Chapter 10: Categorical Data

October 14, 2015

Types of Variables

Variable – a characteristic of an individual. A variable can take different values for different individuals.

Two Types of Variables:

- **Quantitative variable** – takes numerical values for which arithmetic operations such as adding and averaging make sense. {Height, Blood Pressure, Survival Time, etc.}
- **Categorical (Qualitative) variable** – places an individual into one of several groups or categories. {Gender, Blood Type, Type of Treatment, etc.}
  - **Dichotomous/Binary** – Has only two possible responses.
  - **Polytomous** – Has more than two possible responses.
**Dichotomous Response**

**Bernoulli Trial** – A trial that has only 2 possible outcomes – a success (S) or a failure (F).

**Bernoulli Distribution** – Let $X=1$ when a success is observed from a Bernoulli trial and $X=0$ when a failure is observed. If the probability of getting a success is $\pi$, then the probability distribution of $X$ is

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pr(X = x)$</td>
<td>$\pi$</td>
<td>$1 - \pi$</td>
</tr>
</tbody>
</table>

**Binomial Distribution** – Suppose a Bernoulli trial is repeated $n$ times. The random variable $Y$, that counts the number of successes out of $n$, follows the Binomial probability distribution.

$$P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad \text{for } y = 0, 1, \ldots, n$$

**Binomial Example**

Suppose someone claims that more than 60% of the student population of UWL is in favor of making the UWL campus a non-smoking zone. To test this claim, we randomly interviewed 10 UWL students and count the number ($Y$) of students (out of 10) who are in favor of this policy.

Note: $Y \sim \text{Binomial}(n = 10, \pi = .60)$

$$\Rightarrow P(Y = y) = \binom{10}{y} (.6)^y (1 - .6)^{10-y}, \quad \text{for } y = 0, 1, \ldots, 10$$

When $y = 2$, $P(Y = 2) = \binom{10}{2} (.6)^2 (1 - .6)^{10-2} = 45(.6)^2(.4)^8 = 0.0106$

```r
> dbinom(2,size=10,prob=.60)
[1] 0.01061683
> barplot(dbinom(0:10,size=10,prob=.60))
```

To compute $P(Y \leq 2)$,

```r
> pbinom(2,size=10,prob=.60)
[1] 0.01229455
```
Testing Hypothesis

In the previous example, we wanted to test the claim that the proportion of UWL students who are in favor of making the campus a non-smoking zone is at least 60%. That is, we want to test $H_0: \pi = 0.60$ versus $H_1: \pi < 0.60$.

Suppose, out of the 10 students interviewed, only 2 are in favor of this policy. Then the $p$-value $= P(Y \leq 2) = 0.0123$. Therefore, using a 0.05 level of significance, we can conclude that the true proportion must be less than 60%.

Properties of the Binomial Distribution with parameters $n$ and $\pi$:
1. Its mean is $n(\pi)$.
   For example, if $Y \sim \text{Bin}(n=10, \pi=.60)$, then $\mu_Y = 10(0.6) = 6$.
2. Its variance is $n(\pi)(1-\pi)$.
   If $Y \sim \text{Bin}(n=10, \pi=.60)$, then $\sigma^2_Y = 10(0.6)(0.4) = 2.4$.
3. It is symmetric when $\pi = 0.5$.
4. It is skewed to the right when $\pi < 0.5$.
5. It is skewed to the left when $\pi > 0.5$.
6. But as $n$ increases, its distribution becomes more symmetric and approximately normal.

Estimation of $\pi$

The best unbiased estimator of $\pi$ is $\hat{\pi} = \frac{y}{n}$, the sample proportion. When $n$ is sufficiently large ($n\pi \geq 5$ and $n(1-\pi) \geq 5$), $\hat{\pi}$ is approximately normal with mean $\pi$ and variance $\left(\frac{\pi(1-\pi)}{n}\right)$. That is, $\hat{\pi} \sim N \left(\mu_\hat{\pi} = \pi, \sigma_\hat{\pi} = \sqrt{\frac{\pi(1-\pi)}{n}}\right)$.

$Z = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$

The $(1-\alpha)100\%$ confidence interval for $\hat{\pi}$ is given below:

$\hat{\pi} \pm \frac{z_{\alpha/2}\hat{\sigma}_\pi}{\sqrt{n}} \quad \text{or} \quad \left(\hat{\pi} - \frac{z_{\alpha/2}\hat{\sigma}_\pi}{\sqrt{n}}, \hat{\pi} + \frac{z_{\alpha/2}\hat{\sigma}_\pi}{\sqrt{n}}\right)$

where

$\hat{\pi} = \frac{y}{n}$ \quad and \quad $\hat{\sigma}_\pi = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

This is known as the Wald confidence interval.
Example

EXAMPLE 10.1 (p. 501)
Researchers in the development of new treatments for cancer patients often evaluate the effectiveness of new therapies by reporting the proportion of patients who survive for a specified period of time after completion of the treatment. A new genetic treatment of 870 patients with a particular type of cancer resulted in 330 patients surviving at least 5 years after treatment. Estimate the proportion of all patients with the specified type of cancer who would survive at least 5 years after being administered this treatment. Use a 90% confidence interval.

Solution
For these data,
\[
\hat{p} = \frac{330}{870} = 0.38
\]
\[
\hat{\sigma}_p = \sqrt{\frac{(0.38)(0.62)}{870}} = 0.016
\]

\[
\text{lwr} = 0.38 - \text{qnorm}(0.95) \times 0.016 \quad \# 0.3536823
\]
\[
\text{upr} = 0.38 + \text{qnorm}(0.95) \times 0.016 \quad \# 0.4063177
\]
\[
\text{prop.test}(330,870,\text{conf.level}=0.90)
\]
1-sample proportions test with continuity correction
90 percent confidence interval:
0.3521005 0.4072844

Other Adjustments:
1. When \( n \) is not large enough, we could use the Wilson-Agresti-Coull (WAC) adjustments.

\[
\hat{\pi}_{\text{adj}} = \frac{y + \frac{3}{8}}{n + \frac{3}{4}} \quad \text{when } y = 0,
\]
\[
\hat{\pi}_{\text{adj}} = \frac{y + \frac{1}{2}}{n + 1} \quad \text{when } y = n
\]

When \( y = 0 \), the confidence interval is \((0, 1 - (\alpha/2)^{1/2})\).
When \( y = n \), the confidence interval is \((\alpha/2)^{1/2}, 1)\).

For example, when \( y = 50 \) and \( n = 50 \),
\[
y = y + 5z_{0.975}^2 = 50 + (1.645)^2 = 50 + 2.706\quad \# 52.706
\]
\[
\hat{\pi}_{\text{adj}} = \frac{y + \frac{3}{8}}{n + \frac{3}{4}} \quad \# 0.9926
\]
\[
\text{lwr} = \frac{(0.05/2)^{1/50}}{\hat{\pi}_{\text{adj}}} \quad \# 0.93
\]

Hence, the 95% CI for \( \pi \) is (0.93, 1).

For example, when \( y = 43 \) and \( n = 50 \),
\[
y = y + 5z_{0.975}^2 = 43 + (1.645)^2 = 44.9208\quad \# 44.9208
\]
\[
\hat{\pi}_{\text{adj}} = \frac{y + \frac{3}{8}}{n + \frac{3}{4}} \quad \# 0.9926
\]
\[
\text{lwr} = \frac{(0.05/2)^{1/50}}{\hat{\pi}_{\text{adj}}} \quad \# 0.93
\]

Hence, the 95% CI for \( \pi \) is (0.93, 1).
Large-Sample Test for $\pi$

$H_0 : \pi = 0.7$ versus $H_1 : \pi \neq 0.7$

> prop.test(43,50,p=.7)

1-sample proportions test with continuity correction
data: 43 out of 50, null probability 0.7
X-squared = 5.3571, df = 1, p-value = 0.02064
alternative hypothesis: true p is not equal to 0.7
95 percent confidence interval:
0.7264362 0.9372299
sample estimates:
p 0.86

Therefore, we are 95% confident that the true proportion ($\pi$) is between (0.7264, 0.9372).

$H_0 : \pi = 0.7$ versus $H_1 : \pi > 0.7$

> prop.test(330,870,p=.7,alternative="greater")

1-sample proportions test with continuity correction
data: 43 out of 50, null probability 0.7
X-squared = 5.3571, df = 1, p-value = 0.01032
alternative hypothesis: true p is greater than 0.7
95 percent confidence interval:
0.7491613 1.0000000

Using $\alpha=0.05$, Reject $H_0$.

Sample Size Calculations

The $(1-\alpha)100\%$ CI for $\pi$ is $\hat{\pi} \pm ME$

$$ME = Z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}} \Rightarrow ME^2 = \frac{Z_{\alpha/2}^2[\pi(1-\pi)]}{n} \Rightarrow n = \frac{Z_{\alpha/2}^2[\pi(1-\pi)]}{ME^2} \leq \frac{Z_{\alpha/2}^2(0.25)}{ME^2}$$

EXAMPLE 10.4

In Example 10.3, the designer of the new operating system has decided to conduct a more extensive study. She wants to determine how many programs to randomly sample in order to estimate the proportion of Microsoft Windows–compatible programs that would perform adequately using the new operating system. The designer wants the estimator to be within .03 of the true proportion using a 95% confidence interval as the estimator.

$$n = \frac{(1.96)^2[0.5(1-0.5)]}{(0.03)^2} = 1067.1 \quad \text{Rounding up, choose } n=1068.$$

If we are fairly sure that $\pi \geq 0.8$ \Rightarrow $n = \frac{1.96^2[0.8(1-0.8)]}{(0.03)^2} = 682.95 \quad \text{Choose } n=683.$
Two Population Proportion, $\pi_1 - \pi_2$

The best unbiased estimator of $(\pi_1 - \pi_2)$ is $(\hat{\pi}_1 - \hat{\pi}_2)$.

$$\mu_{\hat{\pi}_1 - \hat{\pi}_2} = \pi_1 - \pi_2$$

$$\sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

When $n_1$ and $n_2$ are sufficiently large ($n_1\pi_1 \geq 5$ and $n_1(1-\pi_1) \geq 5$),

$$Z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \approx N(0, 1)$$

The $(1-\alpha)100\%$ CI for $(\pi_1 - \pi_2)$ is

$$\hat{\pi}_1 - \hat{\pi}_2 \pm \frac{z_{\alpha/2}}{\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}}$$

Where,

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Example 10.6

**Example 10.6** (p. 508)

A company tests a market in the Grand Rapids, Michigan, and Wichita, Kansas, metropolitan areas. The company’s advertising in the Grand Rapids area is based almost entirely on television commercials. In Wichita, the company spends a roughly equal dollar amount on a balanced mix of television, radio, newspaper, and magazine ads. Two months after the ad campaign begins, the company conducts surveys to determine consumer awareness of the product.

<table>
<thead>
<tr>
<th>Grand Rapids</th>
<th>Wichita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number interviewed</td>
<td>608</td>
</tr>
<tr>
<td>Number aware</td>
<td>392</td>
</tr>
</tbody>
</table>

Calculate a 95% confidence interval for the regional difference in the proportion of all consumers who are aware of the product (as shown in Table 10.1).

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{(.784)(.216)}{527} + \frac{(.645)(.355)}{608}} = .0264$$

Therefore, the 95% confidence interval is

$$(.784 - .645) - 1.96(.0264) \leq \pi_1 - \pi_2 \leq (.784 - .645) + 1.96(.0264)$$

$$\approx .097 \leq \pi_1 - \pi_2 \leq .191$$
Using R

```r
> prop.test(c(413,392),c(527,608))

2-sample test for equality of proportions with continuity correction
data:  c(413, 392) out of c(527, 608)
X-squared = 25.7596, df = 1, p-value = 3.867e-07
alternative hypothesis: two.sided
95 percent confidence interval:
 0.08537646 0.19251229
sample estimates:
prop 1    prop 2
0.7836812 0.6447368
```

Testing Equality of Two Population Proportions

\[ H_0 : \pi_1 = \pi_2 \text{ versus } H_1 : \pi_1 \neq \pi_2 \]

Since the p-value is 3.87e-07, we reject \( H_0 \). Therefore, we have sufficient evidence that \( \pi_1 \) is not equal to \( \pi_2 \).

---

Fisher Exact Test

When at least one of the conditions, \( n_1\hat{\pi}_1 \geq 5, n_1(1 - \hat{\pi}_1) \geq 5, n_2\hat{\pi}_2 \geq 5, \text{ or } n_2(1 - \pi_2) \geq 5 \), for using the large sample approximation to the distribution of the test statistic for comparing two proportions is invalid, the Fisher Exact Test should be used.

**EXAMPLE 10.8**

A clinical trial is conducted to compare two drug therapies for leukemia: P and PV. Twenty-one patients were assigned to drug P and forty-two patients to drug PV. Table 10.4 summarizes the success of the two drugs:

<table>
<thead>
<tr>
<th>Drug</th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>58</td>
<td>42</td>
<td>100</td>
</tr>
<tr>
<td>PV</td>
<td>11</td>
<td>21</td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Drug</th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>58</td>
<td>15</td>
<td>42</td>
<td>100</td>
</tr>
<tr>
<td>PV</td>
<td>11</td>
<td>7</td>
<td>21</td>
<td>32</td>
</tr>
</tbody>
</table>

Is there significant evidence that the proportion of patients obtaining a successful outcome is higher for drug PV than for drug P?

\[ H_0 : \pi_P \geq \pi_{PV} \text{ versus } H_a : \pi_P < \pi_{PV} \]
Fisher Exact Test using R

```r
> data,eg10.8=matrix(c(38,14,4,7),ncol=2)
> data,eg10.8
   [,1] [,2]
[1,]  38  4
[2,]  14  7
> dimnames(data,eg10.8)=list(c("PV","V"),c("Success","Failure"))
> data,eg10.8
      Success Failure
   PV       38       4
   P        14       7
> fisher.test(data,eg10.8,alternative="greater")
Fisher's Exact Test for Count Data
data:  data,eg10.8
p-value = 0.02537
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
  1.22629    Inf
sample estimates:
   odds ratio
        4.615064
```

Since the p-value (0.02537) is less than 0.05, we reject $H_0$ and conclude $\pi_{PV} > \pi_{P}$. That is, the proportion of successful outcome using drug PV is higher than that of drug P.

Using Raw Data

```r
> data=read.csv("data_example10_8.csv",header=T)
> head(data)
patients drug outcome
1       1    v       s
2       2 pv       s
3       3 pv       s
4       4    v       f
5       5 pv       s
6       6 pv       s
> attach(data)
> table(drug)
drug
  p pv
21 42
> table(drug,outcome)
   outcome
   drug f s
   p  7 14
   pv 4 38
> fisher.test(drug,outcome,alternative="greater")
Fisher's Exact Test for Count Data
data:  drug and outcome
p-value = 0.02537
```

Note that drug P is now in the first row and the first column is now for “failure”. Hence, using alternative="greater" would mean that the proportion of failed outcome using drug P is higher than that of drug PV. That is, $(1-\pi_{PV}) > (1-\pi_{P})$, which is equivalent to $\pi_{PV} > \pi_{P}$.
Polytomous Response

Multinomial Experiment

1. The experiment consists of \( n \) identical trials.
2. Each trial results in one of \( k \) outcomes.
3. The probability that a single trial will result in outcome \( i \) is \( \pi_i \) for \( i = 1, 2, \ldots, k \), and remains constant from trial to trial. (Note: \( \sum \pi_i = 1 \).)
4. The trials are independent.
5. We are interested in \( n_i \), the number of trials resulting in outcome \( i \). (Note: \( \sum n_i = n \)).

The probability distribution for the number of observations resulting in each of the \( k \) outcomes, called the multinomial distribution, is given by the formula

\[
P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_k^{n_k}
\]

\( n! = n(n-1)(n-2) \cdots (2)(1) \), \( \pi_1 + \pi_2 + \cdots + \pi_k = 1 \), and \( n_1 + n_2 + \cdots + n_k = n \).

For example, when \( k = 4 \), we may want to test the following hypotheses:

\( H_0: \pi_1 = .50, \pi_2 = .25, \pi_3 = .10, \pi_4 = .15 \) versus \( H_1: \text{At least one of the probabilities is different from the hypothesized value.} \)

Multinomial Example

**EXAMPLE 10.10** (p. 516)

A laboratory is comparing a test drug to a standard drug preparation that is useful in the maintenance of patients suffering from high blood pressure. Over many clinical trials at many different locations, the standard therapy was administered to patients with comparable hypertension (as measured by the New York Heart Association (NYHA) Classification). The lab then classified the responses to therapy for this large patient group into one of four response categories. Table 10.6 lists the categories and percentages of patients treated on the standard preparation who have been classified in each category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marked decrease in blood pressure</td>
<td>50</td>
</tr>
<tr>
<td>Moderate decrease in blood pressure</td>
<td>25</td>
</tr>
<tr>
<td>Slight decrease in blood pressure</td>
<td>10</td>
</tr>
<tr>
<td>Stationary or slight increase in blood pressure</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Observed Cell Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>120</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

The lab then conducted a clinical trial with a random sample of 200 patients with high blood pressure. All patients were required to be listed according to the same hypertensive categories of the NYHA Classification as those studied under the standard preparation. Use the sample data in Table 10.7 to test the hypothesis that the cell probabilities associated with the test preparation are identical to those for the standard. Use \( \alpha = .05 \).
Chi-Square Goodness-of-Fit Test

\(H_0: \ \pi_1 = .50, \pi_2 = .25, \pi_3 = .10, \pi_4 = .15\) versus
\(H_1: \) At least one of the probabilities is different from the hypothesized value.

### Goodness-of-fit Test

\[
X^2 = \sum_{i=1}^{k} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}
\]

This \(X^2\) statistic follows **approximately** the \(\chi^2\) distribution with \(k - 1\) degrees of freedom.

\[
X_{\text{obs}}^2 = \frac{(120 - 200 \times 0.50)^2}{200 \times 0.50} + \frac{(60 - 200 \times 0.25)^2}{200 \times 0.25} + \frac{(10 - 200 \times 0.10)^2}{200 \times 0.10} + \frac{(10 - 200 \times 0.15)^2}{200 \times 0.15}
\]

\[
= \frac{(120 - 100)^2}{100} + \frac{(60 - 50)^2}{50} + \frac{(10 - 20)^2}{20} + \frac{(10 - 30)^2}{30} = 24.3333
\]

\[> \ p\text{.value}=1-p\text{chisq}(24.3333,\text{df}=3) \] # p.value = 2.128091e-05
\[> \ \text{chisq.test}(c(120,60,10,10),p=c(.5,.25,.10,.15)) \]
Chi-squared test for given probabilities
data: c(120, 60, 10, 10)
X-squared = 24.3333, df = 3, p-value = 2.128e-05

Note: This test is always one-sided right-tailed test.

Chi-Square GoF Test for Normality

Checking Normality. The table below shows the number of values in a sample of size \(n = 200\) observations falling in each category. At the 0.05 level of significance, test whether the sample can be reasonably assumed to have come from the standard normal distribution.

<table>
<thead>
<tr>
<th>z-values</th>
<th>z &lt; -1</th>
<th>-1 ≤ z &lt; -0.5</th>
<th>-0.5 ≤ z &lt; 0</th>
<th>0 ≤ z &lt; 0.5</th>
<th>0.5 ≤ z &lt; 1</th>
<th>z &gt; 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Obs.</td>
<td>30</td>
<td>38</td>
<td>43</td>
<td>38</td>
<td>20</td>
<td>31</td>
</tr>
</tbody>
</table>

Note:
- \(\pi_1 = P(Z < -1) \approx 0.16\)
- \(\pi_2 = P(-1 < Z < -0.5) \approx 0.15\)
- \(\pi_3 = P(-0.5 < Z < 0) \approx 0.19\)
- \(\pi_4 = P(0 < Z < 0.5) \approx 0.19\)
- \(\pi_5 = P(0.5 < Z < 1) \approx 0.15\)
- \(\pi_6 = P(Z > 1) \approx 0.16\)

\(H_0: \) The data come from the standard normal distribution versus
\(H_1: \) At least one of the probabilities is different from the hypothesized value.

\[> \ \text{chisq.test}(c(30,38,43,38,20,31),p=c(.16,.15,.19,.19,.15,.16)) \]
Chi-squared test for given probabilities
data: c(30, 38, 43, 38, 20, 31)
X-squared = 6.2808, df = 5, p-value = 0.2798

Based on the p-value, it is reasonable to assume that the data come from a standard normal distribution.
Poisson Distribution

- Developed by a French Mathematician (Simeon Denis Poisson) in 1837.
- This distribution is useful in modeling occurrence frequency of events over a unit of time or space. For example,
  1. The number of customers arriving at a check out counter, getting in a bank, or cars arriving at a parking lot, toll booth, inspection station, garage repair shop, etc.
  2. The number of deer in an acre.
  3. The number of clumps of algae of a particular species observed in a unit of volume of lake water.

Poisson Distribution – Let \( Y \) be the number of events occurring during a fixed time interval or fixed region of area or volume. Then \( Y \) follows the Poisson probability distribution, provided certain conditions are satisfied:

\[
P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad \text{for } y = 0, 1, 2, \ldots
\]

Properties:
1. \( E(Y) = \mu = \lambda t \)
2. \( V(Y) = E(Y) = \mu \)
3. If \( Y \sim \text{Binomial}(n, \pi) \), when \( \pi \) is small (<0.01) and \( n \) is large (>100), then \( Y \sim \text{Pois}(\mu=n\pi) \)

Chi-Square GoF Test for Poisson Dist.

**Example 10.11 (p. 519)**

Environmental engineers often utilize information contained in the number of different algal species and the number of cell clumps per species to measure the health of a lake. Those lakes exhibiting only a few species but many cell clumps are classified as oligotrophic. In one such investigation, a lake sample was analyzed under a microscope to determine the number of clumps of cells per microscope field. These data are summarized here for 150 fields examined under a microscope. Here \( y_i \) denotes the number of cell clumps per field and \( n_i \) denotes the number of fields with \( y_i \) cell clumps.

\[
\begin{array}{c|cccccccccc}
  y_i & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
n_i & 6 & 23 & 29 & 31 & 27 & 13 & 8 & 6 & 5 & 2 \\
\end{array}
\]

Use \( \alpha = .05 \) to test the null hypothesis that the sample data were drawn from a Poisson probability distribution.

\( H_0 \): The data come from a Poisson distribution with \( \mu=3.5 \)

To do this, we need to compute the Poisson probabilities in order to get the expected counts. For example, \( \pi_2 = P(Y=2) = (3.5^2)(e^{-3.5})/2! \approx 0.18496 \)

\[
\begin{align*}
& \texttt{dpois(2,lambda=3.5)} \quad \# 0.184959 \\
& \texttt{ppois(2,lambda=3.5)} \quad \# 0.320847
\end{align*}
\]

\( P(Y \leq 2) \)
Chi-Square GoF Test for Poisson Dist.

**EXAMPLE 10.11**

<table>
<thead>
<tr>
<th>(y_i)</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>(n_i)</td>
<td>6</td>
<td>23</td>
<td>29</td>
<td>31</td>
<td>27</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
y = 0:9
\]
\[
\text{count} = c(6,23,29,31,27,13,8,6,5,2)
\]
\[
\text{barplot(count)}
\]

Since \(n=150\), then the expected counts are
\[
150 * \text{dpois}(0:9, \text{lambda}=3.5)
\]
\[
[1] 4.53 15.85 27.74 32.37 28.32 19.83 11.56 5.78 2.53 0.98
\]

Note that the last 2 expected counts are less than 5. In cases like this, we could collapse some of the cells to achieve the desired expected count of at least 5.

\[
y2 = 0:7
\]
\[
\text{count2} = c(6,23,29,31,27,13,8,13)
\]
\[
\text{pi} = \text{dpois}(0:6, \text{lambda}=3.5)
\]
\[
\text{pi8} = \text{1-ppois}(6, \text{lambda}=3.5)
\]
\[
\text{pies} = c(\text{pi}, \text{pi8})
\]
\[
\text{chisq.test(count2, p=pies)}
\]

Chi-squared test for given probabilities
X-squared = 8.3737, df = 7, p-value = 0.3008

\[
\text{p.value} = 1-\text{pchisq}(8.3737, \text{df}=8-1)
\]
# 0.3007971

Since the p-value is 0.30, then it is reasonable to assume that the data come from a Poisson distribution with mean 3.5.

**Guidelines for Assessing Quality of Model Fit**
- \(p\)-value \(\geq 0.25\) => Excellent fit
- \(0.15 \leq p\)-value < 0.25 => Good fit
- \(0.05 \leq p\)-value < 0.15 => Moderately good fit
- \(0.01 \leq p\)-value < 0.05 => Poor fit
- \(p\)-value < 0.01 => Unacceptable fit
example 10.11  *(when μ is not specified.)*

\[ y_i \begin{array}{c|c|c|c|c|c|c|c|c|c} \hline
0 & 1 & 2 & 3 & 4 & 5 & 7 & 8 & 9 \\
\hline
n_i & 6 & 23 & 29 & 31 & 27 & 13 & 8 & 6 & 5 & 2 \\
\hline
\end{array} \]

\[ > \text{y}=0:7; \text{count}=c(6,23,29,31,27,13,8,6,5,2) \]
\[ > \text{mean.estimate} = \text{sum}(\text{y} * \text{count}) / \text{sum}(\text{count}) \]  # 3.3

\[ \begin{array}{c|c|c|c|c|c|c|c|c|c} \hline
y_i & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \geq 8 \\
\hline
n_i & 6 & 23 & 29 & 31 & 27 & 13 & 8 & 6 & 5 & 2 \\
\hline
\end{array} \]

\[ \text{chisq.test(count2,p=pies)} \]

\[ \chi^2 \text{-squared test for given probabilities} \]
\[ X^2 = 6.9891, \text{df} = 7, p\text{-value} = 0.43 \]

\[ > \text{p.value} = 1-pchisq(6.9891,\text{df}=\text{8-2}) \]

\[ \text{Note that the p-value you get from the chisq.test procedure is wrong due to incorrect df. When μ is unknown and has to be estimated using the data, we lose 1 df.} \]

Testing the Equality of Several Proportions

**Quitting Smoking.** The accompanying table summarizes successes and failures when subjects used different methods in trying to stop smoking. The determination of smoking or not smoking was made five months after the treatment was begun, and the data are based on results from the Centers for Disease Control and Prevention. Use a 0.05 level of significance to test the claim that success is independent of the method used. If someone wants to stop smoking, does the choice of the method make a difference?

\begin{center}
\begin{tabular}{|c|c|c|c|c|}
\hline
& Nicotine Gum & Nicotine Patch & Nicotine Inhaler & Total \\
\hline
Smoking & 191 & 260 & 98 & 549 \\
No smoking & 59 & 57 & 27 & 143 \\
Total & 250 & 320 & 122 & 692 \\
\hline
\end{tabular}
\end{center}

\[ 1 - \hat{π} = 549/692 \]
\[ \hat{π} = 143/692 \approx 20.7\% \]

\[ \begin{align*}
E_{21} &= (143/692) \\
E_{12} &= 320(549/692) \\
\end{align*} \]

\[ \begin{align*}
H_0: & \text{ There is no difference between the three methods.} \\
H_0: & π_1 = π_2 = π_3 \text{ (there is a common π)} \\
H_1: & \text{ There is a difference between the three methods.} \\
H_1: & \text{ At least one } π_i \text{ is different.} \\
\end{align*} \]

\[ \begin{align*}
X^2 &= \sum_{\text{all cells}} \frac{(\text{observed count - expected count})^2}{\text{expected count}} \\
& \sim X^2_{(r-1)(c-1)} \\
\end{align*} \]

\[ \begin{align*}
\Rightarrow \quad X^2_{\text{obs}} &= \frac{(191 - 198.3)^2}{198.3} + \frac{(263 - 253.9)^2}{253.9} + \cdots + \frac{(27 - 25.2)^2}{25.2} \approx 3.06 \\
\end{align*} \]

\[ > \text{p.value} = 1-pchisq(3.06,\text{df}=(2-1)\times(3-1)) \]  # 0.2165357
Testing Several Proportions using R

Quitting Smoking. The accompanying table summarizes successes and failures when subjects used different methods in trying to stop smoking. The determination of smoking or not smoking was made five months after the treatment was begun, and the data are based on results from the Centers for Disease Control and Prevention. Use a 0.05 level of significance to test the claim that success is independent of the method used. If someone wants to stop smoking, does the choice of the method make a difference?

<table>
<thead>
<tr>
<th></th>
<th>Nicotine Gum</th>
<th>Nicotine Patch</th>
<th>Nicotine Inhaler</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>191 (198.3)</td>
<td>263 (253.9)</td>
<td>95 (96.8)</td>
<td>549</td>
</tr>
<tr>
<td>No smoking</td>
<td>59 (51.7)</td>
<td>57 (66.1)</td>
<td>27 (25.2)</td>
<td>143</td>
</tr>
<tr>
<td>Total</td>
<td>250</td>
<td>320</td>
<td>122</td>
<td>692</td>
</tr>
</tbody>
</table>

$H_0$: There is no difference between the three methods.

$H_1$: There is a difference between the three methods.

```r
> data.smoking=matrix(c(191,59,263,57,95,27),ncol=3)
> dimnames(data.smoking)=list(c("S","NS"),c("G","P","I"))
> result.smoking=chisq.test(data.smoking)
> result.smoking
Pearson's Chi-squared test
X-squared = 3.0618, df = 2, p-value = 0.2163
> attributes(result.smoking)

result.smoking$expected

<table>
<thead>
<tr>
<th></th>
<th>Gum</th>
<th>Patch</th>
<th>Inhaler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>198.33815</td>
<td>253.87283</td>
<td>96.78902</td>
</tr>
<tr>
<td>Not Smoking</td>
<td>51.66185</td>
<td>66.12717</td>
<td>25.21098</td>
</tr>
</tbody>
</table>
```

Chi-Square Test for Independence

**Example 10.12** (p. 522)

Suppose a random sample of 216 patients having the skin disease are classified into the four age categories yielding the frequencies shown in Table 10.10.

<table>
<thead>
<tr>
<th>Age Category</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>All Ages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moderate</td>
<td>15</td>
<td>32</td>
<td>18</td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>Mildly Severe</td>
<td>8</td>
<td>29</td>
<td>23</td>
<td>18</td>
<td>78</td>
</tr>
<tr>
<td>Severe</td>
<td>1</td>
<td>20</td>
<td>25</td>
<td>22</td>
<td>68</td>
</tr>
<tr>
<td>All Severities</td>
<td>24</td>
<td>81</td>
<td>66</td>
<td>45</td>
<td>216</td>
</tr>
</tbody>
</table>

Determine if severity of the disease is independent of the age of the patient.

$H_0$: Severity of the disease and age of the patient are independent.

$H_1$: Severity of the disease and age of the patient are not independent (dependent).

```r
> data.eg10.12=matrix(c(15,8,1,32,29,20,18,23,25,5,18,22),ncol=4)
> result.chisq=chisq.test(data.eg10.12)
> result.chisq

Pearson's Chi-squared test
X-squared = 27.135, df = 6, p-value = 0.0001366
```

Since the p-value is < 0.05, we found enough evidence that the age of patients has an effect on the severity of the disease.
Odds of an Event

Let $P(A)$ be the probability that an event $A$ will occur.

For example, if $P(A) = 2/3$, then odds$(A) = (2/3)/(1 − 2/3) = 2$. So the odds of event $A$ happening is 2 to 1. When $P(A)=0.5$, then odds$(A)=1$.

**EXAMPLE 10.15** (p. 531)

Consider both a population in which 1 of every 1,000 people carried the HIV virus and a test that yielded positive results for 95% of those who carry the virus and (false) positive results for 2% of those who do not carry it. If a randomly chosen person obtains a positive test result, should the odds of that person carrying the HIV virus go up or go down? By how much?

**Solution:**
- The initial odds of a person carrying the HIV virus is odds$(H)=(.001)/(.999)$.
- $P(\text{positive result})=P(+) = (.001*.95)+(.999*.02)=0.02093$.
- $P(H|+) = (.001*.95)/0.02093 = 0.04539$.
- Odds$(H|+) = P(H|+)/ (1 – P(H|+)) = 0.04539 / (1 − 0.04539) = 0.0475$

Odds Ratio

**Odds Ratio of an Event for Two Groups**

If $A$ is any event with probabilities $P(A|\text{group 1})$ and $P(A|\text{group 2})$, the odds ratio (OR) is

$$OR = \frac{P(A|\text{group 1})/(1 − P(A|\text{group 1}))}{P(A|\text{group 2})/(1 − P(A|\text{group 2}))}$$

The odds ratio equals 1 if the event $A$ is statistically independent of group.

<table>
<thead>
<tr>
<th>Condition $A$</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
<th>Proportion Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_1$</td>
<td>$p_1 = n_{11}/n_1$</td>
</tr>
<tr>
<td>Group 2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>$n_2$</td>
<td>$p_2 = n_{21}/n_2$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n$</td>
<td>$p = n_{11}/n$</td>
</tr>
</tbody>
</table>

Given the sample data above, we can estimate OR by

$$OR = \frac{p_1/(1 − p_1)}{p_2/(1 − p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}p_{22}}{n_{21}p_{12}}$$
Sampling Distribution of \( \log(OR) \)

For large sample sizes the sampling distribution of the log odds ratio, \( \ln(OR) \), is approximately normal with

\[
\mu_{\ln(OR)} = \ln\left(\frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}\right)
\]

where \( \pi_1 \) and \( \pi_2 \) are the population proportions for the two groups, and

\[
\sigma_{\ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}
\]

From the above results we obtain an approximate 100(1 - \( \alpha \)) confidence interval for the population log odds ratio, \( \ln\left(\frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}\right) \):

\[
\left(\ln(OR) - z_{\alpha/2}\sigma_{\ln(OR)}\right), \ln(OR) + z_{\alpha/2}\sigma_{\ln(OR)}
\]

The (1-\( \alpha \))100% confidence interval for OR can be obtained from the above interval by taking the exponent of the limits. If this interval does not include an odds ratio of 1, we conclude with (1-\( \alpha \))100% confidence level condition A is related to the groups.

Odds Ratio Example

**EXAMPLE 10.16** (p. 533)

A study was conducted to determine if the level of stress in a person's job affects his or her opinion about the company's proposed new health plan. A random sample of 3,000 employees yields the responses shown in Table 10.19.

<table>
<thead>
<tr>
<th>Job Stress</th>
<th>Employee Response</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Favorable</td>
<td>Unfavorable</td>
</tr>
<tr>
<td>Low</td>
<td>250</td>
<td>750</td>
</tr>
<tr>
<td>High</td>
<td>400</td>
<td>1,600</td>
</tr>
<tr>
<td>Total</td>
<td>650</td>
<td>2,350</td>
</tr>
</tbody>
</table>

Estimate the conditional probabilities of a favorable and unfavorable response given the level of stress. Compute an estimate of the odds ratio of a favorable response for the two groups and determine if type of response is related to level of stress.
Odds Ratio Example

Solution The estimated conditional probabilities are given in Table 10.20.

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Job Stress} & \text{Favorable} & \text{Unfavorable} & \text{Total} \\
\hline
\text{Low} & .25 & .75 & 1.0 \\
\text{High} & .20 & .80 & 1.0 \\
\hline
\end{array}
\]

Hence, an estimate of \( OR = \frac{.25/.75}{.20/.80} = \frac{4}{3} = 1.3333 \). Or

\[
OR = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}
\]

\[
\ln(OR) = \ln(1.333) = 0.2874
\]

\[
\hat{\sigma}_{\ln(OR)} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{250} + \frac{1}{750} + \frac{1}{400} + \frac{1}{1,800}} = 0.0084583 = .0920
\]

The 95% confidence interval for \( \ln(OR) \) is

\( (.2874 - 1.96(0.0920), .2874 + 1.96(0.0920)) = (0.1071, 0.4677) \)

The 95% confidence interval for \( OR \) is \( (e^{0.1071}, e^{0.4677}) = (1.113, 1.5963) \)

Combining Sets of 2x2 Tables

**EXAMPLE 10.17** (p. 536)

The pharmaceutical study discussed previously was extended to three clinics. In each clinic, as patients qualified for the study and gave their consent to participate, they were assigned to either the drug or placebo groups according to a predetermined random code. Each clinic was to treat 50 patients per group. The study results are summarized in Table 10.25. Use these data to test the null hypothesis of no difference in the improvement rates, on the average.

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Improved</th>
<th>Not Improved</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40 (80%)</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>15 (30%)</td>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>45</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>35 (70%)</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>20 (40%)</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>55</td>
<td>45</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>43 (86%)</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>31 (62%)</td>
<td>19</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>74</td>
<td>26</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>184</td>
<td>116</td>
<td>300</td>
</tr>
</tbody>
</table>
Cochran-Mantel-Haenszel (CMH)

Let OR be the ratio of the odds of response 1 in treatment 1 to the odds of response 1 in treatment 2.

\( H_0: \text{OR}=1 \) vs. \( H_1: \text{OR}>1 \)

\[
\chi^2_{\text{MH}} = \left( \sum_n \frac{(n_{h11} - n_{h1}n_{h,1})}{n_{h,1}} \right)^2 \sum_n \frac{1}{n_{h,1}n_{h,2}n_{h,3}n_{h,4}} \left( \frac{n_{h,1}n_{h,2}n_{h,3}n_{h,4}}{n_{h,1}n_{h,2}n_{h,3}n_{h,4}} \right)^2 \\
\sim \chi^2_{\text{df}=1}
\]

Example 10.17

Solution: The necessary row and column totals in each clinic are given in Table 10.25. The numerator of the CMH statistic is

\[
\left( \sum_n \left( n_{h11} - n_{h1}n_{h,1} \right) \right)^2 = \left( 40 - \frac{50(55)}{100} \right)^2 + \left( 35 - \frac{50(55)}{100} \right)^2 + \left( 43 - \frac{50(74)}{100} \right)^2 \\
= (12.5 + 7.5 + 6)^2 = 676,
\]

whereas the denominator is

\[
\sum_n \frac{n_{h1}n_{h2}n_{h3}n_{h,4}}{n_{h,1}(n_{h,1} - 1)} = \frac{50(50)(55)(45)}{100} + \frac{50(50)(55)(45)}{100} + \frac{50(50)(74)(26)}{100} = 6.25 + 6.25 + 4.8586 = 17.3586
\]

\[
\chi^2_{\text{MH}} = \frac{676}{17.3586} = 38.9432
\]

\[
> \text{p.value}=1-\text{pchisq}(38.94, \text{df}=1) \quad \# 4.370333e-10
\]
Mantel-Haenszel Test in R

```r
> data.eg10.17 = array(c(40,15,10,35,35,20,15,30,43,31,7,19),dim=c(2,2,3),dimnames=list(Treatment=c("Drug","Placebo"),Response=c("Improved","Not Improved"),Clinic=c(1,2,3)))
> mantelhaen.test(data.eg10.17,alternative="greater")

Mantel-Haenszel chi-squared test with continuity correction
data:  data.eg10.17
Mantel-Haenszel X-squared = 37.4598, df = 1,
p-value = 4.666e-10
alternative hypothesis: true common odds ratio is greater than 1
95 percent confidence interval:
 3.174667   Inf
sample estimates:
common odds ratio
  4.898051
```

Simpson’s Paradox

Simpson’s paradox, or the Yule–Simpson effect, is a paradox in probability and statistics, in which a trend appears in different groups of data but disappears or reverses when these groups are combined.

One of the best known real-life example of Simpson’s paradox occurred when the University of California, Berkeley was sued for bias against women who had applied for admission to graduate school.

<table>
<thead>
<tr>
<th>Department</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
</tr>
<tr>
<td>A</td>
<td>625</td>
<td>62%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
</tr>
<tr>
<td>F</td>
<td>272</td>
<td>6%</td>
</tr>
</tbody>
</table>

This 1973 data shows that men were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to random chance.