Caltech Department of Mathematics

Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2017

Lecture 23: Specification Tests

Relevant textbook passages:

Larsen–Marx [18]: 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 [23], and is known as the **chi-square test**.

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

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.

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 [18]. But you can find it discussed by Breiman [8, pp. 213–217], van der Waerden [31, § 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 7.10.3 asserts that if F is the common cumulative distribution function of the X_is, then

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

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 11.1.2) 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,

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

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

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

(See van der Waerden [31, pp. 67–75], esp. p. 73, or Feller [11].) Smirnov [28, 29] 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 test 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 ad test $F \neq F_0$.) Under the null hypothesis (G is uniform),
$$P\left(\max_x G(x) - x > \varepsilon\right) \approx e^{-2n\varepsilon^2},$$
n is the sample size. (See van der Waerden [31, 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 valuesfor small n, and they may be found for instance in van der Waerden [31, 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 [10, pp. 86–89].) With Mathematica, use the `KolmogorovSmirnovTest` command.

	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 [31, Tables 4 and 5, pp. 344–345]

23.2.1 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 θ_0, typically by MLE, and test the simple hypothesis. Breiman [8, 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 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 case may be found in Pearson and Hartley [22, Tables 54, 55, pp. 117–123].

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

◦ The Shapiro–Wilk [27] 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 [26] and Stephens [30] 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.

23.3 Reminder of the multinomial distribution

Pitman [25]: p. 155
Larsen–Marx [18]: 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

$$p_X(n_1, \ldots, n_K) = \frac{n!}{n_1! \cdot n_2! \cdots n_K!} p_1^{n_1} \cdot p_2^{n_2} \cdot \cdots \cdot p_K^{n_K} \qquad \left(\sum_k n_k = n \right).$$

It is easy to see that each X_k is Binomial(n, p_k) [18, 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_ks are not independent, since they must sum to n.

23.4 "Goodness of fit" tests

Larsen–Marx [18]: Section 10.3, pp. 499–509

When we have a multinomial model, 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 [23] 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 [18] we need $np_k^0 \geqslant 5$, for all $k = 1, \ldots, K$ to use this approximation, but van der Waerden [31, p. 238] and others think this is too conservative.

23.5 ⋆ Why the χ^2 test works

The following outline of a proof is based on Breiman [8, pp. 187–195] and Cramér [9, pp. 416–419]. For more details, see also van der Waerden [31, § 27, pp. 113-118, § 49, pp. 197–202, and § 51, pp. 207–211].

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 22.10.1, $\boldsymbol{Y}$ is approximately a jointly Normal random vector. So from here on, I will treat $\boldsymbol{Y}$ as if it were 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

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

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

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

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 22.12.2.

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

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

where $\boldsymbol{W}$ is treated as a column matrix. Since $\boldsymbol{W}$ is jointly Normal, therefore so is $\boldsymbol{Z}$. By Proposition 22.3.2, multiplication by 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 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 22.8.4, the covariance matrices satisfy

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

(Since B is orthogonal, $BIB' = BB' = I$).

Now examine the matrix $B\boldsymbol{v}\boldsymbol{v}'B' = (B\boldsymbol{v})(B\boldsymbol{v})'$ that appears in (8). By construction, the last row of B is $\boldsymbol{v}'$, and the other rows of B are orthogonal to $\boldsymbol{v}$. Thus $B\boldsymbol{v}$ is a K-column vector of zeroes except for the last entry, which is 1. So $B\boldsymbol{v}\boldsymbol{v}'B'$ 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.

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

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. You can find a statement of the following theorem in Larsen–Marx [18, Theorem 10.4.1] or Pitman [25, Theorem 6.13, p. 196].

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 [9, § 30.3, pp. 424–434].

Heuristically, the first order conditions for a maximum with respect to an estimated parameter impose another restriction on the relations of 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 set a lower threshold for rejection.

23.6.1 Theorem *Let*

$$H_0 \colon \boldsymbol{p} = \boldsymbol{p}_0,$$

where $\boldsymbol{p}_0$ is a K-dimensional vector of probabilities that sum to one. Estimate s parameters that determine p by the Maximum Likelihood Method. Then the test statistic

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

has an approximately Chi-square distribution with $K - 1 - s$ degrees of freedom. (For a good approximation, you should have each $np_k \geqslant 5$.)

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 [12]. Karl Pearson[24], who invented the χ^2test refused to accept Fisher's argument and wrote:

> 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.

23.7 The wonderfulness of the Chi-square Test

23.7.1 Example (The cookie data) A few years ago I passed out chocolate chip cookies, and asked students to email the the results of their cooki dissections. 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
# 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 μ 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
$\frac{(n_i - np_i)^2}{np_i}$	0.022	1.79	0.87

This give the test statistic = 2.69. There are 3 bins and 1 estimated parameter, μ, 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.

□

23.7.2 Example (The World Series) World Series come in lengths of 4, 5, 6, and 7, so there are $K = 4$ categories of results of this experiment. The probability of length ℓ is

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

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

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

has a $\chi^2(3)$-distribution. We can use this to test the null hypothesis (9).

(To find critical values in Mathematica, use the `InverseCDF` command. It has two arguments, the first is a named distribution, the second the α 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. □

23.7.3 Example (Case Study 10.3.2 [18]: 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 [21]. 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 [18, pp. 502–505]. There are also sets of numbers that disobey Benford'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 [16] has developed a sophisticated probabilistic model that predicts Benford's Law, but it is beyond the scope of this course. See also [17] 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 that predicted by Benford's Law. So deviations from Benford's Law are a sign that something is amiss.

Larsen and Marx [18, 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. Their results won't be ready until July 2015, but preliminary indications are that much the accounting data are not consistent with Benford's Law. □

23.7.4 Example (Did Mendel Cheat?) See Larsen–Marx [18, 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 fit this good 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 had faked his data. □

Larsen–Marx [18]: Section 10.4

23.7.5 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]`) □

23.8 Some practical considerations

I mentions 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_k 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 $\geqslant n$." The probability to assign to this bin is given by the Poisson(μ) probability

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

23.9 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.

Larsen–Marx [18]: § 10.5

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:

	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 with 3 degrees of freedom. 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 [18, 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 0.75. This means that a χ^2 random variable with 1 degree of freedom will be greater than the value of test statistic with probability of about 0.25.

Note that this test is not the same as the test we performed in Example 23.7.4.

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

P. J. Bickel, E. A. Hammel, and J. W. O'Connell [5] 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.

	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? 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 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, a more elaborate χ^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.

23.10.1 Example The University has two departments. Department A has 10 spots and Department B has 20. There are 250 male and 150 female applicants to the U. Here are the admissions data.

	University			Department A			Department B		
	Men	Women	Total	Men	Women	Total	Men	Women	Total
Applicants	250	150	400	100	100	200	150	50	200
Admittees	20	10	30	5	5	10	15	5	20
Selectivity	8%	6.7%	7.5%	5%	5%	5%	10%	10%	10%

Men are admitted at a rate higher than that of women overall, but at the same 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 invert the test 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

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

This estimator is one of the estimators considered by Mosteller's [20] analysis of the World Series.

23.12 Minimum Chi-square estimation and the World Series

Here is the function to minimized based on the 107-game sample of best-of-seven World Series.

Be sure to update these data annually.

[2] Mathematica reports that it calculated to 307 places to the right of the decimal point.

$$D(p) = \frac{21^2}{\binom{3}{3}\big(p^4(1-p)^0 + (1-p)^4p^{0)}\big)} + \frac{25^2}{\binom{4}{3}\big(p^4(1-p)^1 + (1-p)^4p^1\big)} + \frac{24^2}{\binom{5}{3}\big(p^4(1-p)^2 + (1-p)^4p^2\big)} + \frac{37^2}{\binom{6}{3}\big(p^4(1-p)^3 + (1-p)^4p^3\big)}$$

The minimum χ^2 estimate of p is 0.600487, which is remarkable agreement with the MLE of 0.601728, and the method of moments estimate of 0.589607.

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

R. A. Fisher [13, 14] developed an alternative to the chi-square test for contingency tables. The test does not rely on the multivariate central limit theorem, and gives an exact p-value. Hence the name Fisher's Exact Test.

It works like this: given the null hypothesis $\boldsymbol{p}$, we can compute the probability of getting any vector $\boldsymbol{n} = (n_1, \dots, n_K)$ of observed outcomes by

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

So we sum these probabilities for all vectors $\boldsymbol{n}'$ that are more extreme than the observed vector of counts $\boldsymbol{n}$ to get the p-value of the test. A vector $\boldsymbol{n}'$ is more extreme than $\boldsymbol{n}$ if whenever $n_k \geqslant np_k$, then $n'_k \geqslant n_k$, and whenever $n_k \leqslant np_k$, then $n'_k \leqslant n_k$. You can imagine that if n

and K are large, then there are a lot of such vectors $\boldsymbol{n}'$, see, e.g., [15]. In the 1920s this seemed like daunting task, but with today's technology, it is really quite practical and most statistical software will happily perform a Fisher exact test for you.

So why is the test 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, but there is another reason too. Since the set of p-values that can result from a data set is discrete, and the one with p-values less than 0.05 might lead to a much more stringent criterion for rejecting the null hypothesis. (This criticism applies to any test based on a discrete distribution.) There are still active discussions on statistics websites such as http://stats.stackexchange.com about the merits of each test.

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

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

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

[12] 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

[13] ——— . 1934. *Statistical methods for research workers*, 1934 ed. Edinburgh: Oliver and Boyd.

[14] ———— . 1935. The logic of inductive inference. *Journal of the Royal Statistical Society* 98(1):39–82. http://www.jstor.org/stable/2342435

[15] 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

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

[17] ———— . 1998. The first digit phenomenon. *American Scientist* 86(4):358. DOI: 10.1511/1998.4.358

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

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

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

[21] 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

[22] 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.

[23] 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

[24] ———— . 1922. On the χ^2 test of goodness of fit. *Biometrika* 14(1/2):186–191. http://www.jstor.org/stable/2331860

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

[26] 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

[27] 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

[28] N. Smirnov. 1939. Estimation of the discrepancy between emirical distributions for two samples. *Bulletin Mathématique de l'Université Moscou* 2.

[29] ———— . 1948. Table for estimating the goodness of fit of empirical distributions. *Ann-MathStat* 19(2):279–281. Reprint of Tables from [28] DOI: 10.1214/aoms/1177730256

[30] 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

[31] 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.