Unit 9: Inferences for Proportions and Count Data

Statistics 571: Statistical Methods
Ramón V. León

Large Sample Confidence Interval for Proportion

Note that \( \frac{\hat{p} - p}{\sqrt{pq/n}} \approx N(0,1) \) and \( \frac{\hat{p} - p}{\sqrt{pq/n}} \approx N(0,1) \) if \( n \) is large

\( (q = 1 - p, \ \ n\hat{p} \geq 10 \text{ and } n\hat{q} \geq 10) \)

It follows that:

\[
P \left( -z_{\alpha/2} \leq \frac{(\hat{p} - p)}{\sqrt{\hat{p}\hat{q}/n}} \leq z_{\alpha/2} \right) \approx 1 - \alpha
\]

Confidence interval for \( p \):

\[
\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}
\]
A Better Confidence Interval for Proportion

Use this probability statement

\[ P \left( -z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{pq/n}} \leq z_{\alpha/2} \right) \approx 1 - \alpha \]

Solve for \( p \) using quadratic equation

CI for \( p \):

\[ \hat{p} + \frac{z^2}{2n} - \sqrt{\frac{p(1 - p)}{n}} + \frac{z^4}{4n^2} \leq p \leq \hat{p} + \frac{z^2}{2n} + \sqrt{\frac{p(1 - p)}{n}} + \frac{z^4}{4n^2} \]

\[ z = z_{\alpha/2}. \]

EXAMPLE 9.1

Time magazine reported a telephone poll survey of 800 adults, of whom 45% stated that they had guns in their homes. The margin of error for this sample estimate was reported as 3.5%. This means that a 95% CI for the proportion \( p \) of homes in the population that have guns is 45% ± 3.5% = [41.5%, 48.5%]. Verify the calculation of this CI.

Here \( \hat{p} = 0.45, \hat{q} = 0.55, n = 800, z_{0.025} = 1.960 \). Substituting these values in (9.1) yields

\[ \left[ 0.45 \pm 1.960 \sqrt{\frac{(0.45)(0.55)}{800}} \right] = [0.415, 0.485] \]

which is the desired CI with a margin of error of 3.5%.

The 95% confidence limits using (9.3) can be calculated as follows:

\[ \text{Lower limit} = \frac{0.45 + \frac{1.96^2}{800} - \sqrt{\frac{(0.45)(0.55)(1.96)^2}{800} + \frac{1}{4} \left( \frac{1.96^2}{800} \right)^2}}{1 + \frac{1.96^2}{800}} = 0.416. \]

\[ \text{Upper limit} = \frac{0.45 + \frac{1.96^2}{800} + \sqrt{\frac{(0.45)(0.55)(1.96)^2}{800} + \frac{1}{4} \left( \frac{1.96^2}{800} \right)^2}}{1 + \frac{1.96^2}{800}} = 0.485. \]

These limits are almost the same as the approximate limits.
CI for Proportion in JMP

Value column has two categories. Can you imagine a situation where one would have more categories in this column?

12/15/2008 Unit 9 - Stat 571 - Ramón V. León

CI for Proportion in JMP

12/15/2008 Unit 9 - Stat 571 - Ramón V. León
Sample Size Determination for a Confidence Interval for Proportion

Want \((1-\alpha)\)-level two-sided CI:
\[
\hat{p} \pm E \quad \text{where } E \text{ is the margin of error.} \quad \text{Then } E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}.
\]

Solving for \(n\) gives
\[
n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}\hat{q} \quad \text{(Formula 9.4)}
\]

Largest value of \(p\hat{q} = \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right) = \frac{1}{4}\) so conservative sample size is:
\[
n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \frac{1}{4} \quad \text{(Formula 9.5)}
\]

Example 9.2: Presidential Poll

Suppose that a nationwide survey is planned to estimate the proportion \(p\) of people who have a favorable approval rating of the President. This proportion is to be estimated within a margin of error of \(\text{three percentage points}\), using a 95% CI. What sample size should be planned for the survey?

If reliable information is not available to make a prior guess \(p^*\), then, using (9.5), the required sample size is
\[
n = \left(\frac{1.96}{.01}\right)^2 \frac{1}{4} = 9604
\]

If a prior guess \(p^* = 0.65\) is available from the previous survey, then, using (9.4), the required sample size is
\[
n = \left(\frac{1.96}{0.03}\right)^2 (0.65)(0.35) = 971.07 \quad \text{or} \quad 972.
\]

Note from formula (9.5) that the sample size varies inversely with the square of the margin of error \(E\). Thus if \(E\) is specified to be 1%, then the sample size would increase ninefold to \(n = 9604\). Many nationwide surveys use sample sizes in the range of 1000 to 10,000, which roughly correspond to a margin of error between 3% to 1%, respectively, for a 95% CI.

Threefold increase in precision requires ninefold increase in sample size
Largest Sample Hypothesis Test on Proportion

\[ H_0 : p = p_0 \ vs. \ H_1 : p \neq p_0 \]

Best test statistics: \( z = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}} \)

• Dual relationship between CI and test of hypothesis holds if the better confidence interval is used.

• There is an exact test that can be used when the sample size is small (given in Section 9.1.3). We do not cover it.

Basketball Problem: z-test

**Example 9.3**

A professional basketball player has been a 70% free throw shooter (i.e., he has a long run average of making 70% of the attempted free throws). In the current season he made 300 of the 400 (75%) attempted free throws. Has his free throw percentage really improved, or can this increase be attributed to chance?

We view the 400 free throws as a random sample of i.i.d. Bernoulli trials with success probability \( p \), which is the player’s “true” shooting percentage. The hypotheses are formulated as \( H_0 : p = 0.70 \) (the player’s true free throw shooting percentage has stayed the same) vs. \( H_1 : p > 0.70 \) (the player’s true free throw shooting percentage has improved). The \( z \)-statistic for testing \( H_0 \) is

\[ z = \frac{0.75 - 0.70}{\sqrt{(0.70)(0.30)/400}} = \frac{300 - (400)(0.70)}{\sqrt{400(0.70)(0.30)}} = 2.182. \]

Thus his observed free throw shooting percentage this season is 2.182 standard deviations higher than his past percentage. The chance that an improvement this large or larger will be observed even if the true percentage for the current season had stayed the same (= 0.70) is given by the \( P \)-value \( = 1 - \Phi(2.182) = 0.015. \)
Test for Proportion in JMP:
Baseball Problem

“Value” has two categories.

Test for Proportion in JMP:
Basketball Problem
Sample Size for Z-Test of Proportion

\[ H_o : p \leq p_0 \text{ vs. } H_1 : p > p_0 \]

Suppose that the power for rejecting \( H_0 \) must be at least \( 1 - \beta \) when the true proportion is \( p = p_1 > p_0 \).

Let \( \delta = p_1 - p_0 \). Then

\[
n = \left[ \frac{z_{\alpha} \sqrt{p_0 q_0} + z_{\beta} \sqrt{p_1 q_1}}{\delta} \right]^2
\]

Test based on:

\[
z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}
\]

Replace \( z_{\alpha} \) by \( z_{\alpha/2} \) for two-sided test sample size.

Example 9.4: Pizza Testing

\[ H_0 : \text{Can't tell two pizzas Apart}, \]
\[ H_1 : \text{Can tell pizzas apart} \]

\( \alpha = .10, \) We wants \( \beta = .25 \) when \( p = 0.5 \pm 0.1 \)

\[
n = \left[ \frac{z_{\alpha/2} \sqrt{p_0 q_0} + z_{\beta} \sqrt{p_1 q_1}}{\delta} \right]^2
\]

\[
= \left[ \frac{1.645 \sqrt{(0.5)(0.5)} + 0.675 \sqrt{(0.6)(0.4)}}{0.1} \right]^2 = 132.98 \approx 133
\]
Multinomial Test of Proportions

Example 9.10  (Testing the Uniformity of Random Digits)

Algorithms are used in computers to generate random digits. Many tests are available to test for randomness of the digits, e.g., no systematic runs or cyclic patterns. We consider a very simple test that checks whether all 10 digits occur with equal probability. This is equivalent to checking whether the digits follow a uniform distribution on the integers 0, 1, ..., 9. The following counts are observed from 100 successive random digits:

<table>
<thead>
<tr>
<th>Digit</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Count</td>
<td>12</td>
<td>7</td>
<td>12</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>7</td>
<td>13</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

Test, using $\alpha = .10$, whether these counts are in agreement with the uniform distribution model.

Denote by $p_i$ the probability of occurrence of digit $i$ ($i = 0, 1, \ldots, 9$). The null hypothesis corresponding to the uniform distribution model is

$$H_0: p_0 = p_1 = \cdots = p_9 = \frac{1}{10}.$$
Comparing Two Proportion:
Independent Sample Design

If $n_1p_1$, $n_1q_1$, $n_2p_2$, $n_2q_2 \geq 10$, then

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \approx N(0,1)$$

Confidence Interval:

$$\hat{p}_1 - \hat{p}_2 \pm z_{a/2} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{a/2} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Test for Equality of Proportions (Large $n$)
Independent Sample Design

$H_0 : p_1 = p_2$ vs. $H_1 : p_1 \neq p_2$

Test statistics: $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x + y}{n_1 + n_2}$

• There is small sample test called Fisher’s exact test. See JMP output latter.
• See Example 9.5 for application to the Salk Polio Vaccine Trial
• There a test for Matched Pair Design in Section 9.2.2. Please read Example 9.9 to see its application for testing the effectiveness of presidential debates. Do voters change their minds about candidates because of debates?
Example 9.6 – Comparing Two Leukemia Therapies

Table 9.2 Leukemia Trial Data

<table>
<thead>
<tr>
<th>Drug Group</th>
<th>Success</th>
<th>Failure</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prednisone</td>
<td>14</td>
<td>7</td>
<td>21</td>
</tr>
<tr>
<td>Prednisone + VCR</td>
<td>38</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>Column Total</td>
<td>52</td>
<td>11</td>
<td>63</td>
</tr>
</tbody>
</table>

The hypotheses are formulated as $H_0: p_1 = p_2$ vs. $H_1: p_1 \neq p_2$, where $p_1$ is the success probability of Prednisone and $p_2$ is the success probability of Prednisone + VCR. We will use the statistic (9.12). First calculate

$$\hat{p}_1 = \frac{14}{21} = 0.667, \quad \hat{p}_2 = \frac{38}{42} = 0.905, \quad \hat{p} = \frac{14 + 38}{21 + 42} = 0.825, \quad \text{and} \quad \hat{q} = 1 - 0.825 = 0.175.$$

Therefore

$$z = \frac{0.667 - 0.905}{\sqrt{0.825(0.175)\left(\frac{1}{21} + \frac{1}{42}\right)}} = -2.347.$$

The two-sided $P$-value of the $z$-test equals $2[1 - \Phi(2.347)] = 0.019$, which is significant at the .05 level but not at the .01 level.

Test for Equality of Proportions in JMP: Example 9.6
Example 9.6 JMP Output

Test for Equality of Proportions in JMP

Recall that the P-value of the two-sided z-test was calculated to be 0.019

\[ z^2 = (-2.347)^2 \]

Less significant result
Inferences for Two-Way Count Data

Table 4.7 Individuals Cross-Classified by Income and Job Satisfaction

<table>
<thead>
<tr>
<th>x: Income (US $)</th>
<th>y: Job Satisfaction</th>
<th>Very Dissatisfied</th>
<th>Little Dissatisfied</th>
<th>Moderately Satisfied</th>
<th>Very Satisfied</th>
<th>Row Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 6000</td>
<td></td>
<td>20</td>
<td>24</td>
<td>80</td>
<td>82</td>
<td>206</td>
</tr>
<tr>
<td>6000–15,000</td>
<td></td>
<td>22</td>
<td>38</td>
<td>104</td>
<td>125</td>
<td>289</td>
</tr>
<tr>
<td>15,000–25,000</td>
<td></td>
<td>13</td>
<td>28</td>
<td>81</td>
<td>113</td>
<td>235</td>
</tr>
<tr>
<td>&gt; 25,000</td>
<td></td>
<td>7</td>
<td>18</td>
<td>54</td>
<td>92</td>
<td>171</td>
</tr>
<tr>
<td>Column Sum</td>
<td></td>
<td>62</td>
<td>108</td>
<td>319</td>
<td>412</td>
<td>901</td>
</tr>
</tbody>
</table>

Sampling Model 1: Multinomial Model (Total Sample Size Fixed)
Sample of 901 from a single population that is then cross-classified

The null hypothesis is that X and Y are independent:

\[ H_0: p_{ij} = P(X = i, Y = j) = P(X = i)P(Y = j) = p_i p_j \]

for all \(i, j\)

\[ \frac{np}{901} \begin{pmatrix} 62 \\ 901 \end{pmatrix} \begin{pmatrix} 206 \\ 901 \end{pmatrix} = \frac{62 \times 206}{901} = 14.18 \]

Cell 1,1

12/15/2008 Unit 9 - Stat 571 - Ramón V. León 23

Sampling Model 1 (Total Sample Size Fixed)

Table 9.10 Observed and Expected Frequencies for Income-Job Satisfaction Data

<table>
<thead>
<tr>
<th>Income (U.S. $)</th>
<th>Very Dissatisfied</th>
<th>Little Dissatisfied</th>
<th>Moderately Satisfied</th>
<th>Very Satisfied</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 6000</td>
<td>20</td>
<td>24</td>
<td>80</td>
<td>82</td>
<td>206</td>
</tr>
<tr>
<td>6000–15,000</td>
<td>22</td>
<td>38</td>
<td>104</td>
<td>125</td>
<td>289</td>
</tr>
<tr>
<td>15,000–25,000</td>
<td>13</td>
<td>28</td>
<td>81</td>
<td>113</td>
<td>235</td>
</tr>
<tr>
<td>&gt; 25,000</td>
<td>7</td>
<td>18</td>
<td>54</td>
<td>92</td>
<td>171</td>
</tr>
<tr>
<td>Column Total</td>
<td>62</td>
<td>108</td>
<td>319</td>
<td>412</td>
<td>901</td>
</tr>
</tbody>
</table>

Estimated Expected Frequency = \( \frac{np}{901} \begin{pmatrix} 62 \\ 901 \end{pmatrix} \begin{pmatrix} 206 \\ 901 \end{pmatrix} = \frac{62 \times 206}{901} = 14.18 \)

Cell 1,1

12/15/2008 Unit 9 - Stat 571 - Ramón V. León 24
Chi-Square Statistics

Next, the \( \chi^2 \)-statistic is calculated as

\[
\chi^2 = \frac{(20 - 14.18)^2}{14.18} + \frac{(24 - 24.69)^2}{24.69} + \frac{(80 - 72.93)^2}{72.93} + \frac{(82 - 94.20)^2}{94.20} \\
+ \frac{(22 - 19.89)^2}{19.89} + \frac{(38 - 34.64)^2}{34.64} + \frac{(104 - 102.32)^2}{102.32} + \frac{(125 - 132.15)^2}{132.15} \\
+ \frac{(13 - 16.17)^2}{16.17} + \frac{(28 - 28.17)^2}{28.17} + \frac{(81 - 83.20)^2}{83.20} + \frac{(113 - 107.46)^2}{107.46} \\
+ \frac{(7 - 11.77)^2}{11.77} + \frac{(18 - 20.50)^2}{20.50} + \frac{(54 - 60.54)^2}{60.54} + \frac{(92 - 78.19)^2}{78.19} \\
= 2.393 + 0.019 + 0.684 + 1.579 \\
+ 0.225 + 0.326 + 0.028 + 0.387 \\
+ 0.622 + 0.001 + 0.058 + 0.286 \\
+ 1.931 + 0.304 + 0.707 + 2.438 \\
= 11.989.
\]

The d.f. for this statistics is \((4-1)(4-1) = 9\). Since \(\chi^2_{0.05} = 16.919\) the calculated \(\chi^2 = 11.989\) is not sufficiently large to reject the hypothesis of independence at \(\alpha = 0.05\) level.

In general df = (r-1)(c-1)
where c is the number of columns and r is the number of rows.

Chi-Square Test

Critical Value

The d.f. for this \( \chi^2 \) - statistics is \((4-1)(4-1) = 9\). Since \(\chi^2_{0.05} = 16.919\) the calculated \(\chi^2 = 11.989\) is not sufficiently large to reject the hypothesis of independence at \(\alpha = 0.05\) level.

In general df = (r-1)(c-1)
where c is the number of columns and r is the number of rows.
JMP Analysis

Column can only take certain values

- 6
- 15
- 25

Shows no significance

Test ChiSquare Prob>ChiSq
Likelihood Ratio 12.037 0.2112
Pearson 11.989 0.2140
JMP Analysis

Note that most of the contribution to the chi-square statistics comes from the corner cells. Lack of significance in the chi-square statistics is the result of the low contribution to the chi-square statistic coming from the center cells.

<table>
<thead>
<tr>
<th>Income ($)</th>
<th>Very Dis</th>
<th>Little Dis</th>
<th>Mod Sat</th>
<th>Very Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 6</td>
<td>20</td>
<td>24</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>6 - 15</td>
<td>22</td>
<td>38</td>
<td>104</td>
<td>125</td>
</tr>
<tr>
<td>15 - 25</td>
<td>13</td>
<td>26</td>
<td>51</td>
<td>112</td>
</tr>
<tr>
<td>&gt; 25</td>
<td>7</td>
<td>18</td>
<td>56</td>
<td>92</td>
</tr>
</tbody>
</table>

Restricting our chi-square analysis to the corner cells shows a strong relationship between income and level of satisfaction.
Product Multinomial Model: Row Totals Fixed

<table>
<thead>
<tr>
<th>Type of Cities</th>
<th>Returned</th>
<th>Kept</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Cities</td>
<td>21</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td>Suburbs</td>
<td>18</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>Medium Cities</td>
<td>17</td>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td>Small Cities</td>
<td>24</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>Column Total</td>
<td>80</td>
<td>40</td>
<td>120</td>
</tr>
</tbody>
</table>

Notice that it does not make any sense to say that for the population 21 out of 80 returned wallets came from Big Cities.

Homework Problem 9.28

Product Multinomial Model: Row Totals Fixed

<table>
<thead>
<tr>
<th>Drug Group</th>
<th>Success</th>
<th>Failure</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prednisone</td>
<td>14</td>
<td>7</td>
<td>21</td>
</tr>
<tr>
<td>Prednisone+VCR</td>
<td>38</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>Column Total</td>
<td>52</td>
<td>11</td>
<td>63</td>
</tr>
</tbody>
</table>

Sampling Model 2: Product Multinomial

Total number of patients in each drug group is fixed.

• The null hypothesis is that the probability of column response (success or failure) is the same, regardless of the row population:

\[ H_0 : P(Y = j \mid X = i) = p_j \]
Chi-Square Statistics

\[ \chi^2 = \sum_{i=1}^{c} \left( \frac{(n_i - e_i)^2}{e_i} \right) \]

which has \((2 - 1)(2 - 1) = 1\) d.f. The \(P\)-value equals 0.019, which indicates a statistically significant difference between the success rates for the drug groups, with a higher success rate for the Prednisone+VCR group \((p_2 > p_1)\).

### JMP Analysis

<table>
<thead>
<tr>
<th>Drug Group</th>
<th>Result</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prednisone</td>
<td>Success</td>
<td>14</td>
</tr>
<tr>
<td>Prednisone</td>
<td>Failure</td>
<td>7</td>
</tr>
<tr>
<td>Prednisone+VCR</td>
<td>Success</td>
<td>36</td>
</tr>
<tr>
<td>Prednisone+VCR</td>
<td>Failure</td>
<td>4</td>
</tr>
</tbody>
</table>

12/15/2008 Unit 9 - Stat 571 - Ramón V. León
Recall Slide 19:

\[ z^2 = (-2.347)^2 = 5.5084 \]

Remarks About Chi-Square Test

- The distribution of the chi-square statistics under the null hypothesis is approximately chi-square only when the sample sizes are large
  - The rule of thumb is that all expected cell counts should be greater than 1 and
  - No more than 1/5th of the expected cell counts should be less than 5.
- Combine sparse cell (having small expected cell counts) with adjacent cells. Unfortunately, this has the drawback of losing some information.
Odds Ratio as a Measure of Association for a 2x2 Table

Sampling Model I: Multinomial

$$\psi = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$$

The numerator is the odds of the column 1 outcome vs. the column 2 outcome for row 1, and the denominator is the same odds for row 2, hence the name “odds ratio”

Odds Ratio as a Measure of Association for a 2x2 Table

Sampling Model II: Product Multinomial

$$\psi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

The two column outcomes are labeled as “success” and “failure,” then $\psi$ is the odds of success for the row 1 population vs. the odds of success for the row 2 population.
Odds Ratio as a Measure of Association for a 2x2 Table

\[
\hat{\psi} = \frac{\left( \frac{n_{11}}{n_{11} + n_{12}} \right) \left( \frac{n_{22}}{n_{21} + n_{22}} \right)}{\left( \frac{n_{21}}{n_{11} + n_{12}} \right) \left( \frac{n_{12}}{n_{21} + n_{22}} \right)} = \frac{14}{21} \times \frac{42}{42} = \frac{14}{7} \times \frac{7}{4} = 0.2105
\]

Confidence Interval: [0.053, 0.831]

Case-Control Studies: The Odds Ratio Approximates the Relative Risk if the Disease is Rare

**Case-Control Studies.** The research question was whether there is an association between intramuscular vitamin K and risk of childhood leukemia. The findings were that 69/107 leukemia cases and 63/107 controls had received intramuscular vitamin K. A two-by-two table of these findings is as follows:

<table>
<thead>
<tr>
<th>Predictor Variable: Medication History</th>
<th>Outcome Variable: Diagnosis</th>
<th>Childhood Leukemia</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM vitamin K</td>
<td></td>
<td>69((d))</td>
<td>63((d))</td>
</tr>
<tr>
<td>No IM vitamin K</td>
<td></td>
<td>48((c))</td>
<td>54((c))</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>107</td>
<td>107</td>
</tr>
</tbody>
</table>

Relative risk \(= \frac{ad}{bc} = \frac{69 \times 54}{63 \times 48} = 1.2\)

Because the disease (leukemia in this instance) is rare, the odds ratio provides a good estimate of the relative risk.
JMP Output for Case-Control Study

How to Do It in JMP
Why the Odds Ratio Can Be Used as an Estimate for Relative Risk in a Case-Control Study

The data in a case-control study represent two samples: The cases are drawn from a population of people who have the disease and the controls from a population of people who do not have the disease. The predictor variable is measured, and the following two-by-two table produced:

<table>
<thead>
<tr>
<th>Disease</th>
<th>No Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk factor present</td>
<td>a</td>
</tr>
<tr>
<td>Risk factor absent</td>
<td>c</td>
</tr>
</tbody>
</table>

If this two-by-two table represented data from a cohort study, then the incidence of the disease in those with the risk factor would be \(a \div (a + b)\) and the relative risk would be simply \(a \div c \div (c + d)\). However, it is not appropriate to compute either incidence or relative risk in this way because the two samples are not drawn from the population in the same proportions. Usually, there are roughly equal numbers of cases and controls in the study samples but many fewer cases than controls in the population. Instead, relative risk in a case-control study can be approximated by the odds ratio, computed as the cross-product of the two-by-two table, \(ad \div cb\).

The basis for this extremely useful fact cannot be understood intuitively, but is relatively easy to demonstrate algebraically. Consider the situation for the full population, represented by \(a', b', c',\) and \(d'\).

<table>
<thead>
<tr>
<th>Disease</th>
<th>No Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk factor present</td>
<td>(a')</td>
</tr>
<tr>
<td>Risk factor absent</td>
<td>(c')</td>
</tr>
</tbody>
</table>

Here it is appropriate to calculate the risk of disease among people with the risk factor as \(a' \div (a' + b')\), the risk among those without the risk factor as \(c' \div (c' + d')\), and the relative risk as \([a' \div (a' + b')] \div [c' \div (c' + d')]\). We have already discussed the fact that \(a' \div (a' + b')\) is not equal to \(a \div (a + b)\). However, if the disease is relatively uncommon (as most are), then \(a'\) is much smaller than \(b'\), and \(c'\) is much smaller than \(d'\). This means that \(a' \div (a' + b')\) is closely approximated by \(a' \div b'\) and that \(c' \div (c' + d')\) is closely approximated by \(c' \div d'\). Thus the relative risk of the population can be approximated as follows:

\[
\frac{a'}{(a' + b')} \approx \frac{a'}{b'}
\]

\[
\frac{c'}{(c' + d')} \approx \frac{c'}{d'}
\]
The latter term is the odds ratio of the population (literally, the ratio of the odds of disease in those with the risk factor, \( \frac{a'}{b'} \), to the odds of disease in those without the risk factor, \( \frac{c'}{d'} \)). This can be rearranged as the cross-product:

\[
\frac{(a') d'}{(c') (b')}
\]

However, \( \frac{a'}{c'} \) in the population equals \( \frac{a}{c} \) in the sample if the cases are representative of all cases in the population (i.e., have the same prevalence of the risk factor). Similarly, \( \frac{b'}{d'} \) equals \( \frac{b}{d} \) if the controls are representative.

Thus the population parameters in this last term can be replaced by the sample parameters, and we are left with the fact that the odds ratio observed in the sample, \( \frac{ad}{bc} \), is a close approximation of the relative risk in the population, \( \frac{[a'/ (a' + b')]/[c'/ (c' + d')]}{[a'/ (a' + b')]/[c'/ (c' + d')]} \), provided that the disease is rare and sampling error (systematic as well as random) is small.

---

**Do You Need to Know More?**

578 Categorical Data Analysis (3) Log-linear analysis of multidimensional contingency tables. Logistic regression. Theory, applications, and use of statistical software. Prereq: 1 yr graduate-level statistics, regression analysis and analysis of variance, or consent of instructor. Sp