What kinds of tests can be carried out for such data ?

1. Tests about $p_1, \ldots, p_k$
   - $H_0 \ : \ p_1 = \cdots = p_k = 1/k$
   - $H_0 \ : \ p_1, \ldots, p_k$ determined by some parametric distribution (Normal, Poisson etc.)
2. Tests about the factors A and B
   - are A and B dependent ?
   - i.e. does classification by A influence classification by B.

Chi-Squared Test

For one-way tables: suppose that a null hypothesis **completely specifies** $p_1, \ldots, p_k$, that is, we have H_0 of the form

$$H_0 \; : \; p_1 = p_1^{(0)}, \ldots, p_k = p_k^{(0)}$$

where $p_1^{(0)}, \ldots, p_k^{(0)}$ are fixed probabilities. For example, for $k = 3$,

$$H_0 \; : \; p_1 = p_2 = p_3 = 1/3$$

or

$$H_0 \; : \; p_1 = 1/2, p_2 = p_3 = 1/4$$

To test this hypothesis against the general alternative hypothesis

$$H_a \ : \ H_0 \text{ not true.}$$

we use the test statistic

$$X^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$$

If H_0 is true,

$$X^2 \stackrel{\cdot}{\sim} \text{Chi-squared}(k-1).$$

that is, X^2 is approximately distributed as Chi-squared$(k-1)$.

In this formula

- n_i is the **observed** count in cell i
- $np_i^{(0)}$ is the **expected** count in cell i if H_0 is **true**.

Sometimes the formula is written

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count, and E_i is the expected count.

If

$$X^2 > \text{Chisq}_\alpha(k-1)$$

then we reject H_0 at the α significance level, where $\text{Chisq}_\alpha(k-1)$ is the $1-\alpha$ (right-hand) tail critical value of the Chi-squared distribution with $k-1$ degrees of freedom.

This method can be extended in the one-way case to test distribution assumptions, that is, for example

$$H_0 \ : \ \text{Data Normally distributed}$$

or

$$H_0 \ : \ \text{Data Poisson distributed}$$

Unfortunately this facility is not available in SPSS; direct calculation is possible but involved.

For the **two-way** table, we can test the hypothesis

H_0 : Factor A and Factor B levels are assigned independently

that is, classification by factor A is independent of classification by factor B. We use the same test statistic that can be rewritten

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - \widehat{n}ij)^2}{\widehat{n}ij}$$

where

$$\widehat{n}_{ij} = \frac{n_{i.}n_{.j}}{n} \qquad n_{i.} = \sum_{j=1}^{c} n_{ij} \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}.$$

The terms $n_{i.}$ and $n_{.j}$ are the row and column totals for row i and column j respectively.

If H_0 is true

$$X^2 \mathrel{\dot\sim} \text{Chi-squared}((r-1)(c-1))$$

i.e. the degrees of freedom quantity is $(r-1)(c-1)$.
Otherwise the test proceeds as before: we check whether

$$X^2 > \text{Chisq}_\alpha((r-1)(c-1))$$

and if so, we reject H_0.

Example: DNA Sequence Data.

Counts of Nucleotides A,C,G,T in a genomic segment related to the breast cancer gene BRCA2.

Example: Eye and Hair Colour Data.

The assignment of hair and eye colour in a sample of humans

See handout.

Note: For the Chi-squared test to be valid, we need the expected cell counts

$$np_i^{(0)} \qquad i = 1, \ldots, k$$

or

$$\widehat{n}_{ij} \qquad i = 1, \ldots, r, j = 1, \ldots, c$$

to be sufficiently large. The convention is to require the expected value to be greater than **five**.