Exercises

A taste test

In statistics, contingency tables are used to record and analyse the relationship between two or more categorical variables.

For this example, a taste test was devised to see what kind of food different breeds of cat prefer. The cats, food and preferences are shown in this contingency table:

Food ***Cat***	***Whiskers***	***Kit-E-Feast***	***Chump***	***Mouseful***	***Kit-Cat***
Persian	0	2	2	3	2
Manx	0	6	1	7	0
Pole	3	2	1	3	0

The entries in the table say how many cats of each breed preferred each food (i.e. the frequency with which each food was preferred).

NAG routine G01AF can be used to perform an analysis of the contingency table. In particular, G01AF will return a value for the chi-squared statistic. This can be used to test the *null hypothesis* that there is no association between cat breed and food preference, against an alternative hypothesis that there *is* an association.

Use G01AF to compute the chi-squared statistic for this data set. G01AF also returns a value `NDF`, which is the number of degrees of freedom for the chi-squared test. You can convert these two values to a *p value* by a call of NAG routine G01EC (choose the upper tail probability). Note that a low p value indicates evidence against the null hypothesis. The usual test is whether or not the p value is significant at the 5% level, i.e. if $p < 0.05$ then the null hypothesis is rejected.

Judging by the p value you obtain, what conclusion can you draw about our cats?

An alternative (Monte Carlo) way of calculating the chi-squared p value

As it happens, our contingency table contains many low frequencies. The chi-squared approximation is often inappropriate in these cases, and therefore the p value may be unreliable. An alternative way of getting an estimate for the p value when working on a table with many low frequencies is to use a Monte Carlo method.

Calculate row and column totals for the contingency table, and remember the value of `CHIS` from the call of G01AF (call this value `OCHIS`). Then implement the following pseudo code in MATLAB:

```
CNT = 0
NSIM = 2000
for i = 1, NSIM do
  call G05QD
  call G01AF on the table generated by G05QD.
  If CHIS > OCHIS, set CNT = CNT + 1
end
Compute new p value as (CNT+1) / (NSIM+1)
```

Here, the routine G05QD is used to randomise the entries in the contingency table (hint: use `MODE=2`). The method outlined above is designed to find out how many tables with the same row and column totals are more extreme than the original table. Keeping a count of how many extreme tables are detected lets us get a Monte Carlo approximation to the p value associated with the chi-squared statistic.

So - is the new p value different to the original value you obtained? And do you draw the same conclusion about cat food preferences?