Hypothesis Testing I: $\chi^2$ distribution

Introduction to Data Science Algorithms
Jordan Boyd-Graber and Michael Paul

OCTOBER 4, 2016
Goodness of Fit

Suppose we see a die rolled 36 times with the following totals.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

- $H_0$: fair die
- How far does it deviate from uniform distribution?
Goodness of Fit

Suppose we see a die rolled 36 times with the following totals.

\[
\begin{array}{ccccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
8 & 5 & 9 & 2 & 7 & 5 \\
\end{array}
\]

- \(H_0\): fair die
- How far does it deviate from uniform distribution?
- \(\chi^2\) distribution
Let $Z_1, \ldots Z_n$ be independent random variables distributed $N(0, 1)$. The $\chi^2$ distribution with $n$ degrees of freedom can be defined by

$$\chi_n^2 \equiv Z_1^2 + Z_2^2 + \cdots + Z_n^2$$ (1)
Chi-Square Definition

\[ df = 1 \]

\[ df = 2 \]

\[ df = 3 \]

\[ df = 5 \]

\[ df = 10 \]

\[ x^2 \]
Chi-Square Distributions

### PDF
\[ \frac{1}{2^n \Gamma \left( \frac{n}{2} \right)} x^{\frac{n}{2} - 1} \exp \left\{ -\frac{x}{2} \right\} \]

### CDF
\[ \frac{1}{2^n \Gamma \left( \frac{n}{2} \right)} \gamma \left( \frac{n}{2}, \frac{x}{2} \right) \]

- \( \gamma (s, x) \equiv \int_0^x t^{s-1} \exp \{ -t \} \, dt \)
- \( \Gamma (x) \equiv \int_0^\infty t^{x-1} \exp \{ -t \} \, dt, \Gamma (n) = (n - 1)! \)
Goodness of Fit

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>8</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Expected</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

- If this were a fair die, all observed counts would be close to expected
- We can summarize this with a test statistic

\[
\sum \frac{(O_i - E_i)^2}{E_i}
\] (2)
Goodness of Fit

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>8</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Expected</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

- If this were a fair die, all observed counts would be close to expected
- We can summarize this with a test statistic

$$\sum \frac{(O_i - E_i)^2}{E_i}$$  \hspace{1cm} (2)

- In our example, 5.33
- Approximately distributed as $\chi^2$ with $k-1$ degrees of freedom
Test Statistic and $p$-value

- Expected value of $\chi^2$ with df=5 is 5
- 5.33 is not that far away
- 0.38 probability of rejecting the null
Degrees of Freedom

- We condition on the number of observations (36)
- So after filling in the cells for five observations, one is known
- So total of $k - 1$ degrees of freedom
Degrees of Freedom

- We condition on the number of observations (36)
- So after filling in the cells for five observations, one is known
- So total of \( k - 1 \) degrees of freedom
- Important because it specifies which \( \chi^2 \) distribution to use