10 Goodness-of-Fit with applications to Contingency Table data

10.1 Goodness-of-Fit

10.1.1 Introduction

In many of the statistical procedures discussed so far, it has been assumed that the random samples have been drawn from either a known and specified distribution, or perhaps a particular parametric family of distributions. The aim of this section is to present procedures for testing whether these assumptions are consistent with the observations that arise in practice.

10.1.2 The Multinomial Distribution

Consider a Bernoulli trial. There are 2 possible outcomes, success(S) or failure(F), where

\[ P(S) = p \quad \text{and} \quad P(F) = 1 - p = q. \]

Setting \( X(S) = 1 \), and \( X(F) = 0 \), then \( X \sim Bernoulli(p) \sim Bin(1, p) \).

Let us generalize this scenario.

Suppose now that at each trial, there are \( k \) possible outcomes, \( A_1, A_2, \ldots, A_k \).

Setting \( p_j = P(A_j) \), then \( p_j \geq 0 \), \( j = 1, \ldots, k \), and \( \sum_{j=1}^{k} p_j = 1 \).

Suppose that \( n \) independent trials of the experiment are performed, and let \( Y_j \) be the number of times that the outcome \( A_j \) occurs in those \( n \) trials.

Then \( Y_j \sim Bin(n, p_j) \), since at each trial, the outcome is either \( A_j \) (corresponding to a success) or \( A_c^j \) (corresponding to a failure).

It is possible to specify the distribution of the random vector \( \underline{Y} = (Y_1, Y_2, \ldots, Y_k) \), based on \( n \) independent trials, as follows.

Proposition 10.1.1 (The Multinomial Distribution)

The joint p.m.f. of \( \underline{Y} = (Y_1, Y_2, \ldots, Y_k) \) is

\[ p_{\underline{Y}}(\underline{x}|\underline{p}) = P(Y_1 = x_1, Y_2 = x_2, \ldots, Y_k = x_k | \underline{p}) \]

\[ = \frac{n!}{x_1!x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \]

\( 0 \leq x_j \leq n, \ j = 1, \ldots, k, \) and \( \sum_{j=1}^{k} x_j = n \).

i.e. \( \underline{Y} \) has a \( k \)-dimensional MULTINOMIAL DISTRIBUTION with parameters \( n \) and \( \underline{p} = (p_1, p_2, \ldots, p_k) \).
Remarks 10.1.2
Even though $Y$ has been generated by a random sample $X_1, X_2, \ldots, X_n$ (which, by definition, are mutually independent), $Y_1, Y_2, \ldots, Y_k$ are actually dependent since

$$
\sum_{j=1}^{k} Y_j = n.
$$

10.1.3 Goodness-of-fit for Discrete distributions

Example 10.1.3 (Is the die fair?)

A die is rolled 60 times; the results are summarized in the 1st and 2nd lines of the following table.

<table>
<thead>
<tr>
<th>Score</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>11</td>
<td>5</td>
<td>12</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>Expected</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Compare this with the expected results in the final line of the table under the assumption that the die is fair.

Let $p_j$ be the probability that the outcome is $j$, for $j = 1, 2, \ldots, 6$. Then we wish to test $H_0 : p_j = \frac{1}{6}, j = 1, 2, \ldots, 6$, versus $H_1 : p_j \neq \frac{1}{6}$ for at least one $j, j = 1, 2, \ldots, 6$.

For this example, under $H_0$, $Y = (Y_1, Y_2, \ldots, Y_6)$ has the multinomial distribution with parameters $n = 60$ and $\underline{p} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$.

We now need to find an appropriate test statistic for deciding whether to accept or reject $H_0$. The following statistic provides one possibility:

**Definition 10.1.4 (Pearson’s Chi-Square statistic)**

$$
\chi^2 = \sum_{j=1}^{k} \frac{(Y_j - E[Y_j])^2}{E[Y_j]}
$$

is called **Pearson’s Chi-square statistic**.

We have an asymptotic distributional result for this statistic.

**Proposition 10.1.5 (A large sample result for the Chi-square statistic)**

For large (sample size) $n$,

$$
\chi^2 = \sum_{j=1}^{k} \frac{(Y_j - E[Y_j])^2}{E[Y_j]} \sim \chi^2_{k-1}
$$

approximately.

**Note:** Do not confuse $k$ with $n$. 

2
Example 10.1.6 (Is the die fair? continued)

If the observations were consistent with $H_0$, then we would expect the $Y_j$ to be 'close' to the $E[Y_j]$ evaluated under $H_0$, and so $X^2$ would be relatively 'small'. On the other hand, if the observations were inconsistent with $H_0$, we might expect the $Y_j$ to be quite far from the $E[Y_j]$ evaluated under $H_0$, and so $X^2$ could take relatively 'large' values. So a possible decision procedure would be to

Accept $H_0$: if $X^2$ is small,

Reject $H_0$: if $X^2$ is large,

or more formally

Accept $H_0$: if $X^2 \leq c$,

Reject $H_0$: if $X^2 > c$,

for $c > 0$.

The details for a test at the $100\alpha\%$ level are found as follows.

$$\alpha = \mathbb{P}(\text{Reject } H_0 | H_0 \text{ true}) = \mathbb{P}(X^2 > c | p_j = \frac{1}{6}, j = 1, \ldots, 6) \approx \mathbb{P}(Y > c)$$

where $Y$ has a $\chi^2_5$-distribution, using Proposition 10.1.5.

Thus, set $c = \chi^2_{5,0.1}$.

The observed value of $X^2$ here is

$$\frac{1}{10}((8 - 10)^2 + (11 - 10)^2 + (5 - 10)^2 + (12 - 10)^2 + (15 - 10)^2 + (9 - 10)^2)$$

$$= \frac{1}{10}(2^2 + 1^2 + 5^2 + 2^2 + 5^2 + 1^2) = \frac{60}{10} = 6.$$ 

But $\chi^2_{5,0.1} = 9.236$, and thus the $p$-value is at least 0.1.

Hence, we do not reject $H_0$ (i.e. die is fair) at the 10% level of significance.

Remarks 10.1.7

- $X^2 = \sum_{j=1}^k \frac{(Y_j - E[Y_j])^2}{E[Y_j]} = \sum_{j=1}^k \frac{(Y_j - np_j)^2}{np_j}$.

- If we denote the observed value of $Y_j$ by $o_j$, and $E[Y_j] = np_j$ evaluated under $H_0$ by $e_j$, then the observed value of $X^2$ is:

$$\sum_{j=1}^k \frac{(o_j - e_j)^2}{e_j} = \sum_{j=1}^k \frac{o_j^2}{e_j} - 2 \sum_{j=1}^k o_j \sum_{j=1}^k e_j$$

$$= \sum_{j=1}^k \frac{o_j^2}{e_j} - n$$

since $\sum_{j=1}^k o_j = \sum_{j=1}^k e_j = n$.

Either (1) or (2) can be used for calculation.
10.1.4 Goodness-of-fit for any specified p.m.f./p.d.f.

Following along the lines of the previous section, we will now construct a Goodness-of-fit test that deals with any specified or known p.m.f./p.d.f.

Suppose that we have an observable characteristic for a population, $X$ say. We wish to devise a test for

$H_0 : X \sim f$ (where $f$ is a completely specified p.m.f./p.d.f.)

versus

$H_1: H_0$ is false

on the basis of a random sample $X_1, X_2, \ldots, X_n$.

We partition $R_X$, i.e. the set of all possible values of $X$, into $k$ disjoint cells $A_1, A_2, \ldots, A_k$, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, and $\bigcup_j A_j = R_X$.

Let $Y_j$, $j = 1, \ldots, k$, be the number of $X_i$’s from the random sample which take values in $A_j$ and set $p_j = P(X \in A_j|X \sim f)$. Then, under $H_0$, $Y = (Y_1, Y_2, \ldots, Y_k)$ has a $k$-dimensional multinomial distribution with parameters $n$ and $p = (p_1, p_2, \ldots, p_k)$.

Now consider the statistic

$$\chi^2_{obs} = \sum_{j=1}^{k} \frac{(o_j-e_j)^2}{e_j}$$

where $e_j = nP(X \in A_j|X \sim f)$.

If the observed random sample is ‘consistent’ with $X \sim f$, then $\chi^2$ should be small; on the other hand, if the data is inconsistent with $X \sim f$, then $\chi^2$ could be large.

Example 10.1.8 (Days lost due to staff illness)

It is believed that the number of days taken off per year by individual staff due to illness is distributed as a random variable $X$ from the Po(3) distribution.

Is this a reasonable assumption?

**Solution:**

We wish to test

$H_0: X \sim Po(3)$ versus $H_1: H_0$ false

on the basis of data collected on 40 individuals that are thought to constitute a random sample. The data are summarized in the first two lines of the table below.

<table>
<thead>
<tr>
<th>Days Lost</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>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Prob. under $H_0$</td>
<td>0.050</td>
<td>0.149</td>
<td>0.224</td>
<td>0.224</td>
<td>0.168</td>
<td>0.101</td>
<td>0.050</td>
<td>0.034</td>
</tr>
<tr>
<td>Exptd. Nos. under $H_0$</td>
<td>2.00</td>
<td>5.96</td>
<td>8.96</td>
<td>8.96</td>
<td>6.72</td>
<td>4.04</td>
<td>2.00</td>
<td>1.36</td>
</tr>
<tr>
<td>Pooled (exptd.) $e_j$</td>
<td>7.96</td>
<td>8.96</td>
<td>8.96</td>
<td>6.72</td>
<td></td>
<td></td>
<td>7.40</td>
<td></td>
</tr>
<tr>
<td>Pooled (observed.) $o_j$</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td></td>
<td></td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

According to the way that we’ve ‘pooled’ the data, $Y$ comes from the 5-dimensional multinomial distribution with parameters 40 and $p = (0.050+0.149, 0.224, 0.224, 0.168, 0.101+0.050+0.034)$.
The observed value of $X^2$ is
\[
\sum_{i=1}^{5} \frac{(o_i-e_i)^2}{e_i} = \frac{(4-7.96)^2}{7.96} + \frac{(7-8.96)^2}{8.96} + \frac{(6-8.96)^2}{8.96} + \frac{(10-6.72)^2}{6.72} + \frac{(13-7.40)^2}{7.40} = 9.21545.
\]

From tables, we find that $\chi^2_{4,0.10} = 7.779$ and $\chi^2_{4,0.05} = 9.488$ (and thus the $p$-value lies between 0.05 and 0.1). Therefore, there is evidence to doubt $H_0$ at the 10% level of significance, but not at the 5% level. Not very significant.

Remarks 10.1.9
- The pooling (to form the $A_j$’s) was carried out on the basis of the expected numbers, not the observed numbers.
- As a ‘rule of thumb’, the $A_j$’s should be constructed in a convenient way (before looking at the data) to satisfy the requirement

\[ e_j = np_j = n\mathbb{P}(X \in A_j | X \sim f) \geq 5. \]

10.2 Contingency Tables

10.2.1 Introduction

Example 10.2.1 (Count data presented in a contingency table)
A random sample of 90 chartered accountants working in the South-East was taken. The accountants were classified according to their salary bracket, and whether or not they were university graduates. The data are presented in the table below.

<table>
<thead>
<tr>
<th>Qualification</th>
<th>Salary (\pounds)</th>
<th>16-20</th>
<th>20-25</th>
<th>25-30</th>
<th>30-35</th>
<th>35+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADUATES</td>
<td></td>
<td>6</td>
<td>11</td>
<td>16</td>
<td>14</td>
<td>13</td>
<td>60</td>
</tr>
<tr>
<td>NON-GRADUATES</td>
<td></td>
<td>5</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>11</td>
<td>20</td>
<td>24</td>
<td>20</td>
<td>15</td>
<td>90</td>
</tr>
</tbody>
</table>

This table is known as a $k \times h$ (2-way) contingency table.

For this example there are 2 factors:
- **Qualification**:- that occurs at 2 levels - GRADUATES or NON-GRADUATES, and
- **Salary (\pounds)**:- that occurs at 5 levels - 16-20, 20-25, 25-30, 30-35, 35+.

For this example, therefore, $k = 2$ and $h = 5$.

A question of interest is whether the possession of a university degree really has an effect on salary; is it the case that the factor **Salary** is independent of the factor **Qualification**?
10.2.2 Testing for Independence

Suppose that each member of a random sample of size \( n \) can be classified according to \( k \) mutually exclusive and exhaustive events \( A_1, A_2, \ldots, A_k \) w.r.t. factor \( F_1 \), and \( h \) mutually exclusive and exhaustive events \( B_1, B_2, \ldots, B_h \) w.r.t. factor \( F_2 \).

Let the r.v. \( Y_{ij} \) be the count or tally of the members of the random sample for which the event \( A_i \cap B_j \) occurs, and \( y_{ij} \) be its observed value, \( i = 1, \ldots, k; \ j = 1, \ldots, h \).

Further, set

\[
Y_i = \sum_{j=1}^{h} Y_{ij}, \quad Y_j = \sum_{i=1}^{k} Y_{ij}
\]

with observed values

\[
y_i = \sum_{j=1}^{h} y_{ij}, \quad y_j = \sum_{i=1}^{k} y_{ij}.
\]

Then independence of the two factors can be tested in the following way.

**Proposition 10.2.2 (Chi-square test for independence)**

Suppose that we have a \( k \times h \) contingency table with factors \( F_1 \) and \( F_2 \), each having \( k \) and \( h \) levels respectively, for the data associated with a random sample of size \( n \).

Then we reject the hypothesis that \( F_1 \) and \( F_2 \) are independent at the (approximate) 100\( \alpha \)% level of significance if

\[
Q_{\text{obs}} = \frac{\sum_{j=1}^{h} \sum_{i=1}^{k} (y_{ij} - n(y_i/n)(y_j/n))^2}{n(y_i/n)(y_j/n)} > \chi^2_{(k-1)(h-1),\alpha}
\]

*Derivation:*

Let \( p_{ij} = \mathbb{P}(A_i \cap B_j) \), \( i = 1, 2, \ldots, k \), \( j = 1, 2, \ldots, h \).

Consider the statistic

\[
\tilde{Q} = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{ij} - np_{ij})^2}{np_{ij}}.
\]

Since there are \( kh \) events of the form \( A_i \cap B_j \), then \( \tilde{Q} \) is associated with \( kh - 1 \) degrees of freedom. Thus,

\( \tilde{Q} \sim \chi^2_{kh-1} \) approximately for large \( n \).

Set \( p_i := \mathbb{P}(A_i) = \sum_{j=1}^{h} p_{ij} \) and \( p_j := \mathbb{P}(B_j) = \sum_{i=1}^{k} p_{ij} \)

The hypothesis for independence can be stated as

\[
H_0 : \mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j), \ i = 1, \ldots, k; \ j = 1, \ldots, h
\]

i.e.

\[
H_0 : \ p_{ij} = p_i \times p_j, \ i = 1, \ldots, k; \ j = 1, \ldots, h.
\]
However, we will generally not have knowledge beforehand about the terms in the above hypothesis, and so we need to estimate them. The terms \( p_i \) and \( p_j \) may be estimated by

\[
\hat{p}_i = \frac{Y_i}{n}, \quad \hat{p}_j = \frac{Y_j}{n}
\]

\( i = 1, \ldots, k \), \( j = 1, \ldots, h \), respectively.

Thus, to test \( H_0 \), we replace \( p_{ij} \) by

\[
\hat{p}_{ij} = \frac{Y_i}{n} \times \frac{Y_j}{n}
\]

in \( \tilde{Q} \), yielding the new test statistic

\[
Q = \sum_{j=1}^{h} \sum_{i=1}^{k} \left[ Y_{ij} - n(Y_i/n)(Y_j/n) \right]^2 / n(Y_i/n)(Y_j/n)
\]

which has a \( \chi^2_{(k-1)(h-1)} \)-distribution approximately, for large \( n \).

[Degrees of freedom given by \((kh - 1) - (k - 1) - (h - 1) = kh - k - h + 1 = (k - 1)(h - 1)\).]

If the hypothesis of independence is untrue, then we could expect \( Q \) to take large values. Thus we reject \( H_0 \) at the (approximate) 100\( \alpha \)% level if

\[ Q_{obs} > \chi^2_{(k-1)(h-1), \alpha} \]

Example 10.2.3 (Example 10.2.1 continued)

The table of data is presented again, but this time with the estimated expected frequencies, the \( n(y_i/n)(y_j/n) \), included (for comparison with the actual observed frequencies), which are in brackets.

<table>
<thead>
<tr>
<th>Qualification</th>
<th>Salary(£)</th>
<th>16-20</th>
<th>20-25</th>
<th>25-30</th>
<th>30-35</th>
<th>35+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADUATES</td>
<td></td>
<td>6</td>
<td>11</td>
<td>16</td>
<td>14</td>
<td>13</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(7.333)</td>
<td>(13.333)</td>
<td>(16)</td>
<td>(13.333)</td>
<td>(10)</td>
<td></td>
</tr>
<tr>
<td>NON-GRADUATES</td>
<td></td>
<td>5</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>11</td>
<td>20</td>
<td>24</td>
<td>20</td>
<td>15</td>
<td>90</td>
</tr>
</tbody>
</table>

For this example,

\[
Q_{obs} = \frac{[6 - 90(60/90)(11/90)]^2}{90(60/90)(11/90)} + \frac{[11 - 90(60/90)(20/90)]^2}{90(60/90)(20/90)} + \ldots
\]

\[
\ldots + \frac{[6 - 90(30/90)(20/90)]^2}{90(30/90)(20/90)} + \frac{[2 - 90(30/90)(15/90)]^2}{90(30/90)(15/90)}
\]

\[
= \frac{[6 - 7.333]^2}{7.333} + \frac{[11 - 13.33]^2}{13.33} + \ldots + \frac{[6 - 6.667]^2}{6.667} + \frac{[2 - 5]^2}{5}
\]

\[
= 4.752 < 9.488 = \chi^2_{4, 0.05}.
\]

Thus the \( p \)-value of this test is certainly greater than 0.05. Thus we are inclined to think that there is no evidence against the hypothesis of independence (at the 5% level). Graduates do not seem to earn salaries which are markedly different from those of non-graduates.