Goodness of Fit

Many statistical quantities derived from data samples are found to follow the Chi-squared distribution.

Hence we can use it to test whether a population fits a particular theoretical probability distribution.

Multinomial Goodness of Fit

A population is called multinomial if its data is categorical and belongs to a collection of discrete non-overlapping classes.

The null hypothesis for goodness of fit test for multinomial distribution is that the observed frequency f_i is equal to an expected count e_i in each category. It is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

$$\chi^2 = \frac{\sum_i (f_i - e_i)^2}{e_i}$$

Example

In the built-in data set survey, the Smoke column records the survey response about the student's smoking habit.

As there are exactly four proper response in the survey: "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never", the Smoke data is multinomial.

It can be confirmed with the levels function in R.

```
> library(MASS)                # load the MASS package
> levels(survey$Smoke)
[1] "Heavy" "Never" "Occas" "Regul"
```

As discussed in the tutorial Frequency Distribution of Qualitative Data, we can find the frequency distribution with the table function.

```
> smoke.freq = table(survey$Smoke)
> smoke.freq

Heavy Never Occas Regul
   11   189    19    17
```

Problem

Suppose the campus smoking statistics is as below. Determine whether the sample data in survey supports it at .05 significance level.

Heavy	Never	Occas	Regul
4.5%	79.5%	8.5%	7.5%

Solution

We save the campus smoking statistics in a variable named smoke.prob. Then we apply the chisq.test function and perform the Chi-Squared test.

```
> smoke.prob = c(.045, .795, .085, .075)
> chisq.test(smoke.freq, p=smoke.prob)

        Chi-squared test for given probabilities

data:  smoke.freq
X-squared = 0.1074, df = 3, p-value = 0.991
```

Answer

As the p-value 0.991 is greater than the .05 significance level, we do not reject the null hypothesis that the sample data in survey supports the campus-wide smoking statistics.

Chi-squared Test of Independence

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

Assume f_{ij} is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume e_{ij} to be the corresponding expected count if x and y are independent.

The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

$$\chi^2 = \frac{\sum_{i,j} (f_{ij} - e_{ij})^2}{e_{ij}}$$

Example

In the built-in data set survey, the Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

We can tally the students smoking habit against the exercise level with the table function in R.

The result is called the contingency table of the two variables.

```
> library(MASS)                          # load the MASS package
> tbl = table(survey$Smoke, survey$Exer)
> tbl                                    # the contingency table
```

```
       Freq None Some
Heavy     7    1    3
Never    87   18   84
Occas    12    3    4
Regul     9    1    7
```

Problem

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

Solution

We apply the chisq.test function to the contingency table tbl, and found the p-value to be 0.4828.

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  table(survey$Smoke, survey$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(table(survey$Smoke, survey$Exer)) :
  Chi-squared approximation may be incorrect
```

Answer

As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

Enhanced Solution

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl, and save it in a new table named ctbl. Then we apply the chisq.test function against ctbl instead.

```
> ctbl = cbind(tbl[,"Freq"], tbl[,"None"] + tbl[,"Some"])
> ctbl
      [,1] [,2]
Heavy    7    4
Never   87  102
Occas   12    7
Regul    9    8

> chisq.test(ctbl)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571
```

We used exploratory techniques to identify 92 stars from the Hipparcos data set that are associated with the Hyades. We did this based on the values of right ascension, declination, principal motion of right ascension, and principal motion of declination. We then excluded one additional star with a large error of parallax measurement:

```
#> hip <- read.table("http://astrostatistics.psu.edu/datasets/HIP_star.dat",
#+ header=T, fill=T)
> hip <- read.table("HIP_star.dat", header=T, fill=T)
> attach(hip)
> filter1 <-  (RA>50 & RA<100 & DE>0 & DE<25)
> filter2 <- (pmRA>90 & pmRA<130 & pmDE>-60 & pmDE< -10)
> filter  <-  filter1 & filter2 & (e_Plx<5)
> sum(filter)

> bvcat <- cut(color, breaks=c(-Inf,.5,.75,1,Inf))

> boxplot(Vmag~bvcat, varwidth=T,
+ ylim=c(max(Vmag),min(Vmag)),
+ xlab=expression("B minus V"),
+ ylab=expression("V magnitude"),
+ cex.lab=1.4, cex.axis=.8)
```

The cut values for bvcat are based roughly on the quartiles of the B minus V variable. We have created, albeit artificially, a second categorical variable ("filter", the Hyades indicator, is the first). Here is a summary of the dataset based only on these two variables:

```
> table(bvcat,filter)
```

Note that the Vmag variable is irrelevant in the table above.

To perform a chi-squared test of the null hypothesis that the true population proportions falling in the four categories are the same for both the Hyades and non-Hyades stars, use the chisq.test function:

```
> chisq.test(bvcat,filter)
```

Since we already know these two groups differ with respect to the B.V variable, the result of this test is not too surprising. But it does give a qualitatively different way to compare these two distributions than simply comparing their means.

The p-value produced above is based on the fact that the chi-squared statistic is approximately distributed like a true chi-squared distribution (on 3 degrees of freedom, in this case) if the null hypothesis is true. However, it is possible to obtain exact p-values, if one wishes to calculate the chi-squared statistic for all possible tables of counts with the same row and column sums as the given table. Since this is rarely practical computationally, the exact p-value may be approximated using a Monte Carlo method (just as we did earlier for the permutation test). Such a method is implemented in the chisq.test function:

```
> chisq.test(bvcat,filter,sim=T,B=50000)
```

The two different p-values we just generated a numerically similar but based on entirely different mathematics. The difference may be summed up as follows: The first method produces the exact value of an approximate p-value, whereas the second method produces an approximation to the exact p-value!

The test above is usually called a chi-squared test of homogeneity. If we observe only one sample, but we wish to test whether the categories occur in some pre-specified proportions, a similar test (and the same R function) may be applied. In this case, the test is usually called a chi-squared test of goodness-of-fit.

Exercise

Use the data provided, two datasets, one with a total of m = 290 observations the other with 386 measurements. The former is of flux densities measured at random positions in the sky; the latter of flux densities at the positions of a specified set of galaxies. Using the chi-square test, examine the hypothesis that there is excess flux density at the non-random positions.