relationship to 1 is similar to relative risk
  - An OR = 1indicates that the probabilities of two events are the same
  - An OR > 1 implies that there is increased risk
  - An OR < 1 implies that there is decreased risk

Odds Ratio

Example: Brain Cancer and cell phone use (continued)

		Brain cancer		
		Yes	**No**	**Total**
Cell Phone	**Yes**	18	80	98
	No	7	95	102
	Total	25	175	200

Calculate the odds of having brain cancer (A) for cell phone owners (B) compared to non-cell phone (C) owners.

$\hat{P}$ = (BC|CP) = 18/98 = 0.184

$\hat{P}$ = (BC|NCP) = 7/102 = 0.069

$$\hat{\theta} = \frac{\dfrac{0.184}{0.816}}{\dfrac{0.069}{0.931}} = \frac{0.225}{0.074} = 3.04$$

The odds of having brain cancer is about 3 times greater for cell phone owners when compared to non-cell phone owners.

Odds Ratio

- We could have compared the odds of owning a cell phone given that a patient had brain cancer versus an other cancer (i.e., the *column-wise* probabilities)
  $\hat{P}$ (CP|BC) = 18/25 = 0.72 versus $\hat{P}$ (CP|NBC) = 80/175 = 0.457
  However this does not seem as important scientifically
- But if we did calculate the OR of owning a cell phone given that a patient had brain cancer versus an other cancer we'd get:

$$\hat{\theta} = \frac{\dfrac{0.72}{0.28}}{\dfrac{0.457}{0.543}} = \frac{2.57}{0.842} = 3.05$$

  - Note that this OR comes out to be approximately equal!

Odds Ratio

- Shortcut formula for an odds ratio:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Now it is easier to see why the OR would be the same for the *rowwise* and *columnwise* probabilities!

Where the table structure looks like:

n_{11}	n_{12}
n_{21}	n_{22}

Relative Risk vs. Odds Ratio

- The formula and reasoning for the relative risk is a little bit easier to follow
  - In most cases the two measures are roughly equal to each other
- Odds ratios have an advantage over relative risk because they can be calculated no matter the row or column comparison
  - Relative risk runs into problems when the study design is a cohort study or a case-control design
  - Odds ratios are an approximation of relative risk
    OR = RR*(1-P_2)/(1-P_1)

Relative Risk vs. Odds Ratio

Example: Suppose a group of 200 people who have experienced a heart attack and 200 with no heart attack were asked if they were ever smokers.

		Heart Attack (HA)?	
		Yes	No
Ever Smoker (SMK)?	Yes	33	18
	No	167	182
	Total	200	200

- We can reasonably calculate $\hat{P}$(SMK|HA) = 33/200 = 0.165 and $\hat{P}$ (SMK|NHA) = 18/200 = 0.09
- However, the *row-wise* probabilities (incidence of heart attacks given that someone is a smoker or non-smoker) should not be estimated
  - Because the number of subject with and without heart attacks were predetermined in the study design
  - We have no information about the incidence of heart attacks

Relative Risk vs. Odds Ratio

- If we calculate the *column-wise* probabilities and the odds of smoking for those with heart attacks compared to no heart attacks, using $\hat{P}$ (SMK|HA) = 0.165 and $\hat{P}$ (SMK|NHA) = 0.09

  OR comes out to about 2.0

- Had we incorrectly calculated the *rowwise* probabilities and then the odds of heart attacks for people who smoke versus non-smokers, using $\hat{p}$ (HA|SMK) = 33/51 = 0.65 and $\hat{P}$ (HA|NSMK) = 167/349 = 0.48

However, the OR comes out to about 2.0
- Remember that the OR will come out to be approximately equal for row and column comparisons

Relative Risk vs. Odds Ratio

- Because these estimates of the odds ratio are the same for column-wise and row-wise probabilities (see p. 449)
- And we know that the odds ratio is an approximation of relative risk
- We can say that we estimate the relative risk of a heart attack is about 2 twice as great for those who smoke versus who do not smoke
  - Without incorrectly calculating the *row-wise* probabilities

Odds Ratio Confidence Interval

- Common to report odds ratios along with their CI
- One problem with our estimate of the odds ratio $\hat{\theta}$ is that it' sampling distribution is not normally distributed
  - To solve this we take the log of $\hat{\theta}$ and so that the sampling distribution of **ln($\hat{\theta}$)** is normally distributed

SE of ln ($\hat{\theta}$): $$SE_{\ln(\hat{\theta})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Where the table structure looks like:

n_{11}	n_{12}
n_{21}	n_{22}

Odds Ratio Confidence Interval

- A 100(1- α) CI for θ is

table structure to remember

n_{11}	n_{12}
n_{21}	n_{22}

$$\ln(\hat{\theta}) \pm Z_{\alpha/2}(SE_{\ln(\hat{\theta})})$$

Where $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$ and $SE_{\ln(\hat{\theta})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

HINT: You can use your t table to find certain values of $Z_{\alpha/2}$

Odds Ratio Confidence Interval

Example: Smoking and heart attack (continued)

		Heart Attack?	
		Yes	No
Ever Smoker?	Yes	33	18
	No	167	182
	Total	200	200

Calculate a 90% confidence interval for odds of smoking for heart attack subjects and non-heart attack subjects.

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{33*182}{18*167} = 1.998$$

$$SE_{\ln(\hat{\theta})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$
$$= \sqrt{\frac{1}{33} + \frac{1}{18} + \frac{1}{167} + \frac{1}{182}}$$
$$= \sqrt{0.09734} = 0.3120$$

Odds Ratio Confidence Interval

So the 90% CI for θ is

$$\ln(\hat{\theta}) = \ln(1.998) = 0.6921$$

$$\ln(\hat{\theta}) \pm Z_{\alpha/2}(SE_{\ln(\hat{\theta})})$$
$$0.6921 \pm Z_{0.05}(0.3120) =$$
$$0.6921 \pm 1.645(0.3120) =$$
$$(0.1789,\ \ 1.2053)$$

But right now this is transformed data (natural log) so we need to untransform it by taking the exponent of the CI

$$(e^{0.1789}, e^{1.2053}) = (1.196,\ \ 3.338)$$

Odds Ratio Confidence Interval

We are confident at the 0.10 level that the true odds of smoking for heart attack subjects and non-heart attack subjects are between 1.196 and 3.338

- So what does this actually mean?
- Does the zero rule work here?
- What if the CI came out to be (0.196,1.338)?