Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2021

Lecture 23: Specification Tests

Relevant textbook passages:

Larsen–Marx [22]: Sections 10.3, 10.4.

23.1 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 θ, we are interested in tests concerning the *function* f. These kinds of tests are usually referred to as **specification tests**.

For instance, as a Southern Californian, I am interested in whether earthquakes follow a Poisson process. If so, the time between main earthquake shocks follows an Exponential(λ) distribution for some λ. Since the exponential distribution is memoryless, the fact that we have not had a major earthquake on the San Andreas fault since 1906 does not mean that we are "overdue" for a major earthquake. But if the distribution is not exponential, we might be overdue for a major earthquake, in which case I would have to move. (Imagine the consequences of an earthquake that would cut off the water supply to and the exit routes from Los Angeles county.) So it is very important for my mental health to have evidence that earthquakes follow a Poisson process. In fact, part of your homework is to figure this out. I'll give you a hint: I still live here.

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(λ) 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 this 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 still 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 are just hypothesis tests. A particularly useful such test was characterized by Karl Pearson[1] in 1900 [28], 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.

[1] Karl Pearson is not the Pearson of the Neyman–Pearson Theorem. That Pearson is Egon Pearson, Karl's son.

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 with continuous data.

23.2 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 [22]. But you can find it discussed by Breiman [8, pp. 213–217], van der Waerden [39, § 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 7 that given independent and identically distributed random variables $X_1, \dots, X_n, \dots$, the **empirical distribution function** is defined by

$$F_n(x) = \frac{\#\{i : i \leqslant n \ \&\ X_i \leqslant x\}}{n},$$

or in terms of indicator functions

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

The Glivenko–Cantelli Theorem 7.11.3 asserts that if F is the common cumulative distribution function of the X_i’s, then

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

This suggests the following test statistic:

$$K(x_1, \dots, x_n) = \sup_{x\in\boldsymbol{R}} |F_n(x) - F(x)|.$$

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 must come at one of the jumps in F_n. Thus we only need to check $|F_n(x_i) - F(x_i)|$ and $|F_n(x_{i-1}) - F(x_i)|$ for $i = 1, \dots, x_n$. 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 12.1.2) 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,

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

So the procedure to use is this:

- Formulate a Null Hypothesis,

$$H_0\colon \text{the cumulative distribution function } F \text{ of } X_i \text{ is } F_0.$$

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

- 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 cumulative distribution function F_U of a Uniform is $F_U(y) = y$ for $0 \leqslant y \leqslant 1$.

- Compute the test statistic

$$K = \sup_{0 \leqslant y \leqslant 1} |G_n(y) - y| = \sup_x |F_n(x) - F_0(x)|,$$

where G_n is the empirical cumulative distribution function of the Y_is:

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

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 points at which the empirical cumulative distribution function jumps are actually the order statistics of $Y_1, \dots, Y_n$ (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 it is manageable. Birnbaum and Tingey [7] derive the following expression

$$P\Big(\sup_y G_n(y) - y > \varepsilon\Big) = \varepsilon \sum_{k=0}^{K} \binom{n}{k} \big(\varepsilon + (k/n)\big)^{k-1} \big(1 - \varepsilon - (k/n)\big)^{n-k}, \quad \text{where } K = \lfloor n(1-\varepsilon) \rfloor.$$

(See van der Waerden [39, pp. 67–75], esp. p. 73, or Feller [15].) Smirnov [35, 36] derived the distribution and came up with a very good approximation that allows us to compute the critical values for one-sided and two-sided tests. (The one-sided test tests the hypothesis that the distribution F satisfies $F(x) \geqslant F_0(x)$, or $G(y) \geqslant y$. The two-sided test is often more interesting and tests $F \neq F_0$.) Under the null hypothesis (G is uniform),

Add a derivation. It's not heinous, given the discussion of order statistics.

$$P\Big(\max_x G(x) - x > \varepsilon\Big) \approx e^{-2n\varepsilon^2},$$

where n is the sample size. (See van der Waerden [39, p. 74].) Thus to for a one-sided test with significance level α with "large" n ($n \geqslant 50$), reject if $K > K_\alpha$, where

$$\alpha = e^{-2nK_\alpha^2} \quad \text{or} \quad K_\alpha = \sqrt{\frac{-\ln \alpha}{2n}}.$$

The two-sided test uses $K_\alpha = \sqrt{\frac{-\ln(\alpha/2)}{2n}}$ for large n. Birnbaum and Tingey [7] have computed exact values for small n, and they may be found for instance in van der Waerden [39, Tables 4 and 5, pp. 344–345] or Breiman [8, p. 212]. I have included them in Table 23.1. The null hypothesis is rejected if K is larger than the cutoff.

But nowadays, you don't need the table. With R, use the `ks.test` command. (See the documentation or Dytham [14, pp. 86–89].) With Mathematica, use the `KolmogorovSmirnovTest` command.

23.2.1 Caveats

Here are some points to keep in mind:

	One-sided		Two-sided	
n	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
5	0.5094	0.6721	0.5633	0.6685
8	0.4096	0.5065		
10	0.3687	0.4566	0.4087	0.4864
15			0.3375	0.4042
20	0.2647	0.3285	0.2939	0.3524
25			0.2639	0.3165
30			0.2417	0.2898
40	0.1891	0.2350	0.2101	0.2521
50	0.1696	0.2107	0.1884	0.2260
60			0.1723	0.2067
70			0.1597	0.1917
80			0.1496	0.1795
90			0.1412	
100			0.1340	
large n	$1.22/\sqrt{n}$	$1.52/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Table 23.1. Critical values for one- and two-sided KS tests. Source: van der Waerden [39, Tables 4 and 5, pp. 344–345]

- The Kolmogorov–Smirnov test is designed to test the null hypothesis $F = F_0$, where F_0 is particular distribution. Often you want to test a composite hypothesis $F \in \Theta_0$. For example, you want to test whether waiting times follow *some* exponential distribution, so $\Theta_0 = \{F : \text{for some } \lambda > 0,\ F(t) = 1 - e^{-\lambda t}\}$. So you first estimate a θ_0, typically by MLE, and test the simple hypothesis $F = F(\theta_0)$. Unfortunately, as Breiman [8, p. 213] points out, "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 θ was estimated."

A typical composite null hypothesis is that $F \sim N(\mu, \sigma^2)$ or $F \sim \text{Exponential}(\lambda)$, where μ, σ, or λ are unknown and have to be estimated. Critical values for these special cases may be found in Pearson and Hartley [27, pp. 117–123, and Tables 54, 55, pp. 359–363].

- The Kolmogorov–Smirnov test is not very powerful, and the power is hard to estimate, but see Birnbaum [6] for some lower bounds.

- If the Kolmogorov–Smirnov test does reject the Null Hypothesis, the Q-Q graph of the quantiles provide useful insights in to 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[2, 3] test statistic is also a measure of the distance of the empirical cdf from the null cdf, but it emphasizes the tails. It is given by

$$A = n \int_{-\infty}^{\infty} \frac{\big(F_n(x) - F_0(x)\big)^2}{F(x)\big(1 - F(x)\big)} f_0(x)\, dx.$$

It can be used with transformed data to get a statistic whose distribution is independent of F_0.

○ The Shapiro–Wilk [33] test is based on order statistics, rather than the empirical cumulative distribution function. It is expressly designed to test whether the data are Normally distributed.

Razali and Wah [32] and Stephens [37] discuss a number of tests and provide references to many others. Razali and Wah argue on the basis of Monte Carlo studies that the Shapiro–Wilk test is more powerful for testing Normality.

23.2.2 The Kolmogorov–Smirnov Test in Mathematica and R

There are some anomalies in the `KolmogorovSmirnovTest` command in Mathematica (at least versions 8 through 12.1). According to the documentation, "`KolmogorovSmirnovTest[data]` tests whether `data` is normally distributed using the Kolmogorov–Smirnov test," where `data` is a list of numbers. The documentation also says that "`KolmogorovSmirnovTest[data, dist]` tests whether `data` is distributed according to `dist` using the Kolmogorov–Smirnov test." So let's check it out:

```
Clear[data1, data2];
SeedRandom[31415];

data1 = RandomVariate[NormalDistribution[0, 1], 100];

KolmogorovSmirnovTest[data1]
KolmogorovSmirnovTest[data1, NormalDistribution[0, 1]]

data2 = RandomVariate[NormalDistribution[0, 1], 100];

KolmogorovSmirnovTest[data2]
KolmogorovSmirnovTest[data2, NormalDistribution[0, 1]]
```

This clears the variables and seeds the random number generator, so that you and I should be able to replicate this result. The `RandomVariate` command creates a random sample of size 100 from a standard normal distribution. According to the documentation, the first `KolmogorovSmirnovTest` command test whether the data are "normal distributed, " and the second `KolmogorovSmirnovTest` test whether it is distributed according to a standard normal distribution. Here are my results.

```
0.494555

0.968586

0.788053

0.597322
```

The output of the `KolmogorovSmirnovTest` command is the p-values of the test.

Clearly something is wrong. If the first test is just testing whether the sample fits some normal distribution, it could be that it would by chance fit a non-standard normal better than a standard normal, so a higher p-valued would be expected. This is the case with the second sample, but not the first.

So let's test the same data with R. Perhaps it will shed some light on what is going on. You can export the data from Mathematica to file like this:

```
Export[NotebookDirectory[] <> "KSTest1.txt", data1]
Export[NotebookDirectory[] <> "KSTest2.txt", data2]
```

and then import the data into R like this:

```
data1 = scan("KSTest1.txt")
data2 = scan("KSTest2.txt")
```

provided you're in the right working directory. Then `ks.test(data1, pnorm)` returns

```
One-sample Kolmogorov-Smirnov test

data:  data1
D = 0.0477, p-value = 0.9768
alternative hypothesis: two-sided
```

and `ks.test(data2, pnorm)` returns

```
One-sample Kolmogorov-Smirnov test

data:  data2
D = 0.0752, p-value = 0.6241
alternative hypothesis: two-sided
```

It seems that Mathematica's `KolmogorovSmirnovTest[data, NormalDistribution[0, 1]]` and R's `ks.test(data, pnorm)` commands do the same thing. Indeed, using

```
h1 = KolmogorovSmirnovTest[data1, NormalDistribution[0, 1],
  "HypothesisTestData"]
h1["TestDataTable"]
```

shows that Mathematica and R use the same value of the test statistic. If you're worried that the *p*-values are slightly different, Mathematica and R use different computational techniques for the complicated distributions involved. But if you use

```
ks.test(data1, pnorm, exact = TRUE)
```

then R returns

```
One-sample Kolmogorov-Smirnov test

data:  data1
D = 0.0477, p-value = 0.9686
alternative hypothesis: two-sided
```

which is in better agreement with Mathematica.

If you want to test the hypothesis that your data are normally distributed, but you are not sure it is a standard normal, then you should use the estimated mean and standard deviation in your KS test. For Mathematica:

```
m = Mean[data]
s = StandardDeviation[data]

h = KolmogorovSmirnovTest[data, NormalDistribution[m, s],
  "HypothesisTestData"]
h["TestDataTable"]
```

or for R:

```
m <- mean(data)
s <- sd(data)
ks.test(data, pnorm(m, s))
```

In my code testing, these two approaches give the same result.

N.B. I do not know what the Mathematica `KolmogorovSmirnovTest[data]` command actually does, but I know it gives the wrong output in some cases, so I recommend avoiding that form and use the `KolmogorovSmirnovTest[data, NormalDistribution[m, s]]` form instead.

23.3 Review of the multinomial distribution

Pitman [31]: p. 155
Larsen–Marx [22]: pp. 494–495

The **multinomial distribution** generalizes the binomial distribution to random experiments with more than two types of outcomes or results or categories. Let there be K categories. Assume category k has probability p_k. Let $X_k = n_k$ be the number of occurrences of category k in $n = n_1 + \cdots + n_K$ independent trials. Then

$$p_X(n_1, \dots, n_K) = \frac{n!}{n_1! \cdot n_2! \cdots n_K!} p_1^{n_1} \cdot p_2^{n_2} \cdots p_K^{n_K} \qquad \left(\sum_{k=1}^{K} n_k = n \right).$$

It is easy to see that each X_k is Binomial(n, p_k) [22, Theorem 10.2.2, p. 496], and so

$$\boldsymbol{E}\, X_k = np_k.$$
$$\boldsymbol{Var}\, X_k = np_k(1 - p_k).$$

But the X_k's are not independent, since they must sum to n.

23.4 "Goodness of fit" tests

Larsen–Marx [22]: Section 10.3, pp. 499–509

When we have a multinomial model with K categories, our null hypothesis often takes the form

$$H_0 \colon \boldsymbol{p} = \boldsymbol{p}^0$$

where $\boldsymbol{p}^0$ is a vector of K probabilities that sum to one. The alternative is typically

$$H_1 \colon \boldsymbol{p} \neq \boldsymbol{p}^0.$$

Karl Pearson [28] proposed the following test statistic for this kind of test,

$$D = \sum_{k=1}^{K} \frac{(n_k - np_k^0)^2}{np_k^0} = \sum_{k=1}^{K} \frac{n_k^2}{np_k^0} - n, \tag{1}$$

or the "sum of squares of (observed − expected) over the expected." What does the distribution of this test statistic look like?

23.4.1 Theorem (The χ^2 Test) *Under the null hypothesis, the distribution of the test statistic D is approximately $\chi^2(K-1)$.*

- According to Theorem 10.3.1 in Larsen and Marx [22] we need $np_k^0 \geqslant 5$, for all $k = 1, \dots, K$ to use this approximation, but Cochran [12], van der Waerden [39, p. 238], and others think this is too conservative. According to Cochran [12, p. 328], "it is customary to recommend, in applications of the test, that the smallest expected number in any class should be 10 or (with some writers) 5. ... The numbers 10 and 5 appear to be arbitrarily chosen." He argues, citing [11], that "there is little disturbance at the 5% level when a *single* expectation is as low as $\frac{1}{2}$."

- Yates [40] points out that the expected number in each cell np_k^0 is rarely an integer, so that the deviation $n_k - np_k^0$ is an overestimate of departure from the model. He shows by comparison with exact probabilities (see Section 23.13⋆ below) that subtracting 1/2 from the absolute deviation improves the hypothesis test. That is, he recommends the use of

$$D' = \sum_{k=1}^{K} \frac{\left(|n_k - np_k^0| - \frac{1}{2} \right)^2}{np_k^0}$$

as the test statistic. This is called the **continuity correction** for the χ^2-test.

23.5 ⋆ Why the χ^2 test works

The following outline of a proof is based on Breiman [8, pp. 187–195] and Cramér [13, pp. 416–419]. For more details, see also van der Waerden [39, § 27, pp. 113-118, § 49, pp. 197–202, and § 51, pp. 207–211]. Cochran [12] also presents a readable exposition.

Sketch of the proof: We start by rewriting $\boldsymbol{X} = (X_1, \dots, X_K)$, the vector of counts by category, in terms of a sum of vectors of indicators. For each of the $i = 1, \dots, n$ independent experiments, and each of the $k = 1, \dots, K$ categories of outcome, let

$$\mathbf{1}_{i,k} = \begin{cases} 1 & \text{if the outcome of experiment } i \text{ is of category } k \\ 0 & \text{otherwise.} \end{cases}$$

And let $\mathbf{1}_i = (\mathbf{1}_{i,1}, \dots, \mathbf{1}_{i,K}) \in \boldsymbol{R}^K$. The vectors $\mathbf{1}_1, \dots, \mathbf{1}_n$ are independent, but within each vector, the components are decidedly not independent, as exactly one is nonzero. In terms of our original counts $\boldsymbol{X}$, we have

$$\boldsymbol{X} = \sum_{i=1}^{n} \mathbf{1}_i.$$

Now define $\boldsymbol{Y} = (Y_1, \dots, Y_K)$ by

$$\boldsymbol{Y} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{1}_i - \boldsymbol{p}) = \frac{1}{\sqrt{n}} (\boldsymbol{X} - n\boldsymbol{p}).$$

Technically, the vector $\boldsymbol{p}$ in the expression above is the vector $\boldsymbol{p}^0$ in the null hypothesis, but that pesky ^{0} just creates visual noise, and we'll omit it. At this point, I will just assert that by the Multivariate Central Limit Theorem 11.3.1, $\boldsymbol{Y}$ is approximately a jointly Normal random vector. So from here on, I will treat $\boldsymbol{Y}$ as if it were truly jointly Normal.

Note that each component of $\boldsymbol{Y}$ is given by

$$Y_k = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{1}_{i,k} - p_k) = \frac{X_k - np_k}{\sqrt{n}},$$

so

$$\sum_{k=1}^{K} Y_k = 0, \tag{2}$$

$$\boldsymbol{E}\, Y_k = 0, \quad (k = 1, \dots, K).$$

Since $\boldsymbol{Y}$ is a sum of the n independent and identically distributed random vectors $(\mathbf{1}_i - \boldsymbol{p})/\sqrt{n}$, the covariance matrix of $\boldsymbol{Y}$ is the same as the covariance matrix of each random vector $\mathbf{1}_i - \boldsymbol{p}$,

$$\boldsymbol{E}(Y_k Y_j) = \boldsymbol{E}(\mathbf{1}_{1,k} - p_k)(\mathbf{1}_{1,j} - p_j) = \begin{cases} p_k(1 - p_k), & k = j \\ -p_k p_j, & k \neq j. \end{cases}$$

(This is because $\mathbf{1}_{1,k}\mathbf{1}_{1,j} = 0$ whenever $j \neq k$ [the outcome can't be of both category j and category k] and $\mathbf{1}_{1,k}\mathbf{1}_{1,k} = 1$ with probability p_k under the null hypothesis.) Let

$$\boldsymbol{v} = (\sqrt{p_1}, \dots, \sqrt{p_K}), \tag{3}$$

where we treat $\boldsymbol{v}$ as a $K \times 1$ column vector. Note that $\boldsymbol{v}'\boldsymbol{v} = 1$, and $\boldsymbol{v}\boldsymbol{v}'$ is the $K \times K$ matrix

$$\boldsymbol{v}\boldsymbol{v}' = \begin{bmatrix} p_1 & \sqrt{p_1 p_2} & \sqrt{p_1 p_3} & \cdots & \sqrt{p_1 p_K} \\ \sqrt{p_2 p_1} & p_2 & \sqrt{p_2 p_3} & \cdots & \sqrt{p_2 p_K} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \sqrt{p_{K-1} p_K} \\ \sqrt{p_K p_1} & \cdots & \cdots & \sqrt{p_K p_{K-1}} & p_K \end{bmatrix}.$$

Next define the random variables W_k, $k = 1, \dots, K$ by

$$W_k = \frac{X_k - np_k}{\sqrt{np_k}} = \frac{Y_k}{\sqrt{p_k}}. \tag{4}$$

The vector $\boldsymbol{W} = (W_1, \dots, W_K)$ is jointly Normal since $\boldsymbol{Y}$ is. The covariance matrix of $\boldsymbol{W}$ is easily derived from that of $\boldsymbol{Y}$, as

$$\boldsymbol{E}\, W_k^2 = \frac{1}{p_k} \boldsymbol{E}\, Y_k^2 = 1 - p_k, \qquad \text{and} \qquad \boldsymbol{E}\, W_k W_j = -\frac{1}{\sqrt{p_k p_j}} p_k p_j = -\sqrt{p_k p_j}.$$

Thus

$$\boldsymbol{Var}\, \boldsymbol{W} = \boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}'.$$

Also

$$\sum_{k=1}^K W_k^2 = \sum_{k=1}^K \frac{Y_k^2}{p_k} = \sum_{k=1}^K \frac{(X_k - np_k)^2}{np_k} = D,$$

is the value of the test statistic. It is also the sum of squares of (nonindependent, approximately) normal random variables. We have cleverly set this up so that we can use an orthogonal transformation to transform D into a sum of $K - 1$ *independent* standard Normals, similar to the technique we use to prove the independence of the estimate of the mean and standard deviation for the multivariate Normal in Corollary 11.5.2.

So create a $K \times K$ orthogonal matrix $\boldsymbol{B}$ that has $\boldsymbol{v}'$ as its last row. We can always do this using the Gramm–Schmidt procedure. Define the transformed variables

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{W},$$

where $\boldsymbol{W}$ is treated as a column matrix. Since $\boldsymbol{W}$ is jointly Normal, therefore so is $\boldsymbol{Z}$. By Proposition S2.5.2, multiplication by $\boldsymbol{B}$ preserves inner products, so

$$\sum_{k=1}^K Z_k^2 = \sum_{k=1}^K W_k^2. \tag{5}$$

But Z_K is the dot product of the last row of $\boldsymbol{B}$ with $\boldsymbol{W}$, or

$$Z_K = \boldsymbol{v} \cdot \boldsymbol{W} = \sum_{k=1}^K \sqrt{p_k} W_k = \sum_{k=1}^K Y_k = \frac{1}{\sqrt{n}} \sum_{k=1}^K (X_k - np_k) = 0. \tag{6}$$

So combining (5) and (6), we have

$$\sum_{k=1}^{K-1} Z_k^2 = \sum_{k=1}^K W_k^2 = D. \tag{7}$$

By Proposition 11.1.3, the covariance matrices satisfy

$$\boldsymbol{Var}\, \boldsymbol{Z} = \boldsymbol{B}(\boldsymbol{Var}\, \boldsymbol{W})\boldsymbol{B}' = \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}')\boldsymbol{B}' = I - \boldsymbol{B}\boldsymbol{v}\boldsymbol{v}'\boldsymbol{B}'. \tag{8}$$

(Since $\boldsymbol{B}$ is orthogonal, $\boldsymbol{B}\boldsymbol{I}\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{B}' = \boldsymbol{I}$).

Now examine the matrix $\boldsymbol{B}\boldsymbol{v}\boldsymbol{v}'\boldsymbol{B}' = (\boldsymbol{B}\boldsymbol{v})(\boldsymbol{B}\boldsymbol{v})'$ that appears in (8). By construction, the last row of $\boldsymbol{B}$ is $\boldsymbol{v}'$, and the other rows of $\boldsymbol{B}$ are orthogonal to $\boldsymbol{v}$. Thus $\boldsymbol{B}\boldsymbol{v}$ is a K-column vector of zeroes except for the last entry, which is 1. So $\boldsymbol{B}\boldsymbol{v}\boldsymbol{v}'\boldsymbol{B}'$ is a $K \times K$ matrix of zeroes, except for the K, K entry, which is 1. So by (8), the covariance matrix of $\boldsymbol{Z}$ has all of its entries equal to 0, except for the $K-1$ diagonal entries, $1, 1 \dots, K-1, K-1$, which are 1.

$$\boldsymbol{Var}\, \boldsymbol{Z} = \left[\begin{array}{ccccc|c} 1 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 1 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & & \ddots & 1 & 0 & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \\ \hline 0 & \cdots & \cdots & \cdots & 0 & 0 \end{array}\right]$$

As a consequence, the random vector $(Z_1, \dots, Z_{K-1})$ is a vector of independent standard Normal random variables. So by (7), D is a $\chi^2(K-1)$ random variable. ∎

23.6 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. For instance, we might want to know if the number of earthquakes in a year is governed by a Poisson(μ) distribution for some $\mu > 0$. In this case we first have to estimate μ from the data before we can calculate $p_k = e^{-\mu}\mu^k/k!$. It turns out the same test statistic can be used, but it has a different limiting distribution. But not too different. It is still a χ^2 distribution, but it has one less degree of freedom for each parameter we estimate.

That is, there is a vector $(\theta_1, \dots, \theta_m)$ of parameters that determine a multinomial probability vector $\boldsymbol{p}(\theta)$ and we wish to test the null hypothesis Let

$$H_0 \colon \boldsymbol{p} = \hat{\boldsymbol{p}},$$

where $\hat{\boldsymbol{p}}$ is the maximum likelihood estimate of $\boldsymbol{p}$, that is, $\boldsymbol{p}(\hat{\theta}_{\text{MLE}})$. We make the following technical assumptions: Let A be an interval in $\boldsymbol{R}^{\text{m}}$ and let $\boldsymbol{p} \colon A \to \boldsymbol{R}^{\text{K}}$, where $m < K$, satisfy

1. For each $\theta = (\theta_1, \dots, \theta_m) \in A$, $\boldsymbol{p}(\theta)$ is a probability vector in $\boldsymbol{R}^{\text{K}}$,
2. each p_i, $i = 1, \dots, K$, is uniformly bounded away from 0 on A.
3. $\boldsymbol{p}$ is twice continuously differentiable on A, and
4. the matrix $\left[\frac{\partial p_i}{\partial \theta_j}\right]$ has full rank m.

You can find a statement of the following theorem in Cramér [13, pp. 426–427], or without the technical conditions in Larsen–Marx [22, Theorem 10.4.1] or Breiman [8, Theorem 6.13, p. 196]. Cramér states the theorem for the minimum χ^2 estimator rather than the MLE estimator, but as Cochran [12] points out, they are asymptotically the same, so the theorem applies to any asymptotically efficient estimator.

23.6.1 Theorem (χ^2 test with estimated parameters) *Under the technical assumptions above, let $\hat{\boldsymbol{p}}$ be the MLE estimate of $\boldsymbol{p}$, where the MLE $\hat{\theta}_{\text{MLE}}$ is interior to A. Then under the null hypothesis*

$$H_0 \colon \boldsymbol{p} = \hat{\boldsymbol{p}},$$

the test statistic

$$D = \sum_{k=1}^{K} \frac{(X_k - n\hat{p}_k)^2}{n\hat{p}_k} = \sum_{k=1}^{K} \frac{X_k^2}{n\hat{p}_k} - n,$$

has an approximately Chi-square distribution with $K - 1 - m$ degrees of freedom.

Why do we lose a degree of freedom? The proof of this theorem is rather involved, but you can find a nine-page proof in Cramér [13, § 30.3, pp. 424–434]. See also Cochran [12, Section 4]. Basically, the first order conditions for a maximum with respect to an estimated parameter impose another linear restriction on the variables we are squaring in our test statistic. That is, the deviations from the estimated expectations are further constrained. Since the df represents the number of standard normals we are squaring, the critical values for a test with estimated probabilities will be smaller.

Note that with fewer degrees of freedom the critical value of the test becomes smaller. That is, we become tougher on the null hypothesis. This is not surprising. By estimating the vector $\boldsymbol{p}$, we are essentially choosing $\boldsymbol{p}$ to give the best fit, so the sum of squared deviations from the predictions should decrease. Thus we should set a lower threshold for rejection.

It is worth noting that the change in the degrees of freedom due to estimating parameters was not initially recognized. It was first pointed out by R. A. Fisher [16]. Karl Pearson, who invented the χ^2-test refused to accept Fisher's argument and wrote [29]:

> The above redescription of what seem to me very elementary considerations would be unnecessary had not a recent writer in the *Journal of the Royal Statistical Society* appeared to have wholly ignored them. He considers that I have made serious blunders in not limiting my degrees of freedom by the number of moments I have taken; for example he asserts (p.93) that if a frequency curve be fitted by the use of four moments then the n' of the tables of goodness of fit should be reduced by 4. I hold that such a view is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the *Journal of the Royal Statistical Society*.

I take comfort in the fact that even brilliant original thinkers can be wrong.

23.6.1 The Chi-square test and continuous distributions

Chernoff and Lehmann [9] raise an interesting question with regard to using a χ^2-test to test whether the data follow a continuous distribution. For example, to test whether $X_1, \dots, X_n$ are distributed according to a Normal(μ, σ^2) distribution, you can construct a set of K bins, and count the number of observations in each bin. Given μ and σ^2, the expected number in each bin can be calculated, and the resulting test statistic D has an approximate χ^2 distribution with $K-1$ degrees of freedom. There is still the question of how to choose the number of bins. This is much the same question as to how to choose the bins for a histogram, and several methods are discussed in Section 7.13 ⋆.

But when μ and σ^2 are unknown, there are two ways to estimate them. Suppose you know only the occupancies n_k of each bin k. Then you would use maximum likelihood or minimum chi-square (see Section 23.11 below) estimator to get an estimate $(\tilde{\mu}, \tilde{\sigma}^2)$ of the parameter, which you could use to calculate the expected occupancies of each cell. The resulting test statistic $\tilde{D}$ would have an approximate an approximate χ^2 distribution with $K-3$ degrees of freedom.

But suppose you knew all the values $x_1, \dots, x_n$. You could use these to get the standard MLE $(\hat{\mu}, \hat{\sigma}^2)$. This gives a test statistic $\hat{D}$ that has a slightly different distribution, between $\chi^2(K-3)$ and $\chi^2(K-1)$. Chernoff and Lehmann show that using this statistic with the cutoff for $\tilde{D}$ gives a larger probability of Type I error than $\tilde{D}$.

23.7 The grooviness of the Chi-square Test

23.7.1 Example (The World Series) World Series come in lengths of 4, 5, 6, and 7 games, so there are $K = 4$ categories of results of this experiment. For our binomial-based model, the probability of length ℓ is

$$H_0\colon p_\ell = \binom{\ell-1}{3}\left(p^4(1-p)^{\ell-4} + (1-p)^4 p^{\ell-4}\right) \qquad (\ell = 4, \dots, 7). \tag{9}$$

Thus if there are k_ℓ series of length ℓ in our sample, the test statistic

$$\sum_{\ell=4}^{7} \frac{(k_\ell - n\hat{p}_\ell)^2}{n\hat{p}_\ell}$$

Larsen–Marx [22]: Section 10.4

has a χ^2-distribution. Since we have to estimate $\hat{p}$, then our degrees of freedom are reduced by 1, so the test statistic has a $\chi^2(2)$-distribution $(2 = (4-1)-1)$. The critical value of the $\chi^2(2)$ distribution is 5.99 at the 5% level of significance. □

23.7.2 Example (Case Study 10.3.2 [22]: 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 [25]. It became known as **Benford's Law** because Frank Benford [4] independently rediscovered and then popularized it. He had investigated among other things: baseball statistics, surface areas of rivers, and molecular weights of chemicals [22, pp. 502–505]. There are also sets of numbers that disobey Benford's Law. Phone directories may disobey the law 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 [19] has developed a sophisticated probabilistic model that predicts Benford's Law, but it is beyond the scope of this course. See also [20] for an accessible discussion.

Here is the table of probabilities:

Digit d	$\log_{10}(d+1) - \log_{10}(d)$
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046

Notice that 30% of wild numbers start with 1!

Forensic accounting

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

Larsen and Marx [22, pp. 502–505] cite as an example, the University of West Florida's budget. The present counts of the leading digits and use a χ^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 χ^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 χ^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. For political reasons, their analysis has not yet been published, but preliminary indications are that much of the accounting data are not consistent with Benford's Law. □

23.7.3 Example (Did Mendel Cheat?) See Larsen–Marx [22, Case Study 10.3.3, 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 present 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.

Phenotype	Obs.	Mendel	Pred.
ry	315	9/16	312.75
rg	108	3/16	104.25
ay	101	3/16	104.25
ag	32	1/16	34.75

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. This means that even if the data are independent, the probability of getting a better fit than this is only 0.075. Or in other words, the p-value of D is about 0.925. This led R. A. Fisher to believe Mendel (or one of his minions) had faked his data. □

23.8 The χ^2 test and the Law of Small Numbers

23.8.1 Example (The cookie data) When I read Chung's [10, p. 196] claim that the number of raisins in cookies follows a Poisson distribution, I immediately wondered who would possibly have done the research to bear that out. Upon reflection I realized that Nabisco, or Famous Amos, or any number of large bakeries would have some incentive to monitor the number of raisins in their cookies. Nevertheless it seemed worth investigating, so every now and then I pass out Mother's brand chocolate chip cookies, and ask students to voluntarily email the results of their cookie dissections. (One of the virtues of the Law of Small Numbers is that it applies equally well to chocolate chips as raisins.)

Here are the data on the number of chocolate chips in the cookies:

# of chips	0	1	2	3	4	5	6	7	8	9	10	11	12
# of cookies	0	1	3	3	5	13	17	13	12	7	7	2	2

That is, 3 cookies had 2 chips, 13 cookies had 7 chips, etc. This makes a total of 569 chips in $n = 85$ cookies. If the number of chips per cookie does follow a Poisson(μ) distribution, then the MLE of μ is $569/85 = 6.7$. (See Example 19.1.3.)

Now the Law of Small Numbers is not about how many chips there are in each cookie, it's about how many cookies have a given number of chips. (You may want to go back to Section 13.6 and Proposition 13.6.1.) According to the Law of Small Numbers, the expected number of cookies having k chips is $np_\mu(k)$, where $p_\mu(k)$ is the Poisson probability

$$p_\mu(k) = e^{-\mu}\frac{\mu^k}{k!}.$$

Using the estimated $\hat{\mu}_{\text{MLE}}$ we can add a row for expected numbers of cookies. But there is a problem:

a Poisson distribution has infinitely many possible outcomes, so I shall add a category, 13 +, for cookies with thirteen or more chips. And how many cookies are expected to fall in that category? The answer is,

$$n - \sum_{k=0}^{12} np_\mu(k) = n\Big(1 - \sum_{k=0}^{12} p_\mu(k)\Big).$$

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13 +
n_k	0	1	3	3	5	13	17	13	12	7	7	2	2	0
$np_{\hat\mu}(k)$	0.1	0.7	2.4	5.3	8.8	11.8	13.2	12.6	10.5	7.8	5.2	3.2	1.8	1.7

Recall the remark after Theorem 23.4.1 that, as a rule, to apply the χ^2 test we should have an expected count of at least 5 in each category. Let's ignore that for now and compute and compute the test statistic

$$D = \sum_{k=0}^{12} \frac{\big(n_k - np_\mu(k)\big)^2}{np_\mu(k)} + (1.7)^2/1.7 = 7.3.$$

Since there are 14 categories, D has a χ^2 distribution with $12 = 14 - 1 - 1$ degrees of freedom (since we estimated μ by MLE). The critical value for a test at the 5% significance level is $\chi^2_{0.95,12} = 21.03$. (The p-value of D is 0.84.) Thus we do not reject the null hypothesis that the number of cookies is governed by the Law of Small Numbers.

Now in order to get at least 5 expected cookies in each category, I grouped them into bins as follows:

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Bin 8	Bin 9
# of chips	0–3	4	5	7	7	8	9	10	11 +
Actual # of cookies	7	5	13	17	13	12	7	7	4
Expected	8.4	8.8	11.8	13.2	12.6	10.5	7.8	5.2	6.6

Treating the bins as the units of analysis gives the test statistic $D = 5.1$. There are 9 bins and 1 estimated parameter, μ, so the test statistic is $\sim \chi^2(7)$. The p-value for this the value $D = 5.1$ is 0.64, so we decisively *fail* to reject the Poisson hypothesis. But you have to admit that the data are crumby. □

So to reiterate, when dealing with count data, where the counts are believed to have a Poisson distribution, you should group counts into bins, so that each bin has an expected number of at least 5 occupants. 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.*

Also the last bin should be of the form "$\geqslant m$." The probability to assign to this bin is given by the Poisson(μ) probability

$$p_{\geqslant m} = \sum_{k=m}^{\infty} e^{-\mu}\frac{\mu^k}{k!} = 1 - \sum_{k=0}^{m-1} e^{-\mu}\frac{\mu^k}{k!},$$

and its expected occupancy is probability times the sample size n, or equivalently, n minus the sum of the expected occupancy of all other bins.

23.9 Testing independence (categorical data)

We often assume that random variables are independent. But if the data are categorical, we can test this. Even if the data are not categorical, we can create finitely many **bins** for the values, and treat the data as categorical.

Given pairs of observations (x_t, y_t), $t = 1, \dots, n$, we can ask are X_t's and Y_t's independent? Or is the value of Y *contingent* on the value of X, or vice versa?

Larsen–Marx [22]: § 10.5

To answer this, we 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 (x_t, y_t) with $y_t = i$ and $x_t = j$.

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:

	g	y	Total
a	32	101	133
r	108	315	423
Total	140	416	556

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.

$$\begin{aligned} \text{row a has relative frequency } 133/556 &= 0.239 \\ \text{row b has relative frequency } 423/556 &= 0.761 \\ \text{col g has relative frequency } 140/556 &= 0.252 \\ \text{col y has relative frequency } 416/556 &= 0.748 \end{aligned}$$

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

	g	y	$\hat{p}$
a	0.060	0.179	0.239
r	0.192	0.569	0.761
$\hat{p}$	0.252	0.748	1.000

Multiplying these by $n = 556$ gives the expected cell occupancy.

	g	y
a	33.36	99.52
r	106.75	316.36

We can use this last table and the first as the basis for a χ^2 test. First compute the difference between the observed and expected cell counts:

	g	y
a	-1.36	1.48
r	1.25	-1.36

Square them:

	g	y
a	1.8496	2.1904
r	1.5625	1.8496

Divide each by its expected occupancy:

	g	y
a	0.0554436	0.0220096
r	0.014637	0.0058465

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 [22, p. 522].)

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

$$rc - 1 - (r-1) - (c-1) = (r-1)(c-1).$$

Note that this test is not the same as the test we performed in Example 23.7.3.

If the null hypothesis is H_0: X and Y are independent, we should use a one sided χ^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.

23.10 Simpson's paradox

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. This prompted P. J. Bickel, E. A. Hammel, and J. W. O'Connell [5] to examine whether the University of California's graduate school discriminated against women.

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

	Observed		Expected		Difference	
	Admit	Deny	Admit	Deny	Admit	Deny
Men	3738	4704	3460.7	4981.3	277.3	-277.3
Women	1494	2817	1771.3	2549.7	-277.3	277.3

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

Does this mean that UC discriminates against women? At UC, just as at Caltech, graduate admissions are not carried out the university level, but at the departmental level. UC had 101 departments, each overseeing its own admissions, so the question ought to be which, if any, of the departments are discriminating against women. As it turned out, women were more likely to apply to more selective programs, and less likely to apply 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, departmental χ^2-tests showed that women were *more* likely to be admitted than men, at a p-value of 0.0016.

By the way, this analysis does not show that women are not discriminated against in STEM fields. What it does show is that if there is discrimination, it operates earlier or later than the graduate admissions decision. This is often referred to as the "leaky pipeline" problem.

23.10.1 Remark By the way this brings up an important issue in applying the χ^2-test in the situation. The test is usually justified on the basis that the outcomes of each admissions

[2] Mathematica reports that it calculated the p-value to be zero to 307 places to the right of the decimal point.

decision are stochastically independent. That is, for each applicant there is a probability p of being admitted, and that we observe independent draws from this population. But anyone with experience in admissions knows that the probability of admission is not fixed, but rather depends on the number of applicants. Perhaps a better model would be to assume that the number of male and female applicants is give by a Poisson distribution, and then the probability of admission depend son the number of applicants and "slots." But then, conditioning on the value of the Poisson distribution leads to the same analysis that we have done. Cf. The discussion in Section 15.12.1.

The phenomenon that the aggregate average rates can be the reverse of the constituent rates, is known as **Simpson's Paradox** [34].[3] Here is an artificial, but simple and transparent example.

23.10.2 Example The University has two departments. Department A has 18 spots and Department B has 104. There are 370 male and 122 female applicants to the U. Here are the admissions data.

	University			Department A			Department B		
	Men	Women	Total	Men	Women	Total	Men	Women	Total
Apps	240	130	370	80	110	190	160	20	180
Admits	97	25	122	7	11	18	90	14	104
Rate	40.4%	19.2%	33.0%	8.8%	10.0%	9.5%	56.3%	70.0%	57.8%

Men are admitted at double the rate of women overall, but at a lower rate in each department. The women are more likely to apply to the more selective department (Department A). □

23.11 Minimum Chi-square estimators

We can stand the Chi-square test on its head to get an estimator. If $\boldsymbol{p}$ depends on parameter vector θ, we can choose $\hat{\theta}$ to minimize the test statistic, which by (1) is equivalent to

$$\hat{\theta} \text{ minimizes } \sum_{k=1}^{K} \frac{X_k^2}{p_k(\theta)}.$$

This estimator is one of the estimators considered by Mosteller's [24] analysis of the World Series.

[3] Simpson [34, p. 231] in 1951 gives an explicit example of this phenomenon in terms of contingency tables. Wikipedia points to an 1899 paper by Karl Pearson, at al. [30], where on pages 277–278, they describe "spurious correlation" at an aggregate level due to "heterogeneity" at various sublevels, which could be interpreted as a similar phenomenon. The same web page points out that Yule [41] in 1903 also gives an example of the paradox (pp. 132–133) in terms of independence. Yule also cites Pearson.

23.12 Minimum Chi-square estimation and the World Series

Here is the function to minimize based on the 111-game sample of best-of-seven World Series.

$$
\begin{aligned}
D(p) = \frac{21^2}{\binom{3}{3}\left(p^4(1-p)^0 + (1-p)^4 p^0\right)} \\
+ \frac{26^2}{\binom{4}{3}\left(p^4(1-p)^1 + (1-p)^4 p^1\right)} \\
+ \frac{24^2}{\binom{5}{3}\left(p^4(1-p)^2 + (1-p)^4 p^2\right)} \\
+ \frac{40^2}{\binom{6}{3}\left(p^4(1-p)^3 + (1-p)^4 p^3\right)}
\end{aligned}
$$

The minimum χ^2 estimate of p is 0.588295. This agrees well with the MLE of 0.593467, and the method of moments estimate of 0.574641.

Some Mathematica code

Here is the code I used for doing the contingency table analysis in section 23.9 “by hand.”

```
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]
```

23.13 ⋆ Fisher’s Exact Test

In 1934 R. A. Fisher communicated a new approach to analyzing 2 × 2 contingency tables to his associate Frank Yates [40], who published a lengthy analysis of this case when the expected cell values were small. Fisher included this in the 1934 edition of his *Statistical Methods for Research Workers*, see [17, pp. 96–97]. It is called an exact test, since it does not rely on the multivariate central limit theorem, and gives an exact p-value. The test is usually called **Fisher’s Exact Test**, but Lehmann [23, p. 146] refers to the test as the **Fisher–Irwin test**, since J. O. Irwin had the same result in 1933, but it was not published until 1935 [21]. It is also sometimes referred to as a *conditional* exact test, because the distribution of the test statistic is conditional on the values the margins.

One way to think about a 2×2 contingency table is this. We have independent samples of $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ of independent Bernoulli(p_X) and Bernoulli(p_Y) random variables. For instance, the X sample might be women applying to Berkeley, and the Y sample would be men. An outcome is a "success" if the applicant is admitted. We are interested in the null hypothesis that the X and Y samples are *homogenous* in that the probability of success (admission) is the same in each sample,

$$H_0 \colon p_X = p_Y.$$

against the alternative hypothesis that $p_X \neq p_Y$. We have already carried out such a test using the χ^2-test. But recall that the rule of thumb for using a χ^2-test is that the expected occupancy of each cell is $\geqslant 5$. Fisher, Irwin, and Yates were interested in small sample sizes where the cell occupancies do not meet that requirement.

We can summarize the data in a 2×2 contingency table:

	Successes	Failures	
Sample X	x	$m - x$	m
Sample Y	y	$n - y$	n
Combined	$x + y$	$m + n - x - y$	$m + n$

That is, there are $X = x$ successes in m trials from the X sample and $Y = y$ successes in n trials from the Y population.

So how do we test the hypothesis that $p_X = p_Y$, when we don't know either? Recall the MLE estimator of p_X is just x/m, and for p_Y it is y/n. Under the null hypothesis $p_X = p_Y = p_*$, the MLE estimator of p_* is $(x + y)/(m + n)$. Irwin [21, method 1, p. 84] argues that we should estimate p_* by maximum likelihood, so

$$\hat{p} = \frac{x + y}{m + n},$$

and then for each of the $(n+1) \cdot (m+1)$ tables with $0 \leqslant x \leqslant m$ and $0 \leqslant y \leqslant n$ use $\hat{p}$ to compute its probability. Then summing over all values (x, y) that give rise to tables less probable than the observed values, gives a p-value for a likelihood ratio test of the null hypothesis. He asserts that "[t]his procedure, however, would be impractical owing to the large number of tables to be enumerated."

Suppose for the moment that we observe $\bar{x}$ success from the X population and $\bar{y}$ successes from the Y population and that $\bar{x}/m < \bar{y}/n$. Then we could contemplate the alternative hypothesis

$$H_1 \colon p_X < p_Y.$$

The method (2) suggested by Irwin is a simple computation would be to compute the probability (using $\hat{p}$ of each table with $x \geqslant \bar{x}$ and $y \leqslant \bar{y}$, that is, those tables with more X success and fewer Y success than observed. That is, we look at table that "more extreme" than the observed table. This seems to fit with our usual notion of likelihood ratio testing. The problem is that in our previous discussion of likelihood ratio tests, in all/most of the examples we had the Monotone Likelihood Ratio Property. That is, there is a one-dimensional sufficient test statistic T that had the property that a "more extreme" value of T corresponds to a lower likelihood ratio. That is, $T' > T$ implies $L(\theta_0; T')/L(\theta_1; T') < L(\theta_0; T)/L(\theta_1; T)$. This then leads to a critical values such that we reject the null hypothesis $H_0 \colon \theta = \theta_0$ in favor of the alternative hypothesis $H_1 \colon \theta = \theta_1$ if $T > t^*$. The question here is what does it mean to be "more extreme?" Irwin [21] gives an example of a contingency table that is not more extreme than the observed data, but is less probable than the observed data, where the probabilities are computed used the MLE of the common parameter value based on the whole sample. That would seem to rule out method (2).

Need better notation here!!!

Method (3) in Irwin is **Fisher's exact test** or the **Fisher–Irwin** test. It makes use of the following observation. Under the null hypothesis $H_0\colon p_X = p_Y$, conditioning on $X+Y$, we have

$$P\big(X = x \bigm| X+Y = t\big) = \frac{\binom{m}{x}\binom{n}{t-x}}{\binom{m+n}{t}}. \tag{10}$$

The amazing thing about this is that it does not depend on the common probability of successYp_*.

To understand this, recall Section 3.10 on sampling without replacement. Think of m bins labeled X and n bins labeled Y. There a total of t balls labeled success and $m+n-t$ balls labeled "failure". The balls are placed at random into the bins, one per bin. The number of ways to get x balls marked success into the m bins labeled X is the same as the number of subsets of size x from a set of size m, namely $\binom{m}{x}$. Likewise there are $\binom{n}{t-x}$ ways to arrange the success in the Y bins. Overall there are $\binom{n+m}{t}$ ways to arrange the successes. Thus $P\big(X = x \bigm| X+Y = t\big)$ is given by the hypergeometric probability described in (10).

So to test of the null hypothesis $H_0\colon p_X = p_Y$ against the one-sided alternative $p_X < p_Y$, compute the conditional probability p of a more extreme (smaller) value of X,

$$\sum_{k=0}^{x} P(X = k \bigm| X+Y = t) = \sum_{k=0}^{x} \frac{\binom{m}{k}\binom{n}{t-k}}{\binom{m+n}{t}},$$

and reject the null hypothesis if this p-value is less than the significance level. Lehmann [23, p. 143] shows that this is in fact a likelihood ratio test. He also shows that the same arguments can be used to develop a test of the equality of two Poisson distributions.

The question arises as to why we should condition on the sum $X+Y$. The answer is that fixing $x+y$, which amount to fixing the marginals in the contingency table makes the notions of "more extreme" and "less probable" coincide again. But it may still be practical to use method (1) given today's technology.

Agresti [1] provides a discussion of the history and development of alternative exact tests, and has extensive references. There are still active discussions on statistics websites such as http://stats.stackexchange.com about the merits of each test.

You can use these ideas for an exact test of the multinomial distribution when the probabilities are given. Given the null hypothesis

$$H_0\colon \boldsymbol{p} = \boldsymbol{p}^0,$$

we can compute the probability of getting any vector $\boldsymbol{n} = (n_1, \dots, n_K)$ of observed outcomes by

$$p_X(n_1, \dots, n_K) = \frac{n!}{n_1! \cdot n_2! \cdots n_K!} p_1^{n_1} \cdot p_2^{n_2} \cdots p_K^{n_K}.$$

So for a two-sided alternative hypothesis $H_1\colon \boldsymbol{p} \neq \boldsymbol{p}^0$ we sum these probabilities for all vectors $\boldsymbol{n}'$ that are less probable under the null hypothesis than the observed vector of counts $\boldsymbol{n}$ to get the p-value of the test. You can imagine that if n and K are large, then there are a lot of such vectors $\boldsymbol{n}'$, see, e.g., [18]. In the 1930s this seemed like daunting task, but with today's technology it is often practical. (See, e.g., Patefield [26].) So why are these tests not used exclusively in practice? Partly because most statistics textbooks have ignored the computer revolution, so it is not as well known as the χ^2-test.

23.13.1 Example (Irwin's example) For the sake of completeness, here is Irwin's example of two contingency tables, one of which is less probable given the estimates from the first table, but is not more extreme.

	Successes	Failures	
Sample X	26	2	28
Sample Y	61	2	63
Combined	87	4	91

On the basis of these data, the probability of success estimated from the combined samples is $\hat{p} = 87/91 = 0.956$. The probability of getting exactly this arrangement under the mull hypothesis is

$$\binom{28}{26}\hat{p}^{26}\,(1-\hat{p})^{2}\cdot\binom{63}{61}\hat{p}^{61}\,(1-\hat{p})^{2} = 0.0528.$$

Now consider the alternative table

	Successes	Failures	
Sample X	3	25	28
Sample Y	51	12	63
Combined	54	37	91

The probability of getting this arrangement under the mull hypothesis is

$$\binom{28}{3}\hat{p}^{3}\,(1-\hat{p})^{25}\cdot\binom{63}{51}\hat{p}^{51}\,(1-\hat{p})^{12} = 4.8\times 10^{-36}.$$

But the second table is not more extreme than the first. In the first table, $X/m = 26/28 = 0.93$ and $Y/n = 61/63 = 0.97$, so a more extreme table would have a smaller X value and a larger Y values. The second table has $X = 3$, which is less than 26, but it also has $Y = 51$, which is also less than the first table's 61, so it is *not* more extreme than the first table.

While we are at it, let's use Irwin's method (1) on this example. There are $29 \times 64 = 1856$ possible tables to consider with sizes 28 and 83. That is large for 1930, but not for 2021.

Mathematica 11.2 reports after 0.03 seconds, that the method (1) probability of getting a table weakly less probable than the first under the null hypothesis is 0.57. Fisher's Exact Test, in this case a test versus the one-sided alternative $p_X \leqslant p_Y$, gives a p-value of

$$\sum_{x=0}^{26}\frac{\binom{28}{x}\binom{63}{87-x}}{\binom{91}{87}} = 0.36.$$

Note that in the sum above, I included $x = 26$, as Fisher recommends. If we omit that case, the probability drops to 0.085. If we were testing at the 10% level, this would indicate the need for randomization (see Cochran [12, p. 327]).

So not only is Irwin's method (1) practical in some cases, it can make a big difference. □

Bibliography

[1] A. Agresti. 1992. A survey of exact inference for contingency tables. *Statistical Science* 7(1):131–153. DOI: 10.1214/ss/1177011454

[2] T. W. Anderson and D. A. Darling. 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics* 23(2):193–212. http://www.jstor.org/stable/2236446

[3] ——— . 1954. A test of goodness of fit. *Journal of the American Statistical Association* 49(268):765–769. http://www.jstor.org/stable/2281537

[4] F. Benford. 1938. The law of anomalous numbers. *Proceedings of the American Philosophical Society* 78(4):551–572. http://www.jstor.org/stable/984802

[5] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187(4175):398–404. http://www.jstor.org/stable/pdfplus/1739581

[6] Z. W. Birnbaum. 1953. On the power of a one-sided test for continuous probability functions. *Annals of Mathematical Statistics* 24(3):484–489.
http://projecteuclid.org/euclid.aoms/1177728989

[7] Z. W. Birnbaum and F. H. Tingey. 1951. One-sided confidence contours for distribution functions. *Annals of Mathematical Statistics* 22(4):592–596.
http://projecteuclid.org/download/pdf_1/euclid.aoms/1177729550

[8] L. Breiman. 1973. *Statistics: With a view toward applications*. Boston: Houghton Mifflin Co.

[9] H. Chernoff and E. L. Lehmann. 1954. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics* 25(3):579–586.
http://www.jstor.org/stable/2236840

[10] K. L. Chung. 1979. *Elementary probability theory with stochastic processes*. Undergraduate Texts in Mathematics. New York, Heidelberg, and Berlin: Springer–Verlag.

[11] W. G. Cochran. 1942. The χ^2 continuity correction. *Iowa State College Journal of Science* 16:421–435.

[12] ——— . 1952. The χ^2 test of goodness of fit. *Annals of Mathematical Statistics* 23(3):315–345. http://www.jstor.org/stable/2236678

[13] H. Cramér. 1946. *Mathematical methods of statistics*. Number 34 in Princeton Mathematical Series. Princeton, New Jersey: Princeton University Press. Reprinted 1974.

[14] C. Dytham. 2011. *Choosing and using statistics: A biologist's guide*, 3d. ed. Wiley–Blackwell.

[15] W. Feller. 1948. On the Kolmogorov–Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics* 19(2):177–189.
http://projecteuclid.org/download/pdf_1/euclid.aoms/1177730243

[16] R. A. Fisher. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85(1):87–94.
http://www.jstor.org/stable/2340521

[17] ——— . 1970. *Statistical methods for research workers*, 14th ed. Darien, Conn.: Hafner.

[18] M. Gail and N. Mantel. 1977. Counting the number of $r \times c$ contingency tables with fixed margins. *Journal of the American Statistical Association* 72(360):859–862.
http://www.jstor.org/stable/2286475

[19] T. P. Hill. 1996. A statistical derivation of the significant-digit law. *Statistical Science* 10(4):354–363. DOI: 10.1214/ss/1177009869

[20] ——— . 1998. The first digit phenomenon. *American Scientist* 86(4):358.
DOI: 10.1511/1998.4.358

[21] J. O. Irwin. 1935. Tests of significance for differences between percentages based on small numbers. *Metron* 12:83–94.

[22] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[23] E. L. Lehmann. 1959. *Testing statistical hypotheses*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley and Sons.

[24] F. Mosteller. 1952. The world series competition. *Journal of the American Statistical Association* 47(259):355–380. http://www.jstor.org/stable/2281309

[25] S. Newcomb. 1881. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* 4(1):39–40. http://www.jstor.org/stable/2369148

[26] W. M. Patefield. 1981. Algorithm AS 159: An efficient method of generating random r × c tables with given row and column totals. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 30(1):91–97. http://www.jstor.org/stable/2346669

[27] E. S. Pearson and H. O. Hartley, eds. 1972. *Biometrika tables for statisticians*, volume 2. Cambridge: Cambridge University Press. Revision of *Tables for statisticians and biometricians*, edited by Karl Pearson.

[28] 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

[29] ——— . 1922. On the χ^2 test of goodness of fit. *Biometrika* 14(1/2):186–191. http://www.jstor.org/stable/2331860

[30] K. Pearson, A. Lee, and L. Bramley-Moore. 1899. Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society A* 192:257–330.

[31] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[32] N. M. Razali and Y. B. Wah. 2011. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics* 2(1):21–33. http://instatmy.org.my/downloads/e-jurnal%202/3.pdf

[33] S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52(3/4):591–611. http://www.jstor.org/stable/2333709

[34] E. H. Simpson. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B.* 13:238–241.

[35] N. Smirnov. 1939. Estimation of the discrepancy between empirical distributions for two samples. *Bulletin Mathématique de l'Univerité Moscou. Série Interationale.* 2(2):3–13.

[36] ——— . 1948. Table for estimating the goodness of fit of empirical distributions. *AnnMathStat* 19(2):279–281. Reprint of Tables from [35] DOI: 10.1214/aoms/1177730256

[37] M. A. Stephens. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69(347):730–737. DOI: 10.1080/01621459.1974.10480196

[38] K. D. Tocher. 1950. Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika* 37(1/2):130–144. http://www.jstor.org/stable/2332156

[39] B. L. van der Waerden. 1969. *Mathematical statistics*. Number 156 in Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete. New York, Berlin, and Heidelberg: Springer–Verlag. Translated by Virginia Thompson and Ellen Sherman from *Mathematische Statistik*, published by Springer-Verlag in 1965, as volume 87 in the series Grundlerhen der mathematischen Wissenschaften.

[40] F. Yates. 1934. Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statisitcal Society* 1(2):217–235. DOI: 10.2307/2983604

[41] G. U. Yule. 1903. Notes on the theory of association of attributes in statistics. *Biometrika* 2(2):121–134. DOI: 10.1093/biomet/2.2.121.