R code for testing Goodness of Fit, Independence and Homogeneity

Goodness of Fit:

Example: (Roses) When crossing certain types of red and white roses, one obtains red, white and pink roses. Theory predicts that the proportion of red to white to pink roses is like 3:2:2. Test the plausibility of this theory when out of a sample of 80 crosses, 35 are red, 31 are white and 14 are pink. (Note: Sampling design is multinomial sampling of one variable and we test to see if the multinomial probabilities are equal to some specified values)

```
> chisq.test(c(35,31,14), p=c(3,2,2)/7)

        Chi-squared test for given probabilities

data:  c(35, 31, 14)
X-squared = 6.3479, df = 2, p-value = 0.04184
```

Conclusion: At a 5% significance level, the data provide sufficient evidence (P-value = 0.0418) that the proportion of red to white to pink roses is different from 3:2:2.

Independence:

Example: (Hair and Eye Color) In a sample of 65 students, we recorded the hair color (categories blond, brown, dark) and eye color (categories bright, dark). The table below summarizes the counts. Null hypothesis: Hair and Eye color are independent. Alternative Hypothesis: Hair and eye color are associated. (Note: Sampling design is multinomial, where two categorical responses are recorded.)

```
> table1=matrix(c(12,2,8,25,6,12),ncol=3)
> colnames(table1)=c("blond","brown","dark")
> rownames(table1)=c("bright","dark")
> table1
       blond brown dark
bright    12     8    6
dark       2    25   12

> chisq.test(table1)

        Pearson's Chi-squared test

data:  table1
X-squared = 15.938, df = 2, p-value = 0.000346
```

Conclusion: There is sufficient evidence (P-value=0.0003) that hair and eye color of students are associated.

Follow-up: What is contributing to the association?

```
> chisq.test(table1)$residuals
           blond      brown       dark
```

```
bright  2.704494 -1.431253 -0.4472136
dark   -2.208210  1.168613  0.3651484
```

There are many more bright-eyed and blond-haired students than would be predicted under independence of eye and hair color (this cell shows the largest residual.) Conversely, there are many fewer students with dark eyes and blond hair than would be predicted under independence.

Homogeneity (Comparing proportions across two groups):

Example : (Egg) We asked n1=25 females (group 1) and n2=17 males (group 2) how they preferred their Sunday morning breakfast egg (Sunny Side Up, Over Easy or Scrambled). The data are summarized in the table below. Is the distribution of egg preference the same for males and females? (Are the proportions homogeneous across the two groups?) Null hypothesis: distributions are the same for females and males; Alternative Hypothesis: Distributions are not the same (Note: Sampling design is product multi(=bi) nomial.)

```
> table2=matrix(c(5,9,12,3,7,5),ncol=3)
> colnames(table2)=c("Sunny","Over Easy","Scrambled")
> rownames(table2)=c("Females","Males")
> table2
        Sunny Over Easy Scrambled
Females     5        12         7
Males       9         3         5
> chisq.test(table2)

        Pearson's Chi-squared test

data:  table2
X-squared = 5.8516, df = 2, p-value = 0.05362

Warning message:
In chisq.test(table2) : Chi-squared approximation may be incorrect
```

When we don't trust the validity of the asymptotic approximation (see Warning), the permutation approach we discussed in class is actually implemented in the chisq.test() function and safer to use here (you can also use it for the previous example about hair and eye color):

```
> chisq.test(table2,simulate.p.value = TRUE, B = 10000)

        Pearson's Chi-squared test with simulated p-value (based on
10000
        replicates)

data:  table2
X-squared = 5.8516, df = NA, p-value = 0.0611
```

Conclusion: We have insufficient evidence (P-value = 0.0611) to conclude that the distribution of egg preference is different for females and males. (Note: The exact test here is based on the proportion of sampled tables (with the same margin) with X2 statistics at least as large or larger than the observed test statistic. Another option is to use fisher.test() which gets the exact P-value using the exact null table probabilities.)