Categorical Data Analysis (Stat 3062)

BY

Getasew Berhanu

Dilla University

Department of Statistics
Chapter One Introduction

**Categorical Response**
This course will focus on categorical outcome data.

- **Outcome**: It can be a result of certain experiment or response of interest.
- **Response**: What is to be measured or obtained as an outcome of interest in the experiment.
- Categorical **Outcome** = Categorical **Response**
- Categorical **Variable**: Variable for which the measurement scale consists of a set of categories.

**Example**: Political Affiliation (Democrat, Republican, Liberal…), Disease Status (Well, Mild, Severe), Public Opinion (Yes or No)

- Each subject can be categorized in only one category

- Categorical Variable Can be:
  - Nominal
  - Ordinal
Nominal: Categorical variable having unordered categories
- Religion (Orthodox, Catholic, Protestant, Muslims)
- Gender (Male or Female)

Ordinal: Categorical variable having ordered categories
- Grading (Excellent, Very good, Good, Satisfactory, Fail)
- Income (High, Medium, Low)

Exercise: List another examples of ordinal and nominal variables (five for each)

Probability Distribution for Categorical Data
- Bernoulli
- Binomial
- Hypergeometric
- Poisson and Multinomial
Bernoulli trial:

- an experiment with two, and only two possible outcomes.

A random variable $X$ has a *Bernoulli$(p)$ distribution* if

$$X = \begin{cases} 
1 & \text{with probability } p \\
0 & \text{with probability } 1 - p,
\end{cases} \quad 0 \leq p \leq 1.$$ 

Binomial distribution:

- A random variable $X$ has *Binomial distribution* and it referred to as a *Binomial random variable* if and only if its probability distribution given by:

$$p(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n.$$
In general binomial distribution has the following characteristics:

- An experiment repeated \( n \) times.
- Only two possible outcomes: success (S) or Failure (F).
- \( P(S) \) (fixed at any trial).
- The \( n \)-trials are independent

**Mean:** \( E(X) = \mu = np \)

**Variance:** \( \text{Var}(X) = E(X - E(x))^2 = np \( 1 - p \) \)
Hypergeometric distribution

Suppose we have a large urn filled with $N$ balls that are identical in every way except that $M$ are red and $N - M$ are green. We reach in, blindfolded, and select $K$ balls at random (the $K$ balls are taken all at once, a case of sampling without replacement). What is the probability that exactly $x$ of the balls are red?

The total number of samples of size $K$ that can be drawn from the $N$ balls is $\binom{N}{K}$, as was discussed in Section 1.2.3. It is required that $x$ of the balls be red, and this can be accomplished in $\binom{M}{x}$ ways, leaving $\binom{N-M}{K-x}$ ways of filling out the sample with $K - x$ green balls. Thus, if we let $X$ denote the number of red balls in a sample of size $K$, then $X$ has a hypergeometric distribution given by

\begin{equation}
(3.2.2) \quad P(X = x|N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}, \quad x = 0, 1, \ldots, K.
\end{equation}
Poisson distribution

• a random variable \( X \) has Poisson distribution with parameter \( \lambda \) and it referred to as a Poisson random variable if and only if its probability distribution given by:

\[
p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}
\]

for \( x = 0, 1, 2, \ldots \).

\( \checkmark \) **Mean:** \( E(X) = \mu = \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} = \lambda \)

\( \checkmark \) **Variance:** \( Var(X) = E((X - E(x))^2) = \lambda \)
Exercise

1. What is the probability that the student will answer 8 out of 10 questions correctly. When the probability of answering a question correctly is 0.5?

2. 10 lamps were brought to an electrician for investigation, but, due to time constraint only 6 of them will be investigated. If only 3 out of the 10 lamps are defective, what is the probability that at least 2 out of the 6 lamps selected for investigation will be found defective?

3. On average 4.6 deaths occur every year at certain intersection due to car accident. What is the probability that at most 5 deaths will occur in any given year?
In general, the aim of this section will be to estimate parameters and to test hypotheses about the parameters.

In some cases, it is possible to get exact distribution for the test statistic or parameters. In other cases we should rely on large sample theory.
Inference for a Proportion

- The **population proportion** \( p \) is the number of elements in the entire population belonging to a certain category of interest divided by the total number of elements in the entire population.

- For instance, the proportion of female students in this class (number of female students in the class divided by the total number of students in the same class).

- The **sample proportion** \( \hat{p} \) is an estimate of the population proportion.

- The **sampling distribution** of \( \hat{p} \) approaches a normal distribution with mean \( p \) and standard error \( \sqrt{p(1-p)/n} \), when the sample size is large.

- A sample is considered large enough when both \( np \) and \( n(1-p) \) are greater than 5.

Large Sample Test

- The first step is **stating the hypotheses**.

  - \( H_0: p = p_0 \)  Vs  \( H_1: p \neq p_0 \)  or  \( H_1: p > p_0 \)  or  \( H_1: p < p_0 \)
Test Statistic

\[ Z_w = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1) \]

Setting Criteria for Rejecting \( H_0 \)

- Small difference: Do not reject
- Large difference: Reject

Collect data and compute the test statistic

Decide about \( H_0 \)

**Example:** A senator decides to have a final survey to decide on the equal rights amendment, if the proportion of registered voters supporting the amendment exceeds 0.6, the senator will vote for it. Test the hypothesis that the senator will indeed vote for it, if out of 750 opinion voters 495 of them support the amendment.

- \( H_0 : p = 0.6 \) Vs \( H_1 : p > 0.6 \)
- \( Z_w = 3.47 \quad \hat{p} = 0.66 \)
Critical value for the test statistic is 1.645
Since calculated value (3.47) > tabulated value(1.645), reject $H_0$
That means, the senator will vote for the amendment

**Confidence Interval**
- A confidence interval is a range of numbers believed to include an unknown population parameter. The interval is also a measure of the confidence we have that the interval does indeed contain the parameter of interest.
- A $(1-\alpha)100\%$ confidence interval for the population proportion $p$ is given by:
  $$\left[ \hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} ; \hat{p} + z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} \right]$$
- **Example**: Construct a 95% confidence interval for the voters who support the equality right amendment.
  - A 95% confidence interval for $p$ is $[0.626, 0.694]$
  - The senator is 95% confident that the proportion of voters who support the amendment will be anywhere in the interval (from 62.6% to 69.4%)
**Exact Test Statistic Confidence Intervals**

- Some times experiments have small samples, and the asymptotic (large sample) test statistic and confidence interval are very poor for properties in small samples. In this situation, exact test statistic and confidence interval can be obtained.

**One-Sided Exact Test Statistic**

- Suppose we want to test
  \[ H_0 : p = p_0 \quad \text{Vs} \quad H_1 : p > p_0 \]
- Test statistic (\(Y\) = the number of success out of \(n\) trials)
- Under \(H_0 : Y \sim \text{Bin}(n, p_0)\)
- If you observe \(y\) successes, the exact \(p\)-value is
  \[
p\text{-value} = P(Y \geq y | H_0 : p = p_0) = \sum_{j=y}^{n} \binom{n}{j} p_0^j (1 - p_0)^{n-j}\]

- **\(p\)-value** is the probability of obtaining a test statistic at least as extreme as the observed one given that the null hypothesis is true.
Other One-Sided Exact Test Statistic

Suppose we want to test

\[ H_0 : p = p_0 \quad \text{Vs} \quad H_1 : p < p_0 \]

If you observe \( y \) successes, the exact p-value is

\[
p\text{-value} = P(Y \leq y | H_0 : p = p_0) = \sum_{j=0}^{\frac{y}{n}} \binom{n}{j} p_0^j (1 - p_0)^{n-j}
\]

Two-Sided Exact Test Statistic

It is easy to compute the 2-sided p-value, for a symmetric distribution.

When the distribution is not symmetric, 2-sided exact p-value is trickier.

Suppose we want to test

\[ H_0 : p = p_0 \quad \text{Vs} \quad H_1 : p \neq p_0 \]

Calculate the probability of the observed result under the null

\[
\pi = \binom{n}{y} p_0^y (1 - p_0)^{n-y}
\]
Calculate the probabilities of all \( n+1 \) values that \( Y \) can take on

\[
\pi_j = \binom{n}{j} p_0^j (1 - p_0)^{n-j} \quad j = 0, \ldots, n
\]

Sum the probabilities \( \pi_j \) that are less than or equal to the observed probability \( \pi \) (to obtain the p-value for the two-sided exact test)

**Example:** Suppose \( n = 5 \), hypothesize \( p = 0.4 \) and we observe \( y = 3 \) successes. Calculate the exact p-values for the following hypotheses

1) \( H_1: p > 0.4 \) \hspace{1cm} 2) \( H_1: p \neq 0.4 \) \hspace{1cm} and \hspace{1cm} 3) \( H_1: p < 0.4 \)

**Solution:** \( P(Y = 0) = 0.0777, \quad P(Y = 1) = 0.2592, \quad P(Y = 2) = 0.3456, \quad P(Y = 3) = 0.2304, \quad P(Y = 4) = 0.0768, \quad P(Y = 5) = 0.0102 \)

1) \( p = 0.2304 + 0.0768 + 0.0102 = 0.3174 \)

2) \( P = 0.0777 + 0.2304 + 0.0768 + 0.0102 = 0.3951 \) \hspace{1cm} and \hspace{1cm} 3) Exercise
Of adults in the United States believe that a pregnant woman should be able to obtain an abortion? Let $\pi$ denote the proportion of the American adult population that responds “yes” to the question, “Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children.” We test $H_0: \pi = 0.50$ against the two-sided alternative hypothesis, $H_a: \pi = 0.50$.

Of 893 respondents to this question in 2002, 400 replied “yes” and 493 replied “no”. The sample proportion of “yes” responses was $p = 400/893 = 0.448$. For a sample of size $n = 893$, the null standard error of $p$ equals $\sqrt{(0.50)(0.50)/893} = 0.0167$. The test statistic is $z = (0.448 - 0.50)/0.0167 = -3.1$
The two-sided $P$-value is the probability that the absolute value of a standard normal variate exceeds 3.1, which is $P = 0.002$. There is strong evidence that, in 2002, $\pi < 0.50$, that is, that fewer than half of Americans favored legal abortion in this situation.

The 95% confidence interval equals $p = 0.448$ for $n = 893$, which is $0.448 \pm 0.033$, or $(0.415, 0.481)$.

We can be 95% confident that the population proportion of Americans in 2002 who favored legalized abortion for married pregnant women who do not want more children is between 0.415 and 0.481.
Chapter Two
Contingency Tables

- **Contingency Table**: A table in which the cells contain frequency counts of outcomes
- **Two-way Table**: A contingency table that cross classifies two variables
- **$I \times J$ Table**: A two-way table having $I$ rows and $J$ columns
- We will spend a lot of time talking about $(2 \times 2)$ tables
- Much of the statistical theory is more easily seen in $(2 \times 2)$ tables, and then generalizes to more complicated problems
- $(2 \times 2)$ tables may arise from different sampling plans (see below)
- **Prospective Study**:
  - Number on each \textit{treatment} fixed by design
  - Individuals are randomized to either of the treatments
- **The interest here is**:
  - To see the effect of the treatment on the response
- **Example**: Cold incidence among French Skiers (Pauling, Proceedings of the national Academy of Sciences, 1971)
<table>
<thead>
<tr>
<th>OUTCOME</th>
<th>COLD</th>
<th>NO COLD</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>VITAMIN C</td>
<td>17</td>
<td>122</td>
<td>139</td>
</tr>
<tr>
<td>NO VITAMIN C</td>
<td>31</td>
<td>109</td>
<td>140</td>
</tr>
<tr>
<td>TOTAL</td>
<td>48</td>
<td>231</td>
<td>279</td>
</tr>
</tbody>
</table>

- **Retrospective Study:** Not ethical to randomize patients in a prospective study
- Number of *cases and controls* (outcomes) are fixed by design and exposures are random
- **The interest here is:**
  To see that whether exposure vary among cases and controls
- **Example:** Alcohol Consumption and occurrence of oesophageal cancer (Tuyns et al., Bulletin of Cancer, 1974)
<table>
<thead>
<tr>
<th>STATUS</th>
<th>CASE</th>
<th>CONTROL</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>80+ (g/day)</td>
<td>96</td>
<td>109</td>
<td>205</td>
</tr>
<tr>
<td>0-79 (g/day)</td>
<td>104</td>
<td>666</td>
<td>770</td>
</tr>
<tr>
<td>TOTAL</td>
<td>200</td>
<td>775</td>
<td>975</td>
</tr>
</tbody>
</table>

- **Cross-sectional or Prevalence Study:**
- Sample subjects (*total number fixed*) and then cross-classify them on the basis of 2 variables
- **The interest here is:**
  To see the association between the two variables
- **Example:** General Social Survey (1984 SPSS Manual)
Joint, Marginal and Conditional Probabilities

- Joint Distribution: \( \pi_{ij} = P( X= i, \ Y= j) \)
- Marginal Distribution: \( \pi_{i+} = \sum_j \pi_{ij} \) and \( \pi_{+j} = \sum_i \pi_{ij} \)
- Conditional Distribution: \( \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \) and \( \pi_{i|j} = \frac{\pi_{ij}}{\pi_{+j}} \)
- Similar notation employed for samples with \( p_{ij} = \frac{n_{ij}}{n} \)
- The following table shows notation for the probabilities in a 2x2 table
<table>
<thead>
<tr>
<th>Row</th>
<th>1</th>
<th>2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_{11} )</td>
<td>( p_{12} )</td>
<td>( p_{1+} )</td>
</tr>
<tr>
<td></td>
<td>( (p_{1</td>
<td>1}) )</td>
<td>( (p_{2</td>
</tr>
<tr>
<td>2</td>
<td>( p_{21} )</td>
<td>( p_{22} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( (p_{1</td>
<td>2}) )</td>
<td>( (p_{2</td>
</tr>
<tr>
<td>Total</td>
<td>( p_{+1} )</td>
<td>( p_{+2} )</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Example (General Social Survey)**

- \( p_{11} = \frac{104}{901} = 0.115 \)
- \( p_{12} = \frac{391}{901} = 0.434 \)
- \( p_{21} = \frac{66}{901} = 0.073 \)
- \( p_{22} = \frac{340}{901} = 0.377 \)
- \( p_{1+} = \frac{495}{901} = 0.549 \)
- \( p_{2+} = \frac{406}{901} = 0.451 \)
- \( p_{11} = \frac{104}{495} = 0.210 \)
- \( p_{21} = \frac{391}{495} = 0.790 \)

**Exercise:** Compute the remaining
• In Table 2.1, belief in the afterlife is a response variable and gender is an explanatory variable. We therefore study the conditional distributions of belief in the afterlife, given gender.

• For females, the proportion of “yes” responses was $\frac{509}{625} = 0.81$ and the proportion of “no” responses was $\frac{116}{625} = 0.19$.

• The proportions $(0.81, 0.19)$ form the sample conditional distribution of belief in the afterlife. For males, the sample conditional distribution is $(0.79, 0.21)$. 
Screening Test

- They are diagnostic tests, almost all of such tests are not perfect.
- Therefore, there are 4 conditional probabilities fundamental to evaluating diagnostic procedures (Sensitivity, Specificity, Positive and Negative predictivity).

- **Sensitivity**: the proportion of diseased individuals detected as positive by the test
  \[ P(\text{Test} = + | \text{Disease} = +) \]
- **Specificity**: the proportion of healthy individuals detected as negative by the test
  \[ P(\text{Test} = - | \text{Disease} = -) \]
- **Positive predictivity**: the proportion of test positive individuals having the disease
  \[ P(\text{Disease} = + | \text{Test} = +) \]
- **Negative predictivity**: the proportion of test negative individuals not having the disease
  \[ P(\text{Disease} = - | \text{Test} = -) \]

- **Example**: Cytological test to screen women for cervical cancer
### Test Result

<table>
<thead>
<tr>
<th>Disease</th>
<th>Negative</th>
<th>Positive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>23362</td>
<td>362</td>
<td>23724</td>
</tr>
<tr>
<td>Positive</td>
<td>225</td>
<td>154</td>
<td>379</td>
</tr>
<tr>
<td>Total</td>
<td>23587</td>
<td>516</td>
<td>24103</td>
</tr>
</tbody>
</table>

- Sensitivity = 154/379 = 0.406
  - If a woman with cervical cancer is tested the chance is that the disease would go undetected
- Specificity = 23362/23724 = 0.985
  - If a woman with out cervical cancer is tested, the result would almost surely negative
- Positive Predictivity = 154/516 = 0.298
- Negative predictivity = 23362/23587 = 0.99
Independence

Two variables are statistically independent if the joint probabilities equal the product of their marginal probabilities:

\[ \pi_{ij} = \pi_i \pi_j \quad \forall i, j \]

or if the conditional distributions of \( Y \) are identical at each level of \( X \):

\[ \pi_{j|i} = \pi_j \quad \forall i \quad \text{or} \quad \pi_{j|1} = \cdots = \pi_{j|I} \quad \forall j \]

Comparing Proportion in 2x2 Tables

In general, we can form the following \((2 \times 2)\) table:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Outcome</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\ 1 \</td>
<td>\ 2 \</td>
</tr>
<tr>
<td>\ 1 \</td>
<td>( Y_1 )</td>
<td>( n_1 - Y_1 )</td>
</tr>
<tr>
<td>\ 2 \</td>
<td>( Y_2 )</td>
<td>( n_2 - Y_2 )</td>
</tr>
</tbody>
</table>

- Individuals are given treatment 1 or treatment 2
- Outcome is success or failure
Facts about the distribution

- $n_1$ and $n_2$ are fixed by design
- $Y_1$ and $Y_2$ are independent with distributions:
  
  $$Y_1 \sim Bin(n_1, \pi_1)$$
  
  $$Y_2 \sim Bin(n_2, \pi_2)$$

Question of Interest

- Does treatment affect outcome?
- Are treatment and outcome associated?
- Is the probability of success the same on both treatments?

Hypotheses

- The null hypothesis is
  
  $$H_0 : \pi_1 = \pi_2 = \pi$$

- and the alternative is
  
  $$H_1 : \pi_1 \neq \pi_2$$
Measuring Treatment Differences

- When $\pi_1 \neq \pi_2$, we want to quantify how the two probabilities are different.

- In other words, we want a single measure of how the treatments differ. The exact interpretation of these measures will be deferred to later.

- The measures:
  - Difference of proportions (Risk difference) \((\pi_1 - \pi_2)\)
  - Relative Risk (Risk Ratio) \((\pi_1/\pi_2)\)
  - Odds Ratio (Relative Odds) \((\pi_1/(1-\pi_1)) \div (\pi_2/(1-\pi_2))\)
General formula for Variance of Treatment Difference

Intuitively, the estimators for $\pi_1$ and $\pi_2$ should be the proportion of successes in the two groups, i.e.

$$p_1 = \frac{Y_1}{n_1}$$

$$p_2 = \frac{Y_2}{n_2}$$

These are the MLE’s. But, you can go through a lot of statistical theory to show that these are the MLE’s.

The MLE of a treatment difference

$$g(\pi_1) - g(\pi_2)$$

is then

$$g(p_1) - g(p_2).$$
Recall, the variance of a difference of two independent random variables is

\[ \text{Var}[g(\pi_1) - g(\pi_2)] = \text{Var}[g(\pi_1)] + \text{Var}[g(\pi_2)] \]

Then, to obtain the large sample variance, we can apply the delta method to \( g(\pi_1) \) to get \( \text{Var}[g(\pi_1)] \) and to \( g(\pi_2) \) to get \( \text{Var}[g(\pi_2)] \) and then sum the two.

The results (for the estimates) are summarized in the following table:

<table>
<thead>
<tr>
<th>TREATMENT</th>
<th>ESTIMATE</th>
<th>\text{Var}(\text{ESTIMATE})</th>
</tr>
</thead>
<tbody>
<tr>
<td>RISK DIFF</td>
<td>( p_1 - p_2 )</td>
<td>( \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} )</td>
</tr>
<tr>
<td>\log(RR)</td>
<td>\log(p_1/p_2)</td>
<td>( \frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2} )</td>
</tr>
<tr>
<td>\log(OR)</td>
<td>\log(\frac{p_1/(1-p_1)}{p_2/(1-p_2)})</td>
<td>\left[\frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)}\right] + \left[\frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}\right]</td>
</tr>
</tbody>
</table>

The estimated variance of the log-odds ratio can also be written as:

\[ \frac{1}{y_1} + \frac{1}{n_1 - y_1} + \frac{1}{y_2} + \frac{1}{n_2 - y_2} \]
Let's consider an example to support concepts discussed so far, and also it will be used as an input in following exercises.

**Example: Aspirin and Heart Attacks in Doctors**

Agresti (1996)

<table>
<thead>
<tr>
<th>Group</th>
<th>Myocardial Infarction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Placebo</td>
<td>189</td>
</tr>
<tr>
<td>Aspirin</td>
<td>104</td>
</tr>
<tr>
<td>Total</td>
<td>293</td>
</tr>
</tbody>
</table>

- Prospective study (Clinical Trial)
- Number on each treatment fixed by design
- Outcome is success or failure
- Rows are independent binomials
- Overall probability of heart attack in Doctors is low:
  \[
  \frac{293}{22071} = 1.33\% 
  \]

The disease is rare
Difference of Proportions or Risk Difference

The risk difference is the difference between the “success” probabilities for the two groups:

$$\pi_1 - \pi_2$$

The estimated risk difference is $$p_1 - p_2$$

Since $$Y_1$$ and $$Y_2$$ are independent binomials, we know (see interludium) that the estimated variance of the risk difference is:

$$\text{Var}(p_1 - p_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Therefore, to test $$H_0$$ or to construct confidence intervals, we can use the following statistic:

$$Z = \frac{(p_1 - p_2) - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1) \text{ (WALD)}$$
Example: Aspirin use and heart attack in doctors
For the aspirine intake example, construct a 95% confidence interval for the true difference \( \pi_1 - \pi_2 \).
Can we conclude that the intake of aspirin appears to diminish the risk of myocardial infarction?

Solution

\[
p_1 = \frac{189}{11034} = 0.0171
\]

\[
p_2 = \frac{104}{11037} = 0.0094
\]

\[
p_1 - p_2 = 0.0077
\]

\[
\sqrt{\frac{(.0171)(.9829)}{11034} + \frac{(.0094)(.9906)}{11037}} = 0.0015
\]

Thus, a 95% confidence interval for the true difference is:

\[
[(p_1 - p_2) - 1.96 \times \sqrt{\text{Var}(p_1 - p_2)}; (p_1 - p_2) + 1.96 \times \sqrt{\text{Var}(p_1 - p_2)}]
\]

\[
= [0.0077 - 1.96 \times 0.0015, 0.0077 + 1.96 \times 0.0015]
\]

\[
= [0.005, 0.011]
\]
The risk difference has the interpretation that the excess risk of a MI on placebo is 0.0077. This “fraction” is not very meaningful for rare diseases, but stated in terms of subjects, we can say that we would expect 77 more MIs in 10000 placebo subjects than in 10000 aspirin users.

data m.aspirin;
input aspirin outcome count;
cards;
0 1 189
0 0 10845
1 1 104
1 0 10933;
run;

proc freq data=m.aspirin order=data;
tables aspirin*outcome/riskdiff;
weight count;
run;

<table>
<thead>
<tr>
<th>Risk</th>
<th>ASE</th>
<th>(Asymptotic)</th>
<th>(Exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>0.017</td>
<td>0.001</td>
<td>0.015</td>
</tr>
<tr>
<td>Row 2</td>
<td>0.009</td>
<td>0.001</td>
<td>0.008</td>
</tr>
<tr>
<td>Total</td>
<td>0.013</td>
<td>0.001</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Difference 0.008 0.002 0.005 0.011
(Row 1 - Row 2)
Relative Risk

The relative risk \((RR)\) is the ratio of the “success” probabilities for the two groups:

\[
\frac{\pi_1}{\pi_2}.
\]

- Can take any nonnegative real number
- Skewed sampling distribution
- The log-relative risk is often used to alleviate the restrictions that the relative risk must be positive, i.e.

\[
\log(RR) = \log\left(\frac{\pi_1}{\pi_2}\right) = \log(\pi_1) - \log(\pi_2),
\]

where

\[-\infty \leq \log(RR) \leq \infty.\]

- Less skewed sampling distribution, closer to normality.
- Therefore, best to construct c.i. for \(\log(RR)\) and then transform back (by exponentiating) to obtain c.i. for \(RR\).

The estimated \(\log(RR)\) is \(\log(p_1/p_2)\).

The estimated variance of \(\log(p_1/p_2)\) is (see interludium):

\[
\frac{1 - p_1}{n_1p_1} + \frac{1 - p_2}{n_2p_2}
\]
Example:

For the aspirine intake example, construct a 95% confidence interval for the true relative risk \( \frac{\pi_1}{\pi_2} \).

**Solution:**

\[ p_1 = \frac{189}{11034} = 0.0171 \]
\[ p_2 = \frac{104}{11037} = 0.0094 \]
\[ RR = \frac{p_1}{p_2} = 1.818 \]
\[ \log(RR) = 0.598 \]

The estimated standard error for \( \log(RR) \) is 0.1212. Hence, a 95% confidence interval for \( \log(RR) \) is \([0.360; 0.836]\). Thus, a 95% confidence interval for the true RR is obtained by exponentiating the interval for \( \log(RR) \):

\[ [1.434; 2.306]. \]

The relative risk has the interpretation that individuals on placebo have almost twice (1.8) the risk (or probability) of a heart attack than individuals on Aspirin.
Odds Ratio

The odds ratio (OR) is the ratio of the "odds" of success versus failure for the two groups:

\[ \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}. \]

- Can take any nonnegative real number
- Skewed sampling distribution
- The log-odds ratio is often used to alleviate the restrictions that the odds ratio must be positive, i.e.

\[
\log(OR) = \log\left(\frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}\right) \\
= \log\left(\frac{\pi_1}{1 - \pi_1}\right) - \log\left(\frac{\pi_2}{1 - \pi_2}\right) \\
= \text{logit}(\pi_1) - \text{logit}(\pi_2),
\]

where

\[-\infty \leq \log(OR) \leq \infty.\]

- Less skewed distribution closer to normality
The estimated OR is: \( \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \).

The estimated variance of \( \log \left( \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \right) \) is
\[
\left[ \frac{1}{y_1} + \frac{1}{n_1-y_1} \right] + \left[ \frac{1}{y_2} + \frac{1}{n_2-y_2} \right].
\]

What does values of OR represent?

- OR=1 corresponds with independence
- When \( 1 < OR < \infty \), the odds of success are higher in group 1 than in group 2. Thus, subjects in the first group are more likely to have successes than subjects in group 2, that is \( \pi_1 > \pi_2 \).
- When \( 0 < OR < 1 \), the odds of success are smaller in group 1 than in group 2. Thus, subjects in the first group are less likely to have successes than subjects in group 2, that is \( \pi_1 < \pi_2 \).
- Values of OR farther from 1 in a given direction represent stronger level of association.
Example:
For the aspirine intake example, construct a 95% confidence interval for the true odds ratio \( \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \).

Solution:
\[
p_1 = \frac{189}{11034} = 0.0171 \\
p_2 = \frac{104}{11037} = 0.0094 \\
OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = 1.832 \\
\log(OR) = 0.605
\]
The estimated standard error for \( \log(OR) \) is 0.1228.
Hence, a 95% confidence interval for \( \log(OR) \) is [0.364; 0.846].
Thus, a 95% confidence interval for the true OR is obtained by exponentiating the interval for \( \log(OR) \):

\[ [1.440; 2.331]. \]
The odds ratio has the interpretation that individuals on placebo have almost twice (1.8) the odds of a heart attack versus no heart attack than individuals on Aspirin.
Properties of OR

- When order of rows or columns is reversed, the new value of OR is the inverse of the original value.
- When the orientation of the table is reversed (rows become columns and vice versa), the OR does not change. This in contrast to the relative risk, which does not treat the variables symmetrically!
- When cell counts are very small or any zero cell counts occur, it is preferred to use an amended estimator for the true OR, by adding 1/2 to each cell count.

Summary

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Estimated Standard Error</th>
<th>Z-Statistic (Est/SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Diff</td>
<td>0.0077</td>
<td>0.00154</td>
<td>5.00</td>
</tr>
<tr>
<td>log(RR)</td>
<td>0.598</td>
<td>0.1212</td>
<td>4.934</td>
</tr>
<tr>
<td>log(OR)</td>
<td>0.605</td>
<td>0.1228</td>
<td>4.927</td>
</tr>
</tbody>
</table>

- In each case, we reject the null, and the Z-statistic is about 5.
- The WALD test statistic using the Risk difference, log OR and log RR are slightly different.
Relation Ship b/n OR and RR

\[ OR = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} \]

\[ = \left( \frac{p_1}{p_2} \right) \left[ \frac{1 - p_2}{1 - p_1} \right] \]

\[ = RR \left[ \frac{1 - p_2}{1 - p_1} \right] \]

- When the disease is “rare” (such as in the aspirin example),

\[ \left[ \frac{1 - p_2}{1 - p_1} \right] \approx 1, \text{ and } OR \approx RR \]

- In the example, the estimated values for OR and RR are 1.832 and 1.818 respectively, i.e. they are almost identical.
The Chi-Squared Distribution

- Specified by its degrees of freedom \((df)\).
  - Mean of chi-squared distribution = \(df\).
  - Variance of chi-squared distribution = \(2df\).

- Defined only for nonnegative values.

- Skewed to the right.

- Becomes more “bell-shaped” and is concentrated around larger values, as \(df\) increases.
Chi-squared Tests of Independence in a Two-way Contingency Tables

The null hypothesis of statistical independence in a general $I \times J$ 2-way table has following form:

$$H_0: \pi_{ij} = \pi_i \pi_j$$

for all $i$ and $j$.

How do we proceed?

Intuitive Outline:
- In order to judge whether the data contradict $H_0$, we will compare the sample cell counts to the expected frequencies under the $H_0$.
- The larger the “Observed - Expected” differences, the stronger the evidence against $H_0$.
- The test statistics used to make such comparisons have large-sample Chi-squared distribution.
Pearson Statistic

The Pearson chi-squared statistic for testing \( H_0 \) is

\[
X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},
\]

where

- \( O_{ij} \) is the observed cell count,
- \( E_{ij} \) is the estimated expected count under the null hypothesis in the \( ij \)-th cell of a \( (I \times J) \) table.

Properties

- Pearson’s Chi-Square measures the discrepancy between the observed counts, and the estimated expected count under the null.
- If they are similar, you expect the statistic to be small, and for us not to reject the null.
- The minimum value of 0 is obtained when all \( O_{ij} \)'s are equal to the \( E_{ij} \)'s.
- The \( X^2 \)-statistic has approximately a chi-squared distribution for “large” sample sizes (where “large” means \( E_{ij} \geq 5 \)).