21. The chi-square test for goodness of fit
Objectives (PSLS Chapter 21)

The chi-square test for goodness of fit

- Idea of the chi-square test
- The chi-square distributions
- Goodness of fit hypotheses
- Conditions for the chi-square goodness of fit test
- Chi-square test for goodness of fit
Idea of the chi-square test

The chi-square ($\chi^2$) test is used when the data are categorical. It measures how different the observed data are from what we would expect if $H_0$ was true.
The chi-square statistic

The chi-square ($\chi^2$) statistic compares observed and expected counts.

- **Observed counts** are the actual number of observations of each type.
- **Expected counts** are the number of observations that we would expect to see of each type if the null hypothesis was true.

\[
\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}
\]

(calculated for each type separately and then summed)

Large values for $\chi^2$ represent strong deviations from the expected distribution under $H_0$, and will tend to be statistically significant.
The chi-square distributions

The $\chi^2$ distributions are a family of distributions that take only positive values, are skewed to the right, and are described by a specific degrees of freedom.

Published tables & software give the upper-tail area for critical values of many $\chi^2$ distributions.
Table D

Ex: df = 6

If $\chi^2 = 15.9$ the $P$-value is between 0.01–0.02.
Goodness of fit hypotheses

The chi-square test can be used to for a categorical variable (1 SRS) with **any number k of levels**. The null hypothesis can be that all population proportions are equal (uniform hypothesis)

Are hospital births uniformly distributed in the week?  
\[ H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = 1/7 \]

or that they are equal to some specific values, as long as the sum of all the population proportions in \( H_0 \) equals 1.

When crossing homozygote parents expressing two co-dominant phenotypes A and B, we would expect in F2  
\[ H_0: p_A = 1/4, \; p_{AB} = 1/2, \; p_B = 1/4 \] where AB is an intermediate phenotype.
For 1 SRS of size \( n \) with \( k \) levels of a categorical variable

When testing

\[ H_0: p_1 = p_2 = \ldots = p_k \] (a uniform distribution)

The expected counts are all \( = \frac{n}{k} \)

When testing

\[ H_0: p_1 = p_{1H_0} \quad \text{and} \quad p_2 = p_{2H_0} \quad \ldots \quad \text{and} \quad p_k = p_{kH_0} \]

The expected counts in each level \( i \) are

\[
\text{expected count}_{i} = n \cdot p_{iH_0}
\]
Conditions for the goodness of fit test

The chi-square test for goodness of fit is used when we have a single SRS from a population and the variable is categorical with $k$ mutually exclusive levels.

We can safely use the chi-square test when:

- all expected counts have values $\geq 1.0$
- no more than 20% of the $k$ expected counts have values $< 5.0$
The chi-square statistic for goodness of fit with \( k \) proportions measures how much observed counts differ from expected counts. It follows the chi-square distribution \textbf{with \( k - 1 \) degrees of freedom} and has the formula:

\[
X^2 = \sum \frac{(\text{count of outcome } i - np_{i0})^2}{np_{i0}}
\]

The \( P \)-value is the tail area under the \( X^2 \) distribution with df = \( k - 1 \).
Aphids evade predators (ladybugs) by dropping off the leaf. An experiment examined the mechanism of aphid drops.

“When dropped upside-down from delicate tweezers, live aphids landed on their ventral side in 95% of the trials (19 out of 20). In contrast, dead aphids landed on their ventral side in 52.2% of the trials (12 out of 23).”

Is there evidence (at significance level 5%) that live aphids’ landing right side up (on their ventral side) is not a chance event? We test

\[ H_0: p_{\text{ventral}} = 0.5 \text{ and } p_{\text{dorsal}} = 0.5 \quad \quad \quad H_a: H_0 \text{ not true} \]

<table>
<thead>
<tr>
<th>Observed counts</th>
<th>Expected counts</th>
<th>( \chi^2 ) contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>20(0.5) = 10</td>
<td>((19 - 10)^2/10 = 8.1)</td>
</tr>
<tr>
<td>1</td>
<td>20(0.5) = 10</td>
<td>((19 - 10)^2/10 = 8.1)</td>
</tr>
</tbody>
</table>

Expected counts are large enough

\[ \chi^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}} = 8.1 + 8.1 = 16.2 \]

\[ \chi^2 \text{cdf}(16.2; 1) 
\quad 5.699411621e-5 \]

\( P\text{-value} 0.00006 \)
“When dropped upside-down from delicate tweezers, live aphids landed on their ventral side in 95% of the trials (19 out of 20).

Is there evidence (at significance level 5%) that live aphids’ landing right side up (on their ventral side) is not a chance event? We test

\[ H_0: p_{\text{ventral}} = 0.5 \text{ and } p_{\text{dorsal}} = 0.5 \quad H_a: H_0 \text{ not true} \]

From Table D, we find \( X^2 > 12.12 \), so \( P < 0.0005 \) (software gives \( P \)-value = 0.00006), highly significant. We reject \( H_0 \).

We have found very strong evidence that the righting behavior of live aphids is not chance (\( P < 0.0005 \)).
Interpreting the $\chi^2$ output

The individual values summed in the $\chi^2$ statistic are the $\chi^2$ components (or contributions). When the test is statistically significant,

- the largest components indicate which condition(s) are most different from the expected $H_0$. Compare the observed and expected counts to interpret the findings.

- You can also compare the actual proportions qualitatively in a graph.

### Observed counts vs. Expected counts

<table>
<thead>
<tr>
<th>Observed counts</th>
<th>Expected counts</th>
<th>$\chi^2$ contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>20(0.5) = 10</td>
<td>$(19 - 10)^2/10 = 8.1$</td>
</tr>
<tr>
<td>1</td>
<td>20(0.5) = 10</td>
<td>$(19 - 10)^2/10 = 8.1$</td>
</tr>
</tbody>
</table>

Aphids land the right side up (ventral side) a lot more often (95%) than chance alone would predict (50%).
Lack of significance: Avoid a logical fallacy

A non-significant $P$-value is not conclusive: $H_0$ could be true, or not.

This is particularly relevant in the $X^2$ goodness of fit test where we are often interested in $H_0$ that the data fit a particular model.

- A significant $P$-value suggests that the data do not follow that model.

- But finding a non-significant $P$-value is NOT a validation of the null hypothesis and does NOT suggest that the data do follow the hypothesized model. It only shows that the data are not inconsistent with the model.
The Frizzle fowl is a chicken with curled feathers. A genetic crossing between Frizzle and Leghorn (straight-feathered) varieties produced chickens in F1 all with slightly frizzled feathers. Here are the observed counts in F2, by phenotype:

<table>
<thead>
<tr>
<th>Phenotype (feather type)</th>
<th>Observed counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frizzled</td>
<td>23</td>
</tr>
<tr>
<td>Slightly frizzled</td>
<td>50</td>
</tr>
<tr>
<td>Straight</td>
<td>20</td>
</tr>
</tbody>
</table>

The most likely genetic model is that of a single gene locus with two codominant alleles producing a 1:2:1 ratio in F2. Are the data consistent with such a model?

\[
H_0: \ p_F = 1/4; \ p_{SF} = 2/4; \ p_S = 1/4 \quad H_a: \ H_0 \text{ is not true}
\]

Test Category | Observed | Proportion | Expected | Contribution to Chi-Sq |
---|---|---|---|---|
F | 23 | 0.25 | 23.25 | 0.002688 |
SF | 50 | 0.50 | 46.50 | 0.263441 |
S | 20 | 0.25 | 23.25 | 0.454301 |

\[
\chi^2 = \frac{(23 - 23.25)^2}{23.25} + \frac{(50 - 46.5)^2}{46.5} + \frac{(20 - 23.25)^2}{23.25} \approx 0.72
\]

\[
\text{N DF Chi-Sq P-Value } 93 \ 2 \ 0.720430 \ 0.698
\]
The most likely genetic model is that of a single gene locus with two codominant alleles producing a 1:2:1 ratio in F2. Are the data consistent with such a model?

$H_0$: $p_F = 1/4; p_{SF} = 2/4; p_S = 1/4 \quad H_a: H_0$ is not true

<table>
<thead>
<tr>
<th>TABLE D</th>
<th>Chi-square distribution critical values</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>.25</td>
</tr>
<tr>
<td>2</td>
<td>2.77</td>
</tr>
</tbody>
</table>

$\chi^2 = 0.72$

<table>
<thead>
<tr>
<th>N</th>
<th>DF</th>
<th>Chi-Sq</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>93</td>
<td>2</td>
<td>0.720430</td>
<td>0.698</td>
</tr>
</tbody>
</table>

Degrees of freedom = $k - 1 = 2$, and $\chi^2 = 0.72, P > 0.25$ (Table D). Software gives $P = 0.698$. This is not statistically significant and we fail to reject $H_0$.

The observed data are consistent with a single gene locus with two codominant alleles (1:2:1 genetic model). The small observed deviations from the model could simply have arisen from the random sampling process alone.