Lecture 21: Specification Tests

Relevant textbook passages:
Larsen–Marx [8]: Sections 10.3, 10.4.

21.1 Specification testing

Today we will take up the topic of deciding whether our parametric data model $f(x; \theta)$ with parameters $\theta \in \Theta$ is a “good” model. That is rather than testing hypotheses about the parameter $\theta$, we are interested in test concerning the function $f$. These kinds of tests are usually referred to as specification tests.

For instance, as a Southern Californian, I happen to be interested in whether earthquakes follow a Poisson process. If so, the time between main earthquake shocks follows an Exponential($\lambda$) distribution for some $\lambda$. So one partial test of the Poisson Process model of earthquakes would be to test whether the time between earthquakes follows an exponential distribution. The straightforward obvious approach to this would be to embed the class of exponential in a larger class, say the Gamma family. We could then use a generalized likelihood ratio test to test the null of an exponential hypothesis against the alternative of a general Gamma distribution. Since the Exponential($\lambda$) distribution is also the Gamma(1, $\lambda$), these hypotheses are nested. In order to use the likelihood ratio test, we need to be able to compute the density of our test statistic, and in the particular case that seems rather do-able as these things go. But suppose we choose as our alternative the Normal family. In this case the distribution is more complicated.

It turns out there is a relatively simple way to test whether the data come from a given continuous distribution. This approach is based on the fact that the quantiles of a continuous distribution are uniformly distributed. If we translate our data into quantiles, we can define test statistics in terms of Q-Q plots that have known (or computable distributions). This gives rise to a number of tests, the best known of which is the Kolmogorov–Smirnov test, and we will take this up in the next section.

For data that are not continuous, the quantile approach is useful. Data of this sort are typically counts of the number of observations that fit into one of a set of categories. Again, going back to earthquakes, if the Poisson Process model is a good model, then the number of earthquakes per year should follow a
Poisson distribution, so we should test that. Such tests are often called **goodness-of-fit** tests, but they just hypothesis tests. A particularly useful such test was characterized by Karl Pearson\(^1\) in 1900 [11], and is known as the **chi-square test**. One of the uses of the chi-square test is testing whether two random variables are stochastically independent. This is a test of the null hypothesis \(f(x, y) = f_X(x) f_Y(y)\) on the distribution, so it too comes under the heading of a specification test.

By “**binning**” continuous data, say by constructing a histogram, one can convert continuous data into categorical data, and the chi-square test is often used on continuous data.

### 21.2 Testing continuous distributions

You may have already forgotten this, but in your homework you have used Normal Q-Q plots to get an “eyeball” test of the hypothesis that the data are normally distributed. But we can define test statistics based on these plots as well. The most familiar is the **Kolmogorov–Smirnov test**. Surprisingly, it does not appear in the textbook [8]. But you can find it discussed by Breiman [3, pp. 213–217], van der Waerden [14, § 16, pp. 60–75], or Wikipedia [http://en.wikipedia.org/wiki/Kolmogorov–Smirnov_test]. The treatment here relies heavily on Breiman’s Chapter 6.

The idea is this. Recall from Lecture 8 that given independent and identically distributed random variables \(X_1, \ldots, X_n, \ldots\), the **empirical distribution function** is defined by

\[
F_n(x) = \frac{|\{i : i \leq n \land X_i \leq x\}|}{n},
\]

or in terms of indicator functions

\[
F_n(x) = \frac{\sum_{i=1}^{n} 1(-\infty, x](X_i)}{n}.
\]

The Glivenko–Cantelli Theorem 8.9.3 asserts that if \(F\) is the common cumulative distribution function of the \(X_i\)s, then

\[
\text{Prob} \left( \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{n \to \infty} 0 \right) = 1.
\]

This suggests the following test statistic:

\[
K(x_1, \ldots, x_n) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|.
\]

If you are worried that finding the global supremum \(|F_n(x) - F(x)|\) may be hard, observe that \(F_n\) is a step function, and \(F\) is continuous, so the maximal difference

\(^1\)Karl Pearson is not the Pearson of the Neyman–Pearson Theorem. That Pearson is Egon Pearson, Karl’s son.
must come at one of the jumps in $F_n$. Thus we only need to check $|F_n(x) - F(x)|$ and $|F_n(x_{i-1}) - F(x)|$ for $i = 1, \ldots, n$. The distribution of this statistic depends on $F$ and so it may a difficult one to use. However, there is a transformation we can use to eliminate this dependence.

Recall (Proposition 10.1.1) that for any random variable $X$ with a continuous cumulative distribution function $F$ that $F(X)$ is a Uniform$[0, 1]$ random variable. Here is a recap of the proof for the simpler case where $F$ is strictly increasing: Let $x_p$ satisfy $F(x_p) = p$. Since $F$ is strictly increasing and continuous,

$$P(F(X) \leq p) = P(X \leq x_p) = F(x_p) = p.$$ 

So the procedure to use is this:

- Formulate a Null Hypothesis,

\[ H_0 : \text{the cumulative distribution function } F \text{ of } X_i \text{ is } F_0. \]

If we want to test the hypothesis that $F$ is some exponential, we should use the MLE of $\lambda$ and take $F_0$ to be the cumulative distribution function of an Exponential$(\hat{\lambda}_{MLE})$.

- Transform each $X_i$ via

\[ Y_i = F_0(X_i). \]

- If the Null Hypothesis, is true, then each $Y_i$ is a Uniform$[0, 1]$ random variable. Recall that the $F_U$ cumulative distribution function of a Uniform is $F_U(y) = y$ for $0 \leq y \leq 1$.

- Compute the test statistic

\[ K = \sup_{0 \leq y \leq 1} |G_n(y) - y| = \sup_x |F_n(x) - F_0(x)|, \]

where $G_n$ is the empirical cumulative distribution function of the $Y_i$s:

\[ G_n(y) = \frac{\sum_{i=1}^n 1_{0,y}(y_i)}{n}. \]

The supremum is actually a maximum and we only need to compare $G_n(y)$ to $y$ (which is the Uniform cumulative distribution function) at only finitely many points.

- Since the values at which the empirical cumulative distribution function jumps are actually the order statistics of $Y_1, \ldots, Y_n$ (which under the null hypothesis are known Beta random variables),

the distribution of the test statistic $K$ is independent of $F_0$!
• This is not to say that the distribution of $K$ is not complicated, but you can find tables for test values in van der Waerden [14, Table 4, p. 344] or Breiman [3, p. 212]. The null hypothesis is rejected if $K$ is larger than the cutoff.

Here is the table of cutoffs from Breiman:

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\alpha = 0.1$</th>
<th>$\alpha = 0.05$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.37</td>
<td>.41</td>
</tr>
<tr>
<td>20</td>
<td>.26</td>
<td>.29</td>
</tr>
<tr>
<td>30</td>
<td>.22</td>
<td>.24</td>
</tr>
<tr>
<td>large</td>
<td>$1.22/\sqrt{n}$</td>
<td>$1.36/\sqrt{n}$</td>
</tr>
</tbody>
</table>

But nowadays, you don’t need the table. With R, use the `ks.test` command with the `pnorm` distribution. (See the documentation or Dytham [5, pp. 86–89].) With Mathematica, use the `KolmogorovSmirnovTest` command.

### 21.2.1 Caveats

Here are some points to keep in mind:

• The Kolmogorov–Smirnov test is designed to test the null hypothesis $F = F_0$. If you want to test a composite hypothesis $F \in \Theta_0$, you first estimate a $\theta_0$, typically by MLE, and test the simple hypothesis. Breiman [3, p. 213] says, “The effect that this has on the on the level of the test is not well known. The evidence we have is that the effect is not very important. For moderate to large sample sizes, it is probably safe to ignore the fact that $\theta$ was estimated.”

• The Kolmogorov–Smirnov test is not very powerful, and the power is hard to estimate. According to Breiman, if

$$K \leq \begin{cases} 
0.65/\sqrt{n} & \alpha = 0.1 \\
0.85/\sqrt{n} & \alpha = 0.05 
\end{cases}$$

the power of the test is too small to be useful.

• If the Kolmogorov–Smirnov test does reject the Null Hypothesis, the Q-Q graph of the quantiles provide useful insights into the nature of the data generating process behind the data.

• While the Kolmogorov–Smirnov test is the best known test for based on the empirical cdf, there are many others.

  • The Anderson–Darling[?] test statistic is also measure of the distance of the empirical cdf from the null cdf, but it emphasizes the tails. It is given by

$$A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F(x) (1 - F(x))} f_0(x) \, dx.$$
It too can be used with transformed data to get a statistic whose distribution is independent of $F_0$.

○ The Shapiro–Wilk \cite{7} test is based on order statistics, rather the empirical cumulative distribution function. It is expressly designed to test whether the data are Normally distributed.

Razali and Wah \cite{13} discuss a number of tests and provide references to many others. They argue on the basis of Monte Carlo studies that the Shapiro–Wilk test is more powerful for testing Normality.

### 21.3 Review of the multinomial distribution

The **multinomial distribution** generalizes the binomial distribution to random experiments with more than two outcomes or results. Let there be $t$ possible results, $r_1, \ldots, r_t$. Assume result $r_i$ has probability $p_i$. Let $X_i$ be the number of occurrences of result $r_i$ in $n$ independent trials. Then

$$p_X(k_1, \ldots, k_t) = \frac{n!}{k_1! \cdot k_2! \cdots k_t!} p_1^{k_1} \cdot p_2^{k_2} \cdots p_t^{k_t} \quad \left( \sum_i k_i = n \right).$$

One can show that each $X_i$ is Binomial$(n, p_i)$ \cite[Theorem 10.2.2, p. 496]{8}, and so

$$E X_i = n p_i.$$

$$\text{Var} X_i = n p_i (1 - p_i).$$

But the $X_i$s are not independent, since they must sum to $n$.

### 21.4 “Goodness of fit” tests

When we have a multinomial model, our null hypothesis can take the form

$$H_0: p = p_0$$

where $p_0$ is a vector of $t$ probabilities that sum to one. The alternative is typically

$$H_1: p \neq p_0.$$

Karl Pearson \cite{11} proposed the following test statistic for this kind of test,

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{t} \frac{X_i^2}{np_i} - n, \quad (1)$$

or the “sum of squares of (observed − expected) over the expected.” What does the distribution of this test statistic look like?
When \( t = 2 \) (the Binomial case), \( X_1 \) is a Binomial \((p_1)\) and \( X_2 = n - X_1 \). Thus \( D \) is of the form

\[
\frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n(1 - p_1))^2}{n(1 - p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(np_1 - X_1)^2}{n(1 - p_1)} = \frac{(1 - p_1)(X_1 - np_1)^2}{np_1(1 - p_1)} + \frac{p_1(np_1 - X_1)^2}{n(1 - p_1)p_1} = \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} = \left( \frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}} \right)^2 = \left( \frac{X_1 - E X_1}{\sqrt{\text{Var}X_1}} \right)^2
\]

Now \( (X_1 - E X_1)/\sqrt{\text{Var}X_1} \) is approximately a standard normal for largish \( n \), so \( D \) has an approximate Chi-square distribution with one degree of freedom.

It can be shown that in general, for \( t \geq 2 \), under the null hypothesis, the distribution of \( D \) is approximately \( \chi^2(t - 1) \). This is perhaps why Pearson titles his article “On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables is Such That it Can Reasonably Supposed to Have Arisen From Random Sampling.” (For a that proof uses the characteristic function, described in the forbidden appendix to Lecture 10, see Cramér [4, pp. 416–419]. For a more elementary but equally hard to follow proof, see van der Waerden [14, § 49, pp. 197–202 and § 51, pp.207–211]. Breiman [3, Problems 6–7, pp. 191–192] provides an exercise with many hints to prove the result.)

According to Theorem 10.3.1 in Larsen and Marx [8] we need \( np_i \geq 5 \), for all \( i = 1, \ldots, t \) to use this approximation, but van der Waerden [14, p. 238] thinks this is too conservative.

### 21.5 Goodness of fit with estimated parameters

The chi-square test described above assumed we had specified the probabilities as part of the null hypothesis. But typically we have to estimate the probabilities. In that case, we can still use a chi-square test, but test statistic has few degrees of freedom. Since the df represents the number of standard normals we are squaring, the critical values for a test with estimated probabilities will be smaller.

#### 21.5.1 Theorem [8, Theorem10.4.1][3, Theorem6.13, p. 196]

Let

\[
H_0: p = p_0.
\]
where $p_0$ is a $t$-dimensional vector of probabilities that sum to one. Estimate $s$ parameters that determine $p$ by the Maximum Likelihood Method. Then the test statistic

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{t} \frac{X_i^2}{np_i} - n,$$

has an approximately Chi-square distribution with $t - 1 - s$ degrees of freedom. (Need each $np_i \geq 5$.)

### 21.6 The wonderfulness of the Chi-square Test

#### 21.6.1 Example (The cookie data)

The reports I received from you on the number of chocolate chips in the cookies you ate were:

| # of chips | 0 1 2 3 4 5 6 7 8 9 10 11 |
| # of cookies | 0 0 1 1 0 7 6 1 1 2 2 1 |

This is a total of 140 chips in 22 cookies, so the MLE of $\mu$ is 6.4. To get at least 5 expected cookies in each bin, I grouped them as follows:

| Bin 1 | Bin 2 | Bin 3 |
| # of chips | 0–6 | 6 | > 6 |
| # of cookies | 9 6 7 |
| Expected | 8.56 | 3.50 | 9.94 |
| $\frac{(m_i - np_i)^2}{np_i}$ | 0.022 | 1.79 | 0.87 |

This gives the test statistic $= 2.69$. There are 3 bins and 1 estimated parameter, $\mu$, so the test statistic is $\sim \chi^2(1)$. The $p$-value for this statistic is 0.90, so we decisively fail to reject the Poisson hypothesis.

#### 21.6.2 Example (The World Series)

World Series come in lengths of 4, 5, 6, and 7, so there are $t = 4$ types of results of this experiment. The probability of length $\ell$ is

$$H_0: p_\ell = \left(\frac{\ell - 1}{3}\right)\left(p^4(1-p)^{\ell-4} + (1-p)^4p^{\ell-4}\right) \quad (\ell = 4, \ldots, 7).$$  \hfill (2)

Thus if there are $k_\ell$ series of length $\ell$ in our sample, the test statistic

$$\sum_{\ell=4}^{7} \frac{(k_\ell - np_\ell)^2}{np_\ell}$$

has a $\chi^2(3)$-distribution. We can use this to test the null hypothesis (2).

(To find critical values in Mathematica, use the InverseCDF command. It has two arguments, the first is a named distribution, the second the $\alpha$ or $1 - \alpha$ level.)

This assumes that we have fixed $p$ as part of the null hypothesis. If we must first estimate $p$, then our degrees of freedom are reduced. We’ll come to that in just a moment. 

KC Border
21.6.3 Example (Case Study 10.3.2 [8]: Benford’s Law)

The distribution of leading digits appears to be like this: The probability that a naturally occurring number in the wild begins with the digit \( d \) is
\[
\log_{10}(d + 1) - \log_{10}(d).
\]
This was first noticed by Simon Newcomb [10]. It became known as Benford’s Law because Frank Benford [1] rediscovered and publicized it. He had investigated among other things, baseball statistics, surface areas of rivers, and molecular weights of chemicals [8, pp. 502–505]. There are also sets of numbers that disobey Benson’s Law. Phone directories may disobey because the numbers often start with the same 3-digit exchange, especially in small communities. [When and where I went to high school all phone numbers began with 653- or 655-]

T. P. Hill [6] has developed a sophisticated probabilistic model that appears to explain Benford’s Law, but it is beyond the scope of this course. See also [7] for an accessible discussion.

Here is the table of probabilities:

<table>
<thead>
<tr>
<th>Digit ( d )</th>
<th>( \log_{10}(d + 1) - \log_{10}(d) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.301</td>
</tr>
<tr>
<td>2</td>
<td>0.176</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
</tr>
<tr>
<td>4</td>
<td>0.097</td>
</tr>
<tr>
<td>5</td>
<td>0.079</td>
</tr>
<tr>
<td>6</td>
<td>0.067</td>
</tr>
<tr>
<td>7</td>
<td>0.058</td>
</tr>
<tr>
<td>8</td>
<td>0.051</td>
</tr>
<tr>
<td>9</td>
<td>0.046</td>
</tr>
</tbody>
</table>

Notice that 30% of wild numbers start with 1!

**Forensic accounting**

Benford’s Law is used by forensic eaccounts to help detect embezzlers. It turns out the naïve embezzlers make up fake numbers where the leading digits tend to be more uniformly distributed that predicted by Benford’s Law. So deviations from Benford’s Law are a sign that something is amiss.

Larsen and Marx [8, pp. 502–505] cite as an example, the University of West Florida’s budget. The present counts of the leading digits and use a \( \chi^2 \) test statistic with 8 degrees of freedom to test Benford’s Law. The test statistic has a value of 2.49, and the CDF of \( \chi^2 \) with 8 degrees of freedom at 2.49 is 0.0378. So for the one-sided square test, we see that 96.2% of the samples would fit the model worse, so we do not reject it. The CDF of \( \chi^2 \) with 8 degrees of freedom at 15.507 is 0.949995. So 15.507 is the critical value of the Chi-square at the 5% level. Since 2.49 < 15.507 we fail to reject the null hypothesis that the data satisfy Benford’s Law.

Closer to home, my colleague Jean Ensminger and Caltech alum Jetson Leder-Luis are examining accounts from grants made by the World Bank to various
projects in Kenya. Their results won’t be ready until July 2015, but preliminary indications are that much the accounting data is not consistent with Benford’s Law.

21.6.4 Example (Case Study 10.3.3 [8]: Did Mendel Cheat?) See [8], pages 507–508.

Gregor Mendel categorized 556 specimens of garden peas on two traits, shape and color. The color could be g (green) or y (yellow), and shape could be a (angular) or r (round). His theory of genetics predicted the relative frequencies of these traits in the population of hybrids. The following table presents the reported number of plants in four categories along with Mendel’s theory’s predictions. We want to test the null hypothesis $H_0$: the data are consistent with Mendel’s theory.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Obs.</th>
<th>Mendel</th>
<th>Pred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ry</td>
<td>315</td>
<td>9/16</td>
<td>312.75</td>
</tr>
<tr>
<td>rg</td>
<td>108</td>
<td>3/16</td>
<td>104.25</td>
</tr>
<tr>
<td>ay</td>
<td>101</td>
<td>3/16</td>
<td>104.25</td>
</tr>
<tr>
<td>ag</td>
<td>32</td>
<td>1/16</td>
<td>34.75</td>
</tr>
</tbody>
</table>

The test statistic is

$$D = \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.47.$$ 

For $\chi^2(3)$ the CDF(0.47) = 0.0745689, so the $p$-value of $D$ for is about 0.925. This led R. A. Fisher to believe Mendel had faked his data.

21.6.5 Example Back to the World Series: 4 categories, so 3 degrees of freedom, but we estimated $p$ by MLE. So this reduces our degrees of freedom by 1. So the appropriate test involves a $\chi^2(2)$. So the critical value is 5.99 at the 5% level of significance.

(Mathematica: InverseCDF[ChiSquareDistribution[2], 0.95])

21.7 Some practical considerations

I mentioned earlier that one of the consequences of a Poisson Process model of earthquakes is that it predicts the number of earthquakes per year follows a Poisson distribution. These are count data, so a Chi-square test seems in order. The problem is this:

a Poisson distribution has infinitely many bins or possible outcomes,

so how do we do a chi-square test, which requires only finitely many categories?
Clearly we have to combine some categories. A rule-of-thumb is that the expected number \(np_i\) in each category should be at least five. Even so, the number and size of the bins leaves room for discretion. Each binning rule leads to a different test statistic with different numbers of degrees of freedom, so the test results may vary. Just remember, *Statistics means never having to say you’re certain.*

So the last bin should be of the form “Number of earthquakes per period is \(\geq n\)” The probability to assign to this bin is given by the Poisson(\(\mu\)) probability

\[
p_{\geq n} = \sum_{k=n}^{\infty} e^{-\mu} \frac{\mu^k}{k!} = 1 - \sum_{k=0}^{n-1} e^{-\mu} \frac{\mu^k}{k!}.
\]

### 21.8 Testing independence

We assumed in testing difference of means (\(t\)-test) that the variables were independent. But we can test this.

Given pairs of observations \((X_k, Y_k), k = 1, \ldots, n\), we can ask are \(X\) and \(Y\) independent?

If the data are already categorical, fine, but if not, we **bin** the data. That means break up the continuous variable into convenient ranges.

We now create a **contingency table**, in which columns correspond to the \(X\) values, and rows to the \(Y\) values. In cell \(i, j\) we put the number \(N_{i,j}\) of observations \(k\) with \(y_k = i\) and \(x_k = j\).

For example, let’s go back to Mendel’s data, where \(Y\) takes on values in \(\{a, r\}\) and \(X\) takes on values in \(\{g, y\}\). The contingency table is:

<table>
<thead>
<tr>
<th></th>
<th>g</th>
<th>y</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>32</td>
<td>101</td>
<td>133</td>
</tr>
<tr>
<td>r</td>
<td>108</td>
<td>315</td>
<td>423</td>
</tr>
<tr>
<td>Total</td>
<td>140</td>
<td>416</td>
<td>556</td>
</tr>
</tbody>
</table>

These data are special because each variable takes on exactly two values. The same methodology applies even if the number of rows and columns are different.

We now compute the relative frequency of each row and each column.

- row \(a\) has relative frequency \(133/556 = 0.239\)
- row \(b\) has relative frequency \(423/556 = 0.761\)
- col \(g\) has relative frequency \(140/556 = 0.252\)
- col \(y\) has relative frequency \(416/556 = 0.648\)

If \(X\) and \(Y\) are independent, then the relative frequency of each cell should be its row frequency times its column frequency.

<table>
<thead>
<tr>
<th></th>
<th>g</th>
<th>y</th>
<th>(\hat{p})</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.060</td>
<td>0.179</td>
<td>0.239</td>
</tr>
<tr>
<td>r</td>
<td>0.192</td>
<td>0.569</td>
<td>0.761</td>
</tr>
<tr>
<td>(\hat{p})</td>
<td>0.252</td>
<td>0.648</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Multiplying these by \( n = 556 \) gives the expected cell occupancy.

\[
\begin{array}{cc}
  g & y \\
  a & 33.36 & 99.52 \\
  r & 106.75 & 316.36 \\
\end{array}
\]

We can use this last table and the first as the basis for a \( \chi^2 \) test with 3 degrees of freedom. First compute the difference between the observed and expected cell counts:

\[
\begin{array}{cc}
  g & y \\
  a & -1.36 & 1.48 \\
  r & 1.25 & -1.36 \\
\end{array}
\]

Square them:

\[
\begin{array}{cc}
  g & y \\
  a & 1.8496 & 2.1904 \\
  r & 1.5625 & 1.8496 \\
\end{array}
\]

Divide each by its expected occupancy:

\[
\begin{array}{cc}
  g & y \\
  a & 0.0554436 & 0.0220096 \\
  r & 0.014637 & 0.0058465 \\
\end{array}
\]

Sum them to get the test statistic

\[ D = 0.0979368. \]

How many degrees of freedom does \( D \) have? There are four cells, but we estimated two parameters (the probability of row a and of column g) so the degrees of freedom are \( 4 - 1 - 2 = 1 \). (Cf. Theorem 10.5.1, part b in Larsen and Marx [8, p. 522].)

When testing the independence of \( r \) rows and \( c \) columns with estimated probabilities, the number of degrees of freedom is

\[ rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1). \]

The \( p \)-value of this statistic for 1 degree of freedom is 0.24568. This means that a \( \chi^2 \) random variable with 1 degree of freedom will be smaller than the value of test statistic with probability of about 0.75.

Note that this test is not the same as the test we performed in Example 21.6.4.

If the null hypothesis is \( H_0: X \) and \( Y \) are independent, we should use a one sided \( \chi^2 \) test. At the 5% level of significance we should reject \( H_0 \) if the test statistic \( D \) exceeds the critical value \( \chi^2_{.95,1} = 3.84 \). Since it does not, we fail to reject the null hypothesis.

For your convenience, I have appended the Mathematica code that I used for these calculations.
21.9 Simpson’s paradox

P. J. Bickel, E. A. Hammel, and J. W. O’Connell [2] examined the claim that the University of California’s graduate school discriminated against women.

In the fall of 1973, UC Berkeley received approximately 12,763 completed applications for admission to the graduate school, of which 8,442 were from men, and 4,321 were from women. About 44% of the men and 35% of the women were admitted.

Here is the contingency table for the null hypothesis that the admissions status is stochastically independent of gender.

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Expected</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admit</td>
<td>Deny</td>
<td>Admit</td>
</tr>
<tr>
<td>Men</td>
<td>3738</td>
<td>4704</td>
<td>4981.3</td>
</tr>
<tr>
<td>Women</td>
<td>1494</td>
<td>2817</td>
<td>1771.3</td>
</tr>
</tbody>
</table>

The $\chi^2$-test statistic has 1 df, and a value of 110.8, for a $p$-value of 0 (according to both R and Mathematica).

Does this mean that UC discriminates against women? There are two assumptions that went into making the contingency table. 1. Male and female applicants are equally qualified on average. 2. The ratio of men to women applicants is the same in each of the 101 departments.

It turns out that the second assumption is violated. Women were “overrepresented” in applications to more selective programs, and “underrepresented” in the applicant pool to less selective programs. [Selectivity is measured by admits/applicants. Highly selective programs have a lower ratio. It may surprise you to learn that by this measure, the STEM departments tended to be less selective.] Five of the departments had only one applicant. Of the remaining 96, a more elaborate $\chi^2$-test showed that women were more likely to be admitted than men, at a $p$-value of 0.0016.

This phenomenon is known as Simpson’s Paradox. Here is a highly artificial, but simple and transparent example.

21.9.1 Example The University has two departments. Department A has 6 spots and Department B has 99. There are 140 male and 140 female applicants to the U. Here are the admissions data.

<table>
<thead>
<tr>
<th>University</th>
<th>Department A</th>
<th>Department B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>Women</td>
<td>Men</td>
</tr>
<tr>
<td>Applied</td>
<td>140</td>
<td>20</td>
</tr>
<tr>
<td>Admitted</td>
<td>61</td>
<td>39</td>
</tr>
<tr>
<td>Rate</td>
<td>43.6%</td>
<td>5.0%</td>
</tr>
</tbody>
</table>

University-wide, men are admitted at a higher rate than women, yet the admission rate is higher for women in each department. The women are more likely to apply to the more selective department (Department A).  

---

Mathematica reports that it calculated to 307 places to the right of the decimal point.
21.10 Minimum Chi-square estimators

We can invert the test to get an estimator. If \( p \) depends on parameter vector \( \theta \), we can choose \( \hat{\theta} \) to minimize the test statistic, which by (1) is equivalent to

\[
\hat{\theta} \text{ minimizes } \sum_{i=1}^{t} \frac{X_i^2}{\mu_i(\theta)}.
\]

This estimator is one of the ones used by Mosteller’s [9] analysis of the World Series.

21.11 Minimum Chi-square estimation and the World Series

Here is the function to minimized based on the 106-game sample of best-of-seven World Series.

\[
D(p) = \frac{21^2}{(3)(p^4(1-p)^0 + (1-p)^4p^0)} + \frac{24^2}{(4)(p^4(1-p)^1 + (1-p)^4p^1)} + \frac{24^2}{(5)(p^4(1-p)^2 + (1-p)^4p^2)} + \frac{37^2}{(6)(p^4(1-p)^3 + (1-p)^4p^3)}
\]

The minimum \( \chi^2 \) estimate of \( p \) is 0.598853, which is remarkable agreement with the MLE of 0.59938, and the method of moments estimate of 0.58622.

Some Mathematica code

Here is the code I used for doing the contingency table analysis in section 21.8 “by hand.”

```mathematica
ct = {{32, 101}, {108, 315}}

nrows = Length[ct]
ncols = Length[Transpose[ct]]
df = (nrows - 1) (ncols - 1)

size = Total[Flatten[ct]]
colsums = Total[ct]
rowsums = Total[Transpose[ct]]
colfreqs = Round[colsums/size, .001]
```
rowfreqs = Round[rowsums/size, .001]

productfreqs = Round[Outer[Times, rowfreqs, colfreqs], .001]

expected = Round[size * productfreqs, .01]

diff = ct - expected
sqdiff = diff * diff
chisqsummands = sqdiff / expected

teststatistic = Total[Flatten[chisqsummands]]
pvalue = 1 - CDF[ChiSquareDistribution[df], teststatistic]
criticalvalue = InverseCDF[ChiSquareDistribution[df], 0.95]

Bibliography


[11] K. Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* 50(302):157–175. DOI: 10.1080/14786440009463897


