Goodness of fit-tests for multinomial data

April 18, 2012

1 All parameters known

Suppose that $X_1, X_2, \ldots, X_k$ has a multinomial distribution with paramters n and $p_1, p_2, \ldots, p_k$. The expectation and variance of each X_i is then

$$E(X_i) = np_i, \qquad \mathrm{Var}(X_i) = np_i(1-p_i), \tag{1}$$

and the covariance between a given X_i and X_j is negative and equal to

$$\mathrm{Cov}(X_i, X_j) = -np_ip_j. \tag{2}$$

Thus,

$$\frac{X_i - np_i}{\sqrt{np_i(1-p_i)}} \tag{3}$$

is approximately $N(0,1)$. Furthermore, it is proved elsewhere that the statistic

$$D = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{4}$$

is approximately chi-square distribution with $k-1$ degrees of freedom, provided that all expectations $E(X_i) = np_i \geq 5$.

To test the goodness of fit of a given null hypothesis of the form

$$H_0 : p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_k = p_{k,0} \tag{5}$$

we can thus be based on the test statistic

$$D = \sum_{i=1}^{k} \frac{(X_i - np_{i,0})^2}{np_{i,0}}. \tag{6}$$

If the deviation of the observed values X_i from their respective expectations $np_{i,0}$ under H_0 is large, D will take a large value. We thus reject H_0 if D is larger than the upper α quantile of the chi-square distribution

$$D > \chi^2_{\alpha,k-1}. \tag{7}$$

Tests of null-hypotheses of this kind where we hypothesize that all p_i's have some particular value are rare. The so call Benford's law provide one example. This law typically applies to positive numerical quantities which follows for example a log-normal distribution and which vary across many orders of magnitude. For example, the brain size of different land mammals vary between 0.14 and 5712 grams, that is by more than 4 orders of magnitude. The law states that the first digit of such numbers follow a distribution where the probability that the first digit is equal to i is given by

$$p_i = \log_{10}(i+1) - \log_{10} i. \tag{8}$$

These probabilities can be computed in R as follows

```
> i <- 1:9
> p <- log10(i+1)-log10(i)
> p
[1] 0.30103000 0.17609126 0.12493874 0.09691001 0.07918125 0.06694679 0.05799195
[8] 0.05115252 0.04575749
```

So most brain sizes (31%) should have 1 as the first digit.

The brain size in gram of the 62 different land mammals (assignment 1) are

```
> print(mammals$brain)
 [1]   44.50   15.50    8.10  423.00  119.50  115.00   98.20    5.50   58.00
[10]    6.40    4.00    5.70    6.60    0.14    1.00   10.80   12.30    6.30
[19] 4603.00    0.30  419.00  655.00    3.50  115.00   25.60    5.00   17.50
[28]  680.00  406.00  325.00   12.30 1320.00 5712.00    3.90  179.00   56.00
[37]   17.00    1.00    0.40    0.25   12.50  490.00   12.10  175.00  157.00
[46]  440.00  179.50    2.40   81.00   21.00   39.20    1.90    1.20    3.00
[55]    0.33  180.00   25.00  169.00    2.60   11.40    2.50   50.40
```

The number of brains sizes starting with the digit 1, 2, ..., 9 are

```
> x <- c(24,7,7,9,7,5,0,2,1)
```

Under the null hypothesis that Benford's law applies these counts are a sample from a multinomial distribution with parameters $n = 62$ and $p_1, p_2, \ldots, p_9$ given by (8). Having computed these probabilities and stored the result in the vector `p`, the chi-square test of this null hypothesis based on (6) can be done in R using the `chisq.test` function

```
> chisq.test(x,p=p)

Chi-squared test for given probabilities

data:  x
X-squared = 10.7747, df = 8, p-value = 0.2148

Warning message:
In chisq.test(x, p = p) : Chi-squared approximation may be incorrect
```

The observed value of chi-square statistic is close to its expected value of 8 and the large p-value indicates that we can not reject the reject the null hypothesis.

R gives a warning message because some of the expected values are smaller than 5. These expected values are available in the `$expected` component of the list returned by `chisq.test` (see the help page)

```
> chisq.test(x,p=p)$expected
[1] 18.663860 10.917658  7.746202  6.008421  4.909237  4.150701  3.595501
[8]  3.171456  2.836964
```

We see that the observed values are fairly close to the expected values based on Benford's law.

2 Unknown parameters

Rather than having some hypothesized value for all the probabilities $p_1, p_2, \ldots, p_k$, we usually want to test the goodness-of-fit of the null hypotheses H_0 that there is a particular mathematical

relationship between the p_i's. Such relationships can in general be represented by assuming that $p_1, p_2, \ldots, p_k$ are functions of a smaller number of s parameters, that is, that

$$\begin{aligned} p_1 &= p_1(\theta_1, \theta_2, \ldots, \theta_s) \\ p_2 &= p_2(\theta_1, \theta_2, \ldots, \theta_s) \\ &\vdots \\ p_k &= p_k(\theta_1, \theta_2, \ldots, \theta_s) \end{aligned} \tag{9}$$

We shall see that we can sometimes easily and sometimes only by numerical methods obtain maximum likelihood estimates of the unknown s parameters $\theta_1, \theta_2, \ldots, \theta_s$ from the observed counts $X_1, X_2, \ldots, X_k$. Either way, the following important theorem applies. Suppose that the maximum likelihood estimators of $\theta_1, \theta_2, \ldots, \theta_s$ are $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$. Under H_0, the test statistic

$$D = \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}. \tag{10}$$

where

$$\hat{p}_i = p_i(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s) \tag{11}$$

is then approximately chi-square distributed with $k - 1 - s$ degrees of freedom provided that $n\hat{p}_i \geq 5$ for all categories i.

2.1 Contingency tables

Testing for independence in a an 2×2 contingency table is a special case of a test of a null hypothesis of the form (9). Suppose that we categorize a sample of $n = 61$ patients as follows.

	Healed	Not Healed	Total
Pirenzepine	23	7	30
Trithiozine	18	13	31
Total	41	20	61

The counts of number of patients in each of the 4 categories now follow a multinomial distribution. The null hypothesis of independence between medical treatment and healing outcome means that the four multinomial probabilities are given by

	Healed	Not Healed	Marginal prob.
Pirenzepine	$p_{11} = pq$	$p_{12} = (1-p)q$	q
Trithiozine	$p_{21} = p(1-q)$	$p_{22} = (1-p)(1-q)$	$1-q$
Marginal prob.	p	$1-p$	

that is, four different functions of $s = 2$ parameters p and q.

Maximum likelihood estimates of p and q can be found be first realising that the total count of patients which are healed is binomially distributed with parameters p and n under H_0. Hence the maximum likelihood estimate of p is $\hat{p} = 41/61 = 0.6721$. Similarly, the maximum likelihood estimate of q becomes $\hat{q} = 30/61 = 0.4918$.

Based on the maximum likelihood estimates of the $s = 2$ parameters p and q we can compute the corresponding maximum likelihood estimates $\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{21}, \hat{p}_{22}$ of the probabilites of observations in the four different categories by applying the functions given in the above table on $\hat{p}$ and $\hat{q}$.

According to (9), given $k = 4$ categories with associated probabilites being functions of $s = 2$ parameters, the statistic

$$D = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(X_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} \tag{12}$$

is now chi-square distributed with $k-1-s=1$ degree of freedom.

R carries out tests of this kind if the first argument to `chisq.test` is a matrix or table containing the counts.

```
> x <- matrix(c(23,7,18,13),2,2,byrow=T)
> chisq.test(x)

Pearson's Chi-squared test with Yates' continuity correction

data:  x
X-squared = 1.6243, df = 1, p-value = 0.2025

> chisq.test(x)$exp
         [,1]      [,2]
[1,] 20.16393  9.836066
[2,] 20.83607 10.163934
```

A generalisation of this test to a $r \times c$ contingency table would involve $s=(r-1)+(c-1)$ parameters and a total number of $k=rc$ categories. The associated chi-square statistic in this case thus have

$$k-1-s=rc-1-(r-1)+(c-1)=rc-r-c+1=(r-1)(c-1) \tag{13}$$

degrees of freedom.

Similarly, the goodness-of-fit chi-square test statistic for complete independence for a three-way $r \times c \times t$ contingency table would have $rct-r-c-t+2$ degrees of freedom. For three-way tables many other hypotheses are of interest, however, and can be tested using addon R-packages.

2.2 Testing Hardy-Weinberg equilibrium

2.2.1 Diallelic loci

Consider a population of a diploid organism and let P_{AA}, P_{Aa}, P_{aa} be the genotype frequencies of the different genotypes at a particular diallelic locus. If we sample n individuals from the population, the counts X_{AA}, X_{Aa}, X_{aa} of number of individuals of the different genotypes in the sample will follow a multinomial distribution with parameters n and P_{AA}, P_{Aa}, P_{aa}.

The population is said to be in Hardy-Weinberg equilibrium at a diallelic locus if there is a certain relationship between the genotype frequencies, namely that all the frequencies are the functions

$$\begin{aligned} P_{AA} &= p^2, \\ P_{Aa} &= 2p(1-p), \\ P_{aa} &= (1-p)^2 \end{aligned} \tag{14}$$

of a single parameter p being the allele frequency of allele A.

A goodness-of-fit test of this null hypothesis can again be based on (10) since the probabilities of observations in the $k=3$ categories again are functions of a smaller number of $s=1$ parameters.

It can be shown (see assignment 5) that the maximum likelihood estimator of the allele frequency p in the population under H_0 is simply equal to the frequency of the allele in the sample, that is, the number of alleles of type A in divided by the total number of alleles (two times the sample size),

$$\hat{p}=\frac{2X_{AA}+X_{Aa}}{2n} \tag{15}$$

For example, if we observe 51, 42 and 7 individuals of genotype AA, Aa and aa in a sample of $n = 100$ individuals, the maximum likelihood estimate of the allele frequency of A is $\hat{p} = (2 \cdot 51 + 42)/200 = 0.72$.

We can carry out the test as follows in R. Letting the three elements of the vector `X` represent the number of individuals of genotype AA, Aa and aa in the sample, $\hat{p}$ can be computed as follows.

```
> X <- c(51,42,7)
> n <- sum(X)
> phat <- (2*X[1]+X[2])/(2*n)
> phat
[1] 0.72
```

The correspoding maximum likelihood estimates of the genotype frequencies are given by

```
> Phat <- c(phat^2,2*phat*(1-phat),(1-phat)^2)
> Phat
[1] 0.5184 0.4032 0.0784
```

The expected numbers of each genotype $n\hat{P}_{AA}, n\hat{P}_{Aa}, n\hat{P}_{aa}$ become

```
> n*Phat
[1] 51.84 40.32  7.84
```

which are not far from the observed values. The observed value of the test statistic based on (10) is

```
> D <- sum((X-n*Phat)^2/(n*Phat))
> D
[1] 0.1736111
```

which is below the expected value of $k - 1 - s = 3 - 1 - 1 = 1$ degree of freedom. The P-value of the test is

```
> pchisq(D,df=1,lower.tail=F)
[1] 0.6769222
```

so we can clearly not reject the null hypothesis that the population is in Hardy-Weinberg equilibrium.

2.2.2 More than 2 alleles

This approach can easily be extended to test for Hardy-Weinberg equilibrium at loci with more than 2 alleles. With three alleles we have $k = 6$ genotypes,

$$\begin{aligned} A_1A_1, A_1A_2, &A_1A_3 \\ A_2A_2, &A_2A_3, \\ &A_3A_3. \end{aligned} \tag{16}$$

Under the null hypothesis of Hardy-Weinberg equilibrium, the population genotype frequencies of these can all be written as functions of at most $s = 2$ parameters, say the allele frequncies p_1 and p_2 of allele A_1 and A_2 since $p_3 = 1 - p_1 - p_2$. The frequency of genotype A_2A_3 is for example

$$P_{23} = 2p_2p_3 = 2p_2(1 - p_1 - p_2). \tag{17}$$

Again, the maximum likelihood estimates of the allele frequencies are equal to their respective sample frequencies. From these maximum likelihood estimates, the corresponding

maximum likelihood estimates of all 6 genotype frequencies, the associated expected values and the observed value of the test statistic can be computed.

Under H_0, this test statistic is again chi-square distributed with $k - 1 - s = 6 - 1 - 2 = 3$ degrees of freedom.

2.2.3 Incomplete data due to dominance (bolk 10)

Blood type in humans is determined by a triallelic loci with two dominant alleles A, B and one recessive allele O as follows

i	Genotype	Phenotype	Probability p_i	Count X_i
1	AA, A0	A	$p_A^2 + 2p_Ap_O$	44
2	BB, B0	B	$p_B^2 + 2p_Bp_O$	27
3	AB	AB	$2p_Ap_B$	4
4	00	0	p_O^2	88

Note that the probabilities of observing different phenotypes becomes equal to the sums of the underlying frequencies of possible genotypes.

The observed counts in the rightmost column is a sample from an African population (Crow 1986, p. 24). The null hypothesis that this population is in Hardy-Weinberg equilibrium can again be tested based on the general theorem (10) since the the probabilities of the $k = 4$ observable phenotypes can all be written as functions of $s = 2$ parameters, say the allele frequencies of alleles A and B, p_A and p_B. Provided that we can compute the maximum likelihood estimates of p_A and p_B the resulting chi-square distributed test statistic will have $k - 1 - s = 4 - 1 - 2 = 1$ degrees of freedom accodring to (10).

The difficulty lies in computing these maximum likelihood estimates. If we treat p_A and p_B as the unknown parameters, and keep in mind that $p_1, p_2, \ldots, p_4$ are functions of p_A and p_B the likelihood function for the data is

$$L(p_A, p_B) = \frac{n!}{x_1!x_2!x_3!x_4!} \prod_{i=1}^{4} p_i^{x_i} \tag{18}$$

and the log likelihood is

$$\ln L(p_A, p_B) = \ln n! - \sum \ln x_i! + \sum_{i=1}^{4} x_i \ln p_i. \tag{19}$$

Substituting the expressions for each p_i into (19) and setting the partial derivatives with respect to p_A and p_B equal to zero leads to a set of two non-linear equations which have no analytic solution.

The likelihood function can be maximised numerically, however, as follows. We first define the two following functions.

```
multinomialprobs <- function(par) {
  pA <- par[1]
  pB <- par[2]
  p0 <- 1-pA-pB
  c(pA^2 + 2*pA*p0, pB^2 + 2*pB*p0, 2*pA*pB, p0^2)
}

lnL <- function(par,x) {
  n <- sum(x)
  -dmultinom(x,prob=multinomialprobs(par),log=T)
}
```

The first function computes the probabilities of the four different phenotypes for given values of the allele frequencies p_A and p_B (represented by the vector argument `par`). For example, for $p_A = 0.5$ and $p_B = 0$ (and hence $p_O = 0.5$) we get, the probabilities of the four different phenotypes are

```
> multinomialprobs(c(.5,0))
[1] 0.75 0.00 0.00 0.25
```

The second function computes the negative log likelihood of the observed data (represented by the second vector argument `x`) given particular values of p_A and p_B (represented by the first argument, the vector `par`).

We can now find the maximum likelihood likelihood estimates of p_A and p_B by minimising the negative log likelihood function numerically using the `optim` function.

```
> x <- c(44,27,4,88)
> fit <- optim(c(.25,.25),lnL,x=x)
> fit
$par
[1] 0.1604618 0.1003531

$value
[1] 6.917786

$counts
function gradient
      65       NA

$convergence
[1] 0

$message
NULL
```

The maximum likelihood estimates are thus $\hat{p}_A = 0.16$ and $\hat{p}_B = 0.10$.

The corresponding estimates of the phenotype probabilities become

```
> Phat <- multinomialprobs(fit$par)
> Phat
[1] 0.26296987 0.15842973 0.03220566 0.54639474
```

and the expected number of individuals of each phenotype

```
> n <- sum(x)
> n*Phat
[1] 42.864089 25.824047  5.249522 89.062342
```

which, again, is fairly close to the observed counts in the above table. The observed value of the chi-square test statistic of the goodness-of-fit test becomes

```
> D <- sum((x-n*Phat)^2/(n*Phat))
> D
[1] 0.3937418
```

which gives a P value of

```
> pchisq(D,df=1,lower.tail=F)
[1] 0.53
```

Hence, we can not reject the null hypothesis that the population is in Hardy-Weinberg equilibrium.