\( \chi^2 \) Applications
(pronounced “kie-square”; text Ch. 15)

“She’s the sweetheart of \( \Sigma X \)”
The XP page from the *Book of Kells*

- Tests for Independence
  - is characteristic R the same or different in different populations?
- Tests for Equality of Population Proportions,
  Three or More Populations
- Tests for “Goodness of Fit”
  - did this sample come from a population with ____ distribution?
- Confidence Intervals and Hypothesis Tests for
  Variances and Standard Deviations
$\chi^2$ Contingency Tables (Tests for Independence)

<table>
<thead>
<tr>
<th>Medication/Outcome</th>
<th>Recovered</th>
<th>Didn’t Recover</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prozan</td>
<td>150</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Drug X</td>
<td>200</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>350</strong></td>
<td><strong>150</strong></td>
<td><strong>500</strong></td>
</tr>
</tbody>
</table>

Research Question: Was the proportion who recovered on Prozan the same as the proportion that recovered on Drug X?

Some difference in proportions is NOT sufficient evidence – the difference could be due to sampling variability.
• So we ask: how probable is the observed difference if there is no actual difference between these populations.
• That probability is governed by the $\chi^2$ distribution
• Chi-Square Calculation:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$r$ is the number of rows; $k$ is the number of columns. The expression directs us to calculate the expected number for each row and column intersection, then calculate the required squares. The expected number is the proportion in the total sample with characteristic R multiplied by the size of the subsample.

The chi-square statistic has $(r-1) \times (k-1)$ degrees of freedom.
How many would we *expect* to recover in the Prozan group?

Of the total sample of 500, 350 recovered so $P(R) = \frac{350}{500} = 0.7$.

There are 200 in the Prozan sample, so $E_{11} = 0.7 \times 200 = 140$.

Formula: $E_{ij} = \left(\frac{C_j}{n}\right) \times R_i$ where $C_j$ is the total in the j-th column and $R_i$ is the total in the i-th row, while $n$ is the size of the total sample.
<table>
<thead>
<tr>
<th>Medication /Outcome</th>
<th>Recovered</th>
<th>Didn’t Recover</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prozan</td>
<td>O(_{11}): 150</td>
<td>O(_{12}): 50</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>E(_{11}): 140</td>
<td>E(_{12}): 60</td>
<td></td>
</tr>
<tr>
<td>Drug X</td>
<td>O(_{21}): 200</td>
<td>O(_{22}): 100</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>E(_{21}): 210</td>
<td>E(_{22}): 90</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>350</td>
<td>150</td>
<td>500</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{10^2}{140} + \frac{10^2}{210} + \frac{10^2}{60} + \frac{10^2}{90} = 3.9683
\]
• The null hypothesis is phrased as “Recovery is independent of drug taken.”
• That is, knowing the patient took Prozan doesn’t change the probability of recovery
• Or, there’s no difference in effectiveness
• To reject $H_0$ means to conclude that there is a difference in effect
• Performing the hypothesis tests:
  • $\chi^2$ table, Appendix E, p. 785
  • In these tests we are concerned only with right-tail area
  • At 5% significance with 1 degree of freedom, critical value = 3.841 < 3.9683
• Excel offers several approaches
• CHIINV(significance level, df) gives the critical value
  – CHIINV(.05, 1) = 3.8415
• CHIDIST($\chi^2$, df) gives the p-value of the test
  – CHIDIST(3.9683, 1) = 0.0463
<table>
<thead>
<tr>
<th>Major/Start Sal</th>
<th>Business</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40,000</td>
<td>120</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>≥ 40,000</td>
<td>280</td>
<td>340</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \left( \frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2$$
the *expected* number of business grads with salary < 40,000

1) 120  
2) 124.44  
3) 400  
4) none of the above
The squared value for “Other < 40000” is

1) 160
2) 6.25
3) 0.1267
4) 155.56
The $\chi^2$ value is

1) 12.35 
2) 900 
3) 0.4139 
4) none of the above
<table>
<thead>
<tr>
<th>Method/Level</th>
<th>Not Proficient</th>
<th>Proficient</th>
<th>Highly Proficient</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sight</td>
<td>64</td>
<td>86</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Phonics</td>
<td>51</td>
<td>78</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Write-to-Read</td>
<td>33</td>
<td>99</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• **Excel Procedures:**
  - Enter the data as a table in Excel
  - Create a table in Excel showing the expected frequencies, as below
  - When cell B9 is copied to B9 to D11, it will create the required expected frequencies
  - Use `chitest(B2:D4,B9:D11);` the result is the p-value of the test

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td><code>(B$5/$E$5)*$E2</code></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Using $\chi^2$ for estimating variances

Text, pp. 343 – 344

• If the population is normally distributed, then
• A C% confidence interval for $\sigma^2$ has the form

$$\frac{(n-1)s^2}{\chi^2_U} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_L}$$

where $s^2$ is the variance calculated from a sample and $n$ is the sample size. The chi-square values are chosen to cut off the upper and lower $(1 - C)/2$ of the chi-square distribution.

CHIINV(probability, df) gives the chi-square value such that the probability entered lies to the right of that value

CHIINV(0.975, df) gives the lower value for a 95% confidence interval or $\chi^2_L$

CHIINV(0.025, df) gives the upper value for a 95% confidence interval or $\chi^2_U$
• **Example:** Speeds for 38 cars were recorded on Stadium Drive in Boone. The mean speed was 28 mph with standard deviation 7 mph. Calculate 95 and 99% confidence intervals for the population variance.

  – Solution: First, find the chi-square values using Excel
  – For 95% confidence interval $\chi^2_U = \text{chidist}(0.025,37) = 55.67$, $\chi^2_L = \text{chidist}(0.975,37) = 22.11$
  – The corresponding values for 99% c.i. are 62.88 and 18.59
  – Applying our formula, we have

\[
\frac{(38-1)7^2}{55.67} \leq \sigma^2 \leq \frac{(38-1)7^2}{22.11}
\]

Or alternatively,

\[
5.71 \leq \sigma \leq 9.1
\]

\[
32.57 \leq \sigma^2 \leq 82
\]
• Find the limits for a 99% confidence interval for the variance.

• **Example:** Among a sample of 63 ASU students, mean family income was $72,123 with standard deviation $12,456. Calculate a 90% confidence interval for the population standard deviation.
The *lower* limit of the 90% confidence interval for the *standard deviation* is

A. 118,202,262
B. 14,638.76
C. 44.89
D. 10,872.09
Hypothesis tests for the population variance (standard deviation)

Text: pp. 393 – 395

The quantity \((n - 1)s^2/\sigma^2_0\) is distributed as a \(\chi^2\) statistic with \(n - 1\) degrees of freedom, so we test hypotheses about the variance by calculating:

\[
\chi^2 = \frac{(n - 1)s^2}{\sigma^2_0}
\]

This may be either a one or two-tailed test. We can find critical values using the chiinv function or p-values using the chidist function.
Example: Engineers developing a windmill cannot use a particular location if the variability of wind speed is too high. If the standard deviation exceeds 12 mph, the site is unusable. On a sample of 40 days, the standard deviation of wind speed is 14 mph. Make the determination; use a significance level of 2-1/2 %.

Solution: The null hypothesis is $H_0: \sigma^2 \leq 144$ vs. $H_1: \sigma^2 > 144$. From chiinv(0.025,39) we find that the critical value of chi-square = 58.12. The calculated value is $\chi^2 = \frac{(39 \times 196)}{144} = 53.08$.

- Alternatively, chidist(53.08,39) = 0.0657 is the p-value for an upper one-tail test

Conclusion: Fail to reject $H_0$. The evidence is insufficient that the winds are too variable, and the engineers can go ahead with this site.
• **Example:** Residents of Stadium Drive claim that speeds are highly variable in their neighborhood. A traffic consultant proposes to test the claim by testing the hypotheses $H_0: \sigma^2 \geq 81$ vs. $H_1: \sigma^2 < 81$. In a sample of 71 cars, the variance is 61. Perform the test at 5% significance.

  – Solution: $\chi^2 = (n - 1)s^2/\sigma^2_0 = (71 - 1) \times 61/81 = 51.85$. From chidist(51.85, 70) we find the probability of values $> 51.85 = 0.9488$. Since this is a lower one-tail test, $p$-value $= 1 - 0.9488 = 0.0412$.

  – Conclusion: reject $H_0$ – variance is less than 81, and the residents’ claim is exaggerated.
• **Example:** We wish to test whether bed springs are within the allowable variation, which is specified as a standard deviation of 8 lbs per square inch. A sample of 33 are carefully measured, and the resulting standard deviation is 10.5 psi. Perform the test at 5% significance level.

  – **Solution:** The null and alternative hypotheses are $H_0: \sigma^2 = 64$ vs. $H_0: \sigma^2 \neq 64$. Critical values of chi-square can be found using chiinv: they are 49.48 and 18.29, so the decision rule is: Reject $H_0$ if $\chi^2 < 18.29$ or $\chi^2 > 49.48$. In the event, we find $\chi^2 = (33 - 1) \times 10.5^2 / 64 = 55.125$.

  – **Conclusion:** Reject $H_0$; the process is out of control

  – Alternatively, chidist(55.125,32) = 0.006732. This is the area under the chi-square curve greater than 55.125. The p-value of the test is $2 \times 0.006732 = 0.0135$