Contingency Tables
(Conover Chapter Four)

STAT 345-01: Nonparametric Statistics *

Fall Semester 2018

Contents

Tuesday 6 November 2018
– Read Section 4.5 of Conover

This presentation is somewhat theoretical. For a complementary approach with more examples, see `https://ccrg.rit.edu/~whelan/courses/2018_1sp_MATH_252/notes14.pdf`

0 Types of Data (See Section 2.1 of Conover)

We haven't placed too much emphasis on the formalities of types of data and measurement scales, but it's worth taking a moment to define them, since we will be concerned in this chapter primarily with categorical data. Roughly speaking, it's convenient to classify possible types of data as follows, from most specific (and probably most familiar) to least specific.

- **Numerical** or **Cardinal Data** are what you might normally think of in a mathematical sense: numbers. Each data value x is a number, and in particular if x and x' are two data values, it's meaningful to talk about things like the

size of the difference between them, $x - x'$. Conover further subdivides this category into and *interval scale* where $x - x'$ is meaningful and a *ratio scale* where both $x - x'$ and x/x' are meaningful. (An example of a measurement on an interval but not a ratio scale is a timestamp. 11:00 is 45 minutes later than 10:15, but it's not meaningful to take the ratio of two times of day.) You can also imagine making a distinction between continuous and discrete data.

- **Ordinal Data** are data for which you can ask whether $x > x'$, but the value of $x - x'$ is not meaningful. Most of the rank-based methods we've discussed can operate on ordinal data (measured on what Conover calls a *ordinal scale*), while most parametric methods (t-tests, ANOVA, etc) require at least an interval if not a ratio scale.
- **Categorical Data** are data which are really not even inherently numerical, like eye color, hair color, political party registration, etc. We can say whether $x = x'$ or $x \neq x'$, but there is no meaningful ordering of different data values. We may find it convenient to label the categories with numbers, but the ordering of the numbers is basically arbitrary. Note that the entries in a Sudoku puzzle, while usually written as integers 1 to 9, are only categorical as far as the rules of the puzzle goes.

1 The Chi-squared Goodness-of-fit Test

Suppose we have a categorical data sample $\{x_I | I, \ldots, N\}$ of size N where there are c categories $\{\mathcal{C}_i | i = 1, \ldots, c\}$. The null hypothesis H_0 specifies probabilities $\{p_i^* | i = 1, \ldots, c\}$ for an observation to fall into each of the categories, i.e., $P(X = \mathcal{C}_i | H_0) = p_i^*$. For consistency, the probabilities must satisfy $\sum_{i=1}^{c} p_i^* = 1$. We don't generally work with the original categorical data, since all of the meaningful inference can be done with the number of observations O_i in each category,

$$O_i = \sum_{I=1}^{N} I[x_I = \mathcal{C}_i] \tag{1.1}$$

where

$$\sum_{i=1}^{c} O_i = N \tag{1.2}$$

If we think of these $\{O_i\}$ as a random vector, the corresponding probability distribution is the multinomial distribution, a generalization of the binomial distribution

$$p(\{O_i\}|N, \{p_i\}) = \frac{N!}{O_1! O_2! \cdots O_c!} p_1^{O_1} p_2^{O_2} \cdots p_c^{O_c}$$
$$O_i = 0, 1, 2, \ldots; \quad O_1 + \cdots + O_c = N \tag{1.3}$$

Each one of these c random variables is a binomial $\mathrm{Bin}(N, p_i)$ with

$$E(O_i) = N p_i \qquad \text{and} \qquad V(O_i) = N p_i (1 - p_i) \tag{1.4}$$

but of course they are not independent, since there is a constraint

$$\sum_{i=1}^{c} O_i = N \tag{1.5}$$

This means that if we know the values of O_1, O_2, …, O_{c-1}, then O_c is determined. The null hypothesis specifies expected number counts of $E_i = N p_i^*$ for each category, which satisfy $\sum_{i=1}^{c} E_i = N$. Note that in general the $\{E_i\}$ will not be integers, although the $\{O_i\}$ are by definition.

As usual, we have the expected and observed values of a set of random variables. If the sample size (here the number of

observations N or more precisely the expected number counts $\{E_i\}$ is large enough that we can approximate things with the normal distribution, we should be able to write a chi-squared statistic, as we did in the Kruskal-Wallis or Friedman test. As a reminder, this is in analogy with the following two cases:

1. For independent normal random variables $X_i \sim N(\mu_i, \sigma_i^2)$,

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2(n) \tag{1.6}$$

2. For a random sample $\{X_i\}$ from $N(\mu, \sigma^2)$, if we define $Y_i = X_i - \overline{X}$, so that there's a constraint $\sum_{i=1}^{n} Y_i = 0$,

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma_i} \right)^2 = \sum_{i=1}^{n} \left(\frac{Y_i - 0}{\sigma_i} \right)^2 \sim \chi^2(n-1) \tag{1.7}$$

As usual, to do this rigorously, we'd need to take into account the covariances $\mathrm{Cov}(O_i, O_j)$ and invert a variance-covariance matrix, but we can motivate the form by looking at the special case $c = 2$. In this case the multinomial distribution is really a binomial distribution, where $p_1 = p$ and $p_2 = 1 - p$, and the random variable $O_2 = N - O_1$ is redundant with $O_1 \sim \mathrm{Bin}(N, p_1)$. In this case, if we apply the normal approximation to the single non-trivial random variable, we get a chi-squared statistic

$$W = \frac{(O_1 - Np_1^*)^2}{Np_1^*(1-p_1^*)} = \frac{(O_1 - Np_1^*)^2}{Np_1^* p_2^*} \tag{1.8}$$

We'd like to rewrite this in a way that treats 1 and 2 more symmetrically, and can be written entirely in terms of O_1, O_2, E_1 and E_2. The denominator can be rewritten using

$$\frac{1}{p_1^*} + \frac{1}{p_2^*} = \frac{p_2^* + p_1^*}{p_1^* p_2^*} = \frac{1}{p_1^* p_2^*} \tag{1.9}$$

so that

$$W = \frac{(O_1 - Np_1^*)^2}{Np_1^*} + \frac{(O_1 - Np_1^*)^2}{Np_2^*} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_1 - E_1)^2}{E_2} \tag{1.10}$$

Now, this still doesn't treat the two categories symmetrically, but if we remember that $O_1 + O_2 = N = E_1 + E_2$, we see that

$$O_1 - E_1 = -(O_2 - E_2) \tag{1.11}$$

which means we can write

$$W = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{1.12}$$

which is approximately chi-squared distributed with $2 - 1 = 1$ degree of freedom when H_0 is true.

This form turns out to extend to larger numbers of categories, and the normal approximation gives us

$$W = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(c-1) \qquad \text{if } H_0 \text{ true} \tag{1.13}$$

expanding the square gives a "shortcut formula"

$$\begin{aligned} W &= \sum_{i=1}^{c} \frac{{O_i}^2 - 2E_i O_i + E_i^2}{E_i} = \sum_{i=1}^{c} \left(\frac{{O_i}^2}{E_i} - 2O_i + E_i \right) \\ &= \frac{{O_i}^2}{E_i} - 2N + N = \frac{{O_i}^2}{E_i} - N \end{aligned} \tag{1.14}$$

Note that although this is nominally simpler, involving fewer operations, it could actually be less convenient if E_i and O_i are both large, since it involves squaring larger numbers. (In principle, this could also run the risk of roundoff error.)

1.1 Testing with Unknown Parameters

Sometimes the null hypothesis doesn't uniquely specify the $\{p_i^*\}$, but includes some unknown parameters $\boldsymbol{\theta} \equiv \{\theta_1, \ldots, \theta_m\}$. For instance, the null hypothesis might state that the number of male puppies in a litter of five is a binomial random variable $\mathrm{Bin}(5, \theta)$, but not specify the value of the parameter θ. The procedure in this case is to estimate the parameters using the observed data (either the categorical number counts or possibly some finer-grained data from which it was resolved). The best guess expected counts $\{\hat{E}_i = N p_i^*(\hat{\boldsymbol{\theta}})\}$ are then used to construct the goodness-of-fit statistic

$$\hat{W} = \sum_{i=1}^{c} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \tag{1.15}$$

where now the expected counts $\{\hat{E}_i\}$ also depend on the data. If we have fit m parameters in this way, we would ideally expect the null distribution of $\hat{W}$ to be $\chi^2(c-1-m)$. (This is exact if everything is linear, but it can actually be slightly *underestimate* the p-value in general, since the method used for estimating the parameters may not actually minimize the chi-squared statistic.)

1.2 Testing a Probability Distribution

One obvious type of categorical data to which the chi-squared test can be applied is a histogram. Take, for example, Problem 6.1.4 from Conover, which provided the results of "candling" eggs, listing the number of rejected eggs by crate as 4, 0, 2, 0, 2, 0, 2, and 0. The hypothesis that the number of rejected eggs is given by a Poisson distribution with mean 1.5, and it was tested in that problem using the Kolmogorov test and found a p-value of greater than 20%. We can also test the hypothesis using the chi-squared test, first collecting the observations by number of eggs rejected:

```
In [1]: from __future__ import division

In [2]: import numpy as np

In [3]: from scipy import stats

In [4]: x_iota = np.array([4, 0, 2, 0, 2, 0, 2, 0])

In [5]: N = len(x_iota); N
Out[5]: 8

In [6]: c = 6

In [7]: ilist = np.arange(c); ilist
Out[7]: array([0, 1, 2, 3, 4, 5])

In [8]: O_i = np.sum(x_iota[None,:]==ilist[:,None],
   axis=-1); O_i
Out[8]: array([4, 0, 3, 0, 1, 0])

In [9]: O_i[-1] = np.sum(x_iota>=(c-1)); O_i
Out[9]: array([4, 0, 3, 0, 1, 0])
```

Note that, although the distribution was discrete, we did have to make a decision in defining the categories, defining the last bin to be $x_i \geq 5$. This didn't matter for the observed number counts, which cut off at some point, but has an impact on the expected numbers, which we estimate using the pmf of the Poisson distribution:

```
In [10]: pstar_i = stats.poisson(1.5).pmf(ilist);
   pstar_i
Out[10]:
array([ 0.22313016, 0.33469524, 0.25102143,
   0.12551072, 0.04706652,
```

```
       0.01411996])

In [11]: pstar_i[-1] = stats.poisson(1.5).sf(c-1.5);
    pstar_i
Out[11]:
array([ 0.22313016, 0.33469524, 0.25102143,
   0.12551072, 0.04706652,
       0.01857594])

In [12]: np.sum(pstar_i)
Out[12]: 1.0

In [13]: E_i = N*pstar_i; E_i
Out[13]:
array([ 1.78504128, 2.67756192, 2.00817144,
   1.00408572, 0.37653215,
       0.14860749])
```

Now we can calculate the test statistic and compare it to a chi-squared with $6 - 1 = 5$ degrees of freedom.

```
In [14]: W = np.sum((O_i-E_i)**2/E_i); W
Out[14]: 8.1008829563997153

In [15]: np.sum(O_i**2/E_i)-N
Out[15]: 8.1008829563997153

In [16]: stats.chi2(df=c-1).sf(W)
Out[16]: 0.1507627106021614
```

We find a p-value of about 15%, still consistent with the hypothesized distribution.

If, on the other hand, we'd been told to expect a Poisson distribution but not told the mean, we could have estimated it from the observed data:

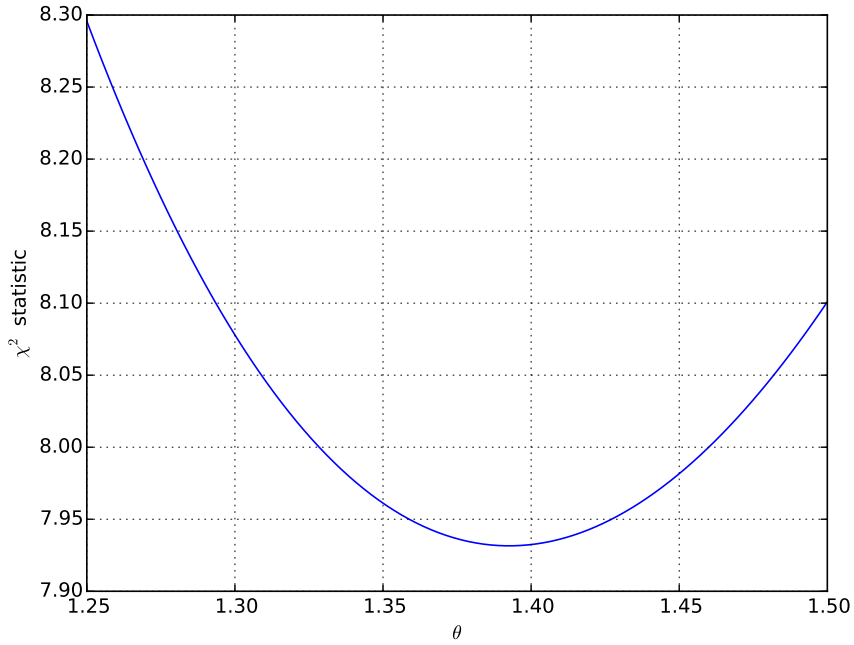
```
In [17]: np.mean(x_iota)
Out[17]: 1.25

In [18]: phatstar_i = stats.poisson(1.25).pmf(ilist)
   ; phatstar_i
Out[18]:
array([ 0.2865048 , 0.358131 , 0.22383187,
   0.09326328, 0.02914478,
       0.00728619])

In [19]: phatstar_i[-1] = stats.poisson(1.25).sf(c
   -1.5); phatstar_i
Out[19]:
array([ 0.2865048 , 0.358131 , 0.22383187,
   0.09326328, 0.02914478,
       0.00912428])

In [20]: np.sum(phatstar_i)
Out[20]: 1.0

In [21]: Ehat_i = N*phatstar_i; Ehat_i
Out[21]:
array([ 2.29203837, 2.86504797, 1.79065498,
   0.74610624, 0.2331582 ,
       0.07299423])

In [22]: What = np.sum((O_i-Ehat_i)**2/Ehat_i); What
Out[22]: 8.2957131997978433
```

Now, since we've used the data to estimate the mean of the distribution, we should compare the value to a chi-squared with $6 - 1 - 1 = 4$ degrees of freedom:

```
In [23]: stats.chi2(df=c-1-1).sf(What)
Out[23]: 0.08132708524626904
```

The p-value is actually lower, 8.1%, but still indicates data plausibly consistent. Actually, something is a little odd here, because the chi-squared value of 8.3 is slightly *higher* than the 8.1 we had with the hypothesized Poisson mean of 1.5. This is an illustration of the pitfall in estimating parameters. We can see what value actually produces the lowest chi-squared:

```
In [24]: theta = np.linspace(1.25,1.5,100)

In [25]: ptheta_i = stats.poisson(theta[:,None]).pmf
   (ilist[None,:])

In [26]: ptheta_i[:,-1] = stats.poisson(theta).sf(c
   -1.5)

In [27]: min(np.sum(ptheta_i,axis=-1))
Out[27]: 0.99999999999999989

In [28]: max(np.sum(ptheta_i,axis=-1))
Out[28]: 1.0000000000000002

In [29]: Etheta_i = N*ptheta_i

In [30]: Wtheta = np.sum((O_i[None,:]-Etheta_i)**2/
   Etheta_i,axis=-1)

In [31]: figure();

In [32]: plot(theta,Wtheta);

In [33]: xlabel(r'$\theta$');

In [34]: ylabel(r'$\chi^2$ statistic');

In [35]: grid(True);

In [36]: xlim(1.25,1.5);

In [37]: savefig('notes04_minchisq.eps',bbox_inches
   ='tight');
```



We see that the chi-squared takes on its minimum value of 8.1 when the Poisson mean is about 1.39:

```
In [38]: theta[np.argmin(Wtheta)]
Out[38]: 1.3914141414141414

In [39]: Wmin = min(Wtheta); Wmin
Out[39]: 7.9315533424663851
```

```
In [40]: stats.chi2(df=c-1-1).sf(Wmin)
Out[40]: 0.094117906644314936
```

The p-value is 9.4%, which is still somewhat lower than we had with the hypothesized Poisson mean. Although the chi-squared is lower, it has to be compared to a distribution with one fewer degree of freedom. (The data are still consistent with the null hypothesis in any event.)

Thursday 15 November 2018
– Read Section 4.2 of Conover

2 Two-Way Contingency Tables

An important generalization of the categorical data problem is when there are multiple sets of categories, and each observation is classified into one category from each set. For instance, we may be randomly drawing individuals from a population and categorizing them based on eye color and hair color. We can think of each observation as a multi-dimensional categorical vector. For concreteness, we focus on the case where there are two sets of categories: $\{\mathcal{R}_1, \ldots, \mathcal{R}_r\} =\equiv \{\mathcal{R}_i | i = 1, \ldots, r\}$ and $\{\mathcal{C}_1, \ldots, \mathcal{C}_c\} =\equiv \{\mathcal{C}_j | j = 1, \ldots, c\}$. The N observations are then a paired categorical data sample $\{(x_I, y_I) | I = 1, \ldots, N\}$, where $x_I \in \{\mathcal{R}_i\}$ and $y_I \in \{\mathcal{C}_j\}$. We can count up the number of observations in each pair of categories

$$O_{ij} = \sum_{I=1}^{N} I[x_I = \mathcal{R}_i, y_I = \mathcal{C}_j] \tag{2.1}$$

and arrange them into what's known as a **contingency table**:

	$j=1$	$j=2$	$\cdots$	$j=c$	
$i=1$	O_{11}	O_{12}	$\cdots$	O_{1c}	r_1
$i=2$	O_{21}	O_{22}	$\cdots$	O_{2c}	r_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$i=r$	O_{r1}	O_{r2}	$\cdots$	O_{rc}	r_r
	c_1	c_2	$\cdots$	c_c	N

We can define the row and column totals (total number of observations in each set of categories)

$$\sum_{i=1}^{r} O_{ij} = c_j \qquad \sum_{j=1}^{c} O_{ij} = r_i \tag{2.2}$$

The total number of observations is given by

$$\sum_{i=1}^{r} r_i = N = \sum_{j=1}^{c} c_j \tag{2.3}$$

Rather than checking a hypothesized set of probabilities for each category, the usual test checks for an association between the categories in the two sets. The null hypothesis says that the column and row in which an observation is categorized have no influence on each other, e.g., that a person with blue eyes is no more likely to have brown as opposed to blond hair than a person with brown eyes. Defining a null distribution for a statistic requires us to define what we'd mean by repeated experiments. There are several different possible assumptions:

1. Only the total number of observations N is fixed. Row totals $\{R_i\}$ and column totals $\{C_j\}$ are random variables. In general, $\{O_{ij} | i = 1, \ldots, r; j = 1, \ldots, c\}$ are a multinomial random vector with probabilities $\{p_{ij}\}$. The null hypothesis says $p_{ij} = p_{i\bullet} p_{\bullet j}$ for some $\{p_{i\bullet} | i = 1, \ldots, r\}$ with $\sum_{i=1}^{r} p_{i\bullet} = 1$ and some $\{p_{\bullet j} | i = 1, \ldots, c\}$ with $\sum_{j=1}^{c} p_{\bullet j} = 1$.

2. The total number r_i in each row is fixed. For each i, we have a multinomial random vector $\{O_{ij}|i = 1,\ldots,r; j = 1,\ldots,c\}$, which in general has probabilities $\{p_1^{(i)}, p_2^{(i)}, \ldots, p_c^{(i)}\} \equiv \{p_j^{(i)}\}$. The null hypothesis says $p_j^{(i)} = p_{\bullet j}$ for all i.
3. The row numbers $\{r_i\}$ and column numbers $\{c_j\}$ are assumed to be fixed. The distribution of $\{O_{ij}\}$ is then purely down to combinatorics: how do we arrange the N observations into rows and columns in a way that respects the marginal totals.

In each case we'll have a null expectation value E_{ij} for the number of observations in row i and column j, and we'll define a chi-squared statistic

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^2}{E_{ij}} - N \tag{2.4}$$

It will turn out that for each of the assumptions about the null distribution, $E_{ij} = r_i c_j/N$, and, under the null hypothesis and the normal approximation, the statistic will be chi-squared distributed with $(r-1)(c-1)$ degrees of freedom. For small samples, however, the details of the null distribution will depend on the assumptions made about the experimental setup.

Tuesday 20 November 2018 – Read Section 4.1 of Conover

2.1 2×2 Contingency Tables

You saw on the last homework that there were slightly different null distributions for the chi-squared statistic from a two-way contingency table, depending on whether we assumed the row totals were fixed or only the total number of observations. We want to explore this in a little more detail, so we look at the simple case where $r = 2 = c$. In that case, the contingency table looks like this:

	$j=1$	$j=2$	
$i=1$	O_{11}	O_{12}	r_1
$i=2$	O_{21}	O_{22}	r_2
	c_1	c_2	N

The exact probability distributions depend on the details of the experimental setup, in particular which quantities (the row and column totals, just the row totals, or just the total number of observations) are assumed to be held fixed in hypothetical repetitions of the experiment.

2.1.1 Fisher's Exact Test

The most restrictive set of assumptions is that the $\{r_i\}$ and $\{c_j\}$ are known in advance, and only the $\{O_{ij}\}$ are random. This means that everything is actually determined by one random variable, say O_{11}, and then we can find $O_{12} = r_1 - O_{11}$, $O_{21} = c_1 - O_{11}$, and

$$O_{22} = r_2 - O_{21} = N - r_1 - c_1 + O_{11} \tag{2.5}$$

The null probability distribution for O_{11} is a hypergeometric distribution, since we are choosing r_1 objects out of N, where c_1 of the N are of a certain type, and finding that O_{11} out of the r_1 are of the type we chose.

As a quick refresher on the hypergeometric distribution, and sampling without replacement, suppose that we have a hand with seven cards, three of them hearts and four spades:

♡ 2 3 4
♠ 2 3 4 5

Suppose we draw four cards, at random, without replacement. What is the probability that we will end up with one heart and three spades? There are

$$\binom{7}{4} = \frac{7!}{4!3!} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = 35 \tag{2.6}$$

possible sets of four cards we can draw out of the seven available. How many of those include one heart and three spades? There are

$$\binom{3}{1} = \frac{3!}{1!2!} = 3 \tag{2.7}$$

possibilities for the heart and

$$\binom{4}{3} = \frac{4!}{3!1!} = 4 \tag{2.8}$$

possibilities for which three spades we have, so there are $3 \times 4 = 12$ different four-card sets with one heart and three spades. So the probability of getting exactly one heart in a four card set when there are three hearts available out of seven cards to choose from is

$$\frac{\binom{3}{1}\binom{4}{3}}{\binom{7}{4}} = \frac{12}{35} \approx 0.3429 \tag{2.9}$$

Returning to the contingency table, the probability of getting O_{11} of the r_1 observations in row 1 to be in column 1, when we know that c_1 of the total N observations will be in column 1 is

$$\frac{\binom{c_1}{O_{11}}\binom{c_2}{O_{12}}}{\binom{N}{r_1}} = \frac{c_1!c_2!r_1!r_2!}{N!O_{11}!O_{21}!O_{12}!O_{22}!} \tag{2.10}$$

If we know that the row and column totals are fixed in an experimental design, "Fisher's exact test" says to compare the one degree of freedom in the table, say O_{11}, against the appropriate hypergeometric distribution. Suppose for example we have the following contingency table

1	6	7
8	2	10
9	8	17

```
In [1]: from __future__ import division

In [2]: import numpy as np

In [3]: from scipy import stats

In [4]: O_ij = np.array([[1,6],[8,2]])

In [5]: r_i = np.sum(O_ij,axis=-1); r_i
Out[5]: array([ 7, 10])

In [6]: c_j = np.sum(O_ij,axis=0); c_j
Out[6]: array([9, 8])

In [7]: N = np.sum(r_i); N
Out[7]: 17

In [8]: O11 = O_ij[0,0]; O11
Out[8]: 1

In [9]: r1 = r_i[0]; r1
Out[9]: 7

In [10]: c1 = c_j[0]; c1
Out[10]: 9

In [11]: x = np.arange(max(0,r1+c1-N),min(r1,c1)+1);
    x
Out[11]: array([0, 1, 2, 3, 4, 5, 6, 7])
```

```
In [12]: px = stats.hypergeom(N,c1,r1).pmf(x); px
Out[12]:
array([ 0.00041135, 0.01295763, 0.10366104,
   0.30234471, 0.36281366,
       0.18140683, 0.03455368, 0.00185109])

In [13]: pxobs = stats.hypergeom(N,c1,r1).pmf(O11);
   pxobs
Out[13]: 0.012957630604689428

In [14]: pval_fisher = np.sum(px[px<=pxobs]);
   pval_fisher
Out[14]: 0.015220074043603449
```

For large N, we can apply a normal approximation to the hypergeometric distribution using the mean and variance

$$E(O_{11}) = \frac{r_1 c_1}{N} \tag{2.11a}$$

$$\mathrm{Var}(O_{11}) = \frac{N - r_1}{N - 1} r_1 \frac{c_1}{N} \frac{N - c_1}{N} = \frac{r_1 r_2 c_1 c_2}{N^2 (N - 1)} \tag{2.11b}$$

and so the test statistic

$$T = \sqrt{\frac{N - 1}{r_1 r_2 c_1 c_2}} (N O_{11} - r_1 c_1) \tag{2.12}$$

should be approximately standard normal. This looks pretty asymmetric, but if we write

$$\begin{aligned} N O_{11} - r_1 c_1 &= (O_{11} + O_{12} + O_{21} + O_{22}) O_{11} - (O_{11} + O_{12})(O_{11} + O_{21}) \\ &= O_{11} O_{22} - O_{12} O_{21} \end{aligned} \tag{2.13}$$

we get the more symmetrical form

$$T = \sqrt{\frac{N - 1}{r_1 r_2 c_1 c_2}} (O_{11} O_{22} - O_{12} O_{21}) \tag{2.14}$$

2.1.2 Test for Equal Proportions

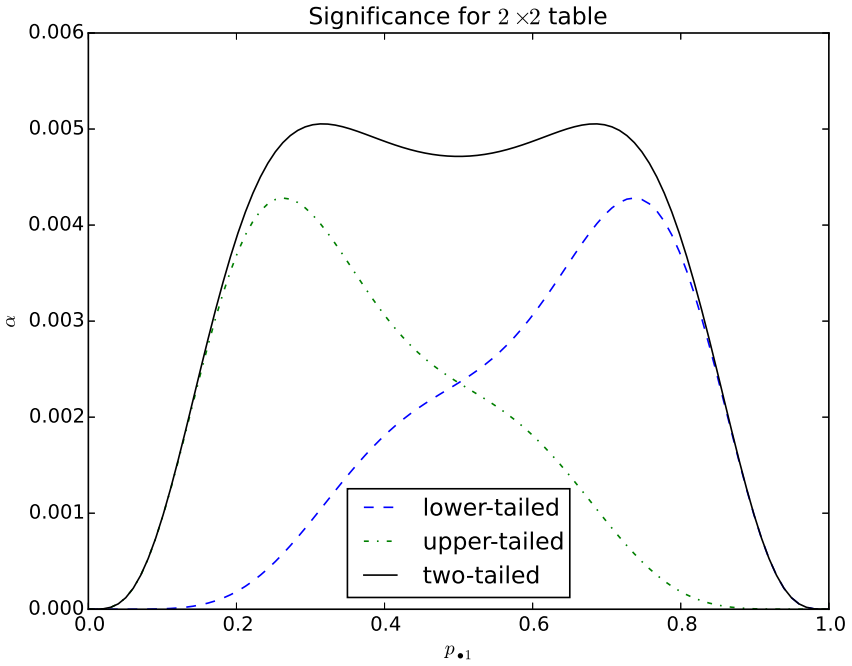
The assumption that the row and column totals are both fixed makes for a somewhat artificial experimental setup. Fisher developed his test in the context of the so-called "lady tasting tea" experiment. A lady claimed to be able to tell by taste whether milk had been added to a cup of tea, or tea had been poured into a cup already containing milk. Fisher tested her ability by preparing eight cups of tea, four by each method, and asking her to determine which four had the tea poured first and which four had the milk first. This setup did indeed guarantee that the row totals (number of cups prepared tea-first and number milk-first) as well as the column totals (number of cups identified as tea-first and number identified as milk-first) were fixed (all to 4 in this case). But if he had not told her how many were prepared each way, but asked her to identify each cup as tea-first or milk-first as it came, the row totals would have been fixed, but not the column totals.

If the row totals are known but not the column totals, there are only two constraints, and thus two (independent) random variables $O_{11} \sim \mathrm{Bin}(r_1, p_1^{(1)})$ and $O_{21} \sim \mathrm{Bin}(r_2, p_1^{(2)})$. The null hypothesis says that $p_1^{(1)} = p_1^{(2)} = p_{\bullet 1}$, but doesn't specify the value of $p_{\bullet 1}$. The joint null pmf for O_{11} and O_{21} is

$$\begin{aligned} p(O_{11}, O_{21}) &= \binom{r_1}{O_{11}} p_{\bullet 1}^{O_{11}} (1 - p_{\bullet 1})^{r_1 - O_{11}} \binom{r_2}{O_{21}} p_{\bullet 1}^{O_{21}} (1 - p_{\bullet 1})^{r_2 - O_{21}} \\ &= \frac{r_1! r_2!}{O_{11}! O_{12}! O_{21}! O_{22}!} p_{\bullet 1}^{c_1} p_{\bullet 2}^{c_2} \end{aligned} \tag{2.15}$$

```
In [15]: r2 = r_i[1]; r2
Out[15]: 10

In [16]: c2 = c_j[1]; c2
```

```
Out[16]: 8

In [17]: O11vals = np.arange(r1+1); O11vals
Out[17]: array([0, 1, 2, 3, 4, 5, 6, 7])

In [18]: O21vals = np.arange(r2+1); O21vals
Out[18]: array([ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9,
   10])

In [19]: c1vals = O11vals[:,None] + O21vals[None,:]

In [20]: c2vals = N - c1vals

In [21]: statvals = (N*O11vals[:,None]-r1*c1vals)/np
   .sqrt(c1vals*c2vals+1e-6)

In [22]: mystat = (N*O11-r1*c1)/np.sqrt(c1*c2);
   mystat
Out[22]: -5.4211519890968649

In [23]: np.sum(statvals<=mystat)
Out[23]: 8

In [24]: np.sum(statvals>=-mystat)
Out[24]: 8

In [25]: p1 = np.linspace(0,1,100)

In [26]: pmfvals = stats.binom(r1,p1[None,None,:]).
   pmf(O11vals[:,None,None]) * stats.binom(r2,p1[
   None,None,:]).pmf(O21vals[None,:,None])

In [27]: plower = np.sum((statvals<=mystat)[:,:,None
   ]*pmfvals,axis=(0,1))

In [28]: pupper = np.sum((statvals>=-mystat)[:,:,
   None]*pmfvals,axis=(0,1))

In [29]: plot(p1,plower,'b--',label='lower-tailed');

In [30]: plot(p1,pupper,'g-.',label='upper-tailed');

In [31]: plot(p1,plower+pupper,'k-',label='two-
   tailed');

In [32]: legend(loc='lower center');

In [33]: xlabel(r'$p_{\bullet 1}$');

In [34]: ylabel(r'$\alpha$');

In [35]: title(r'Significance for $2\times 2$ table
   ');

In [36]: savefig('notes04_alpha2x2.eps',bbox_inches
   ='tight');

In [37]: pval_row = max(plower+pupper); pval_row
Out[37]: 0.0050535355899577281
```



We see that, if the column totals are not actually fixed, Fisher's exact test has overestimated the p-value.

For large sample sizes, we can also apply a normal approximation, assuming that the two random variables O_{11} and O_{21} are independent and approximately normal with the means and variances given by the binomial distribution:

$$E(O_{11}) = r_1 p_1^{(1)} \quad \text{and} \quad E(O_{21}) = r_2 p_1^{(2)} \tag{2.16a}$$

$$\mathrm{Var}(O_{11}) = r_1 p_1^{(1)}(1-p_1^{(1)}) \quad \text{and} \quad \mathrm{Var}(O_{21}) = r_2 p_1^{(2)}(1-p_1^{(2)}) \tag{2.16b}$$

To test the null hypothesis that $p_1^{(1)} = p_1^{(2)} = p_{\bullet 1}$, we should use the test statistic $O_{11}/r_1 - O_{21}/r_2$. Assuming H_0, this has mean

$$E\left(\frac{O_{11}}{r_1} - \frac{O_{21}}{r_2}\right) = 0 \tag{2.17}$$

and variance

$$\mathrm{Var}\left(\frac{O_{11}}{r_1} - \frac{O_{21}}{r_2}\right) = \left(\frac{1}{r_1} + \frac{1}{r_2}\right) p_{\bullet 1}(1-p_{\bullet 1}) \tag{2.18}$$

We can replace the unknown $p_{\bullet 1}$ with its estimate

$$\hat{p}_{\bullet 1} = \frac{O_{11}+O_{21}}{r_1+r_2} = \frac{C_1}{N} \tag{2.19}$$

and likewise

$$1-\hat{p}_{\bullet 1} = 1 - \frac{O_{11}+O_{21}}{r_1+r_2} = \frac{C_2}{N} \tag{2.20}$$

We then assume that the standardized statistic[1]

$$T_1 = \frac{O_{11}/r_1 - O_{21}/r_2}{\sqrt{\left(\frac{1}{r_1}+\frac{1}{r_2}\right)\frac{C_1}{N}\frac{C_2}{N}}} \tag{2.21}$$

is approximately standard normal. We can simplify the statistic as

$$\begin{aligned} T_1 &= \frac{r_2 O_{11} - r_1 O_{21}}{\sqrt{r_1 r_2 (r_2+r_1)\frac{C_1}{N}\frac{C_2}{N}}} = \frac{(O_{21}+O_{22})O_{11} - (O_{11}+O_{12})O_{21}}{\sqrt{r_1 r_2 (r_2+r_1)\frac{C_1}{N}\frac{C_2}{N}}} \\ &= \sqrt{\frac{N}{r_1 r_2 C_1 C_2}}\,(O_{11}O_{22} - O_{12}O_{21}) \end{aligned} \tag{2.22}$$

Note that this is $\sqrt{\frac{N}{N-1}}$ times the large-sample normal approximation to the test statistic in Fisher's exact test, (2.14). As you'll show on the homework, the square of T_1 appearing in (2.22) is the standard chi-squared statistic for a two-way contingency table. (This makes sense, since the square of a standard normal random variable is a $\chi^2(1)$ random variable.)

[1] This is just the usual test for a difference in population proportions, as in e.g., section 9.4 of Devore, *Probability and Statistics for Engineering and the Sciences*.

Tuesday 27 November 2018
– Read Section 4.3 of Conover

2.2 Mood's Median Test

We now describe a test which is an application of the chi-squared test for a $2 \times c$ contingency table. Suppose that you have c samples drawn from different populations and wish to test whether they have the same median. To keep in line with the notation of this chapter, we write the sizes of the c samples as $\{c_j | j = 1, \ldots, c\}$, and the total number of data values as $\sum_{j=1}^{c} c_j = N$ and we can write the data as $\{x_{\iota j} | \iota = 1, \ldots, c_j; j = 1, \ldots c\}$ where we use the Greek letter iota (ι) to avoid confusion with the column label i. Now, we know that we can use the Kruskal-Wallis test to check whether the samples all come from the same distribution, but suppose we want to allow for different distributions with the same (unspecified) median. An obvious estimate of that median is the median of all the $\{x_{\iota j}\} \equiv \{X_I | I = 1 \ldots, N\}$, known as the "grand median". We then categorize each point in each sample by whether it's above or below the grand median. Since it's possible to have some values exactly equal to the grand median (if N is odd and/or some data values are equal), we define the categories as $>$ and $\leq$. We can then arrange the number counts in these categories into a $2 \times c$ contingency table. (We'll follow Conover's somewhat counterintuitive convention of listing the number of values greater than the grand median in the first row.)

	$j=1$	$j=2$	$\cdots$	$j=c$	
$>$	O_{11}	O_{12}	$\cdots$	O_{1c}	r_1
$\leq$	O_{21}	O_{22}	$\cdots$	O_{2c}	r_2
	c_1	c_2	$\cdots$	c_c	N

If the samples are all drawn from distributions with the same median, the probability of a value landing above or below the median will be the same (nominally one-half, although this might be slightly different for discrete distributions with a non-zero probability of landing right at the median). So we can construct the standard chi-squared statistic for independence (this time of columns) in the contingency table:

$$T = \sum_{j=1}^{c} \sum_{i=1}^{2} \frac{N}{r_i c_j} \left(O_{ij} - \frac{r_i c_j}{N} \right)^2 \tag{2.23}$$

We can simplify this somewhat, along the same lines that we used when considering the chi-squared goodness-of-fit test with two categories) by noting that

$$O_{2j} - \frac{r_2 c_j}{N} = c_j - O_{1j} - \frac{(N - r_1) c_j}{N} = -\left(O_{1j} - \frac{r_1 c_j}{N} \right) \tag{2.24}$$

and so

$$T = N \sum_{j=1}^{c} \frac{\left(O_{1j} - r_1 c_j / N \right)^2}{c_j} \left(\frac{1}{r_1} + \frac{1}{r_2} \right) \tag{2.25}$$

but

$$\frac{1}{r_1} + \frac{1}{r_2} = \frac{r_2 + r_1}{r_1 r_2} = \frac{N}{r_1 r_2} \tag{2.26}$$

so

$$T = \frac{N^2}{r_1 r_2} \sum_{j=1}^{c} \frac{\left(O_{1j} - r_1 c_j / N \right)^2}{c_j} \tag{2.27}$$

If none of the original sample values are equal to the grand median, so that $r_1 = r_2 = N/2$ by definition, we have the further simplification

$$T = 4 \sum_{j=1}^{c} \frac{\left(O_{1j} - c_j/2 \right)^2}{c_j} = \sum_{j=1}^{c} \frac{\left(O_{1j} - O_{2j} \right)^2}{c_j} \quad \text{if } r_1 = r_2 \tag{2.28}$$

Note that if the sample size is not large enough to use the chi-squared approximation, the setup of the test means that the

column totals (the sizes of the c samples) are fixed, and the row totals are approximately fixed since we know $r_1 \approx \frac{N}{2} \approx r_2$. The null distribution that's appropriate is thus the one associated with fixed marginal totals.

Thursday 29 November 2018 – Read Section 4.6 of Conover

2.3 Cochran's Q Test

We now turn to a scenario which is somewhere between a contingency table and a complete block design. We consider a contingency table in which all of the counts are either 0 or 1, and are assumed to correspond to the yes/no response of r different subjects to c different treatments. We call these responses $\{X_{ij}\}$ and the data table looks like

	Treatment				
	1	2	$\cdots$	c	
$i=1$	X_{11}	X_{12}	$\cdots$	X_{1c}	r_1
$i=2$	X_{21}	X_{22}	$\cdots$	X_{2c}	r_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$i=r$	X_{r1}	X_{r2}	$\cdots$	X_{rc}	r_r
	c_1	c_2	$\cdots$	c_c	N

If we imagine repeating the experiment, each observation is a Bernoulli random variable. i.e., a binomial with one trial, $X_{ij} \sim \text{Bin}(1, p^{(ij)})$. We write the probability as $p^{(ij)}$ rather than p_{ij} to stress that there is no constraint placed on any sum of the probabilities, just a requirement that $0 \leq p^{(ij)} \leq 1$ for all i and j. We can also see that the marginal totals are all random statistics in this picture, since they represent the total numbers of successes that happen to occur:

	Treatment				
	1	2	$\cdots$	c	
$i=1$	X_{11}	X_{12}	$\cdots$	X_{1c}	R_1
$i=2$	X_{21}	X_{22}	$\cdots$	X_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$i=r$	X_{r1}	X_{r2}	$\cdots$	X_{rc}	R_r
	C_1	C_2	$\cdots$	C_c	N

The null hypothesis H_0 is that each subject responds the same way to all the treatments, i.e., $p^{(ij)} = p^{(i\bullet)}$ for each i, but we don't make any statements about the $\{p^{(i\bullet)}\}$. Under H_0, the row totals might be quite different, but the column totals (number of successes for each treatment) should be similar. We're thus interested in the statistical properties of $C_j = \sum_{i=1}^{r} X_{ij}$. It has mean

$$E(C_j) = \sum_{i=1}^{r} E(X_{ij}) = \sum_{i=1}^{r} p^{(i\bullet)} \tag{2.29}$$

and variance

$$\text{Var}(C_j) = \sum_{i=1}^{r} \text{Var}(X_{ij}) = \sum_{i=1}^{r} p^{(i\bullet)}(1 - p^{(i\bullet)}) \tag{2.30}$$

To construct a statistic, we need to replace the unknown $p^{(i\bullet)}$ with the estimator R_i/c. Note that this means the estimator of $E(C_j)$ is

$$\sum_{i=1}^{r} \frac{R_i}{c} = \frac{N}{c} = \frac{1}{c} \sum_{j=1}^{c} C_j \tag{2.31}$$

which is what you'd have written down as your best guess (the expected column total is the average of the column totals). Each column's contribution to the statistic should then be

$$\frac{[C_j - E(C_j)]^2}{\text{Var}(C_j)} \sim \frac{(C_j - N/c)^2}{\sum_{i=1}^{r} (R_i/c)(1 - R_i/c)} \tag{2.32}$$

Because both the mean and the variance are estimated, the relevant correlations turn out to give a statistic

$$Q = c(c-1)\frac{\sum_{j=1}^{c}(C_j - N/c)^2}{\sum_{i=1}^{r}(R_i)(c-R_i)} \tag{2.33}$$

whose null distribution is approximately $\chi^2(c-1)$.

2.3.1 Importance of Blocking Information

Note that, if we didn't know that all of the observations in the same block were related, we would pool them together and create a $2 \times c$ contingency table where the two rows were just successes and failures. The observations would be $O_{1j} \equiv$ the number of successes in column j (which is C_j in the current notation) and $O_{2j} \equiv$ the number of failures in column j (which is $r - C_j$ in the current notation). The contingency table would look like

	Treatment				
	1	2	$\cdots$	c	
Success	C_1	C_2	$\cdots$	C_c	N
Failure	$r - C_1$	$r - C_2$	$\cdots$	$r - C_c$	$rc - N$
	r	r	$\cdots$	r	rc

As you'll see on the homework, this test is less sensitive if the blocking carries important information.

2.3.2 McNemar's Test (see Conover Section 3.5)

In the case where $c = 2$, Cochran's test is equivalent to a test performed using a 2×2 contingency table, known as McNemar's test. When there are only two columns, the information in each block consists of whether the pair (X_{i1}, X_{i2}) is $(0,0)$, $(0,1)$, $(1,0)$, or $(1,1)$, and the important information is how many blocks of each kind we have. We're then dealing with a 2×2 contingency table

	$X_{i2} = 0$	$X_{i2} = 1$
$X_{i1} = 0$	a	b
$X_{i1} = 1$	c	d

The interpretation of this table is different from the usual two-way contingency table, though. If the treatments behave differently, b and c will differ from each other. The McNemar test statistic is

$$\frac{(b-c)^2}{b+c} \tag{2.34}$$

which is approximately $\chi^2(1)$ distributed if the treatments are equivalent.