Chi-squared Tests for Categorical Data

In a **multinomial** experiment, the independent experimental units are classified to one of k categories determined by the levels of a discrete factor. Let $n_1, n_2, \ldots, n_k$ be the counts of the numbers of experimental units in the k categories, where $n_1 + n_2 + \cdots + n_k = n.$.

The probability that an experimental unit is classified to category i is p_i, for $i = 1, \ldots, k$, so that

$$p_1 + p_2 + \cdots + p_k = 1.$$

- The **one-way** classification table can be displayed as follows:

Category	1	2	$\cdots$	k
Count	n_1	n_2	$\cdots$	n_k
Probability	p_1	p_2	$\cdots$	p_k

  We can test a hypothesis H_0 that fully specifies $p_1, \ldots, p_k$, for example

$$H_0 : p_1 = p_1^{(0)}, p_2 = p_2^{(0)}, \ldots, p_k = p_k^{(0)}$$

  so that, for $k = 3$, we might have

$$H_0 : p_1 = p_2 = p_3 = 1/3 \qquad \text{or} \qquad H_0 : p_1 = 1/2, p_2 = p_3 = 1/4.$$

  We use the test statistic

$$X^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} = \sum_{i=1}^{k} \frac{(\text{Observed Count in Cell } i - \text{Expected Count in Cell } i)^2}{\text{Expected Count in Cell } i}$$

  We sometimes write $\widehat{n}_i = np_i^{(0)}$. If H_0 is true, $X^2 \stackrel{\cdot}{\sim} \text{Chi-squared}(k-1)$.

- The **two-way** classification table can also be constructed to represent the cross-classification for two discrete factors A and B with r and c levels respectively.

		Factor B			
		1	2	$\cdots$	c
Factor A	1	n_{11}	n_{12}	$\cdots$	n_{1c}
	2	n_{21}	n_{22}	$\cdots$	n_{2c}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

  To test the hypothesis

$$H_0 \ : \ \text{Factor A and Factor B levels are assigned independently}$$

  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

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

EXAMPLE 1: DNA Sequence Data
The counts of the numbers of nucleotides (A,C,G,T) in the DNA sequence of the cancer-related gene BRCA 2 are presented in the table below.

Category	1	2	3	4	Total
Nucleotide	A	C	G	T	
Count	38514	24631	25685	38249	127079

so that $k = 4$. To test the hypothesis

$$H_0 \ : \ p_1 = p_2 = p_3 = p_4 = 1/4$$

We use the one-way table chi-squared test: here

$$\widehat{n}_i = np_i^{(0)} = \frac{127079}{4} = 31769.75$$

so the test statistic is

$$\begin{aligned} X^2 &= \frac{(38514 - 31769.75)^2}{31769.75} + \frac{(24631 - 31769.75)^2}{31769.75} + \frac{(25685 - 31769.75)^2}{31769.75} + \frac{(38249 - 31769.75)^2}{31769.75} \\ &= 5522.597 \end{aligned}$$

We compare this with the Chi-squared$(k-1) \equiv$ Chi-squared(3) distribution. From McClave and Sincich, p. 898,

$$\text{Chisq}_{0.05}(3) = 7.815 < X^2$$

so H_0 is **rejected**.

EXAMPLE 2: Eye and Hair Colour Data
The table below contains counts of the number of people in a study with a combination of eye and hair colour.

		Hair				
		Black	Brunette	Red	Blonde	$n_{i.}$
Eyes	Brown	68	119	26	7	220
	Blue	20	84	17	94	215
	Hazel	15	54	14	10	93
	Green	5	29	14	16	64
	$n_{.j}$	108	286	71	127	592

so $r = c = 4$. To test the hypothesis

$$H_0 \ : \ \text{Eye and Hair colour are assigned independently}$$

we use the X^2 statistic

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

Here, for example, for $i = 2$ and $j = 3$

$$\widehat{n}_{23} = \frac{n_{2.} \times n_{.3}}{n} = \frac{215 \times 71}{592} = 25.785.$$

In fact, on complete calculation, we find that

$$X^2 = 138.2898.$$

We compare this with the Chi-squared$((r-1)(c-1)) \equiv$ Chi-squared(9) distribution. From McClave and Sincich, p. 898,

$$\text{Chisq}_{0.05}(9) = 16.919 < X^2$$

so H_0 is **rejected**

Chi-Squared test for the nucleotide count data

Use

Analyze → Nonparametric Tests → Chi-Square

pulldown menus.

For the test of

$$H_0 : p_1 = p_2 = p_3 = p_4 = 1/4$$

First null hypothesis

Nucleotide

	Observed N	Expected N	Residual
A	38514	31769.8	6744.3
C	24631	31769.8	-7138.8
G	25685	31769.8	-6084.8
T	38249	31769.8	6479.3
Total	127079		

Chi-squared Statistic = 5522.597

Test Statistics

	Nucleotide
Chi-Square(a)	5522.597
df	3
Asymp. Sig.	.000

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 31769.8.

p-value < 0.001

For the test of

$$H_0 : p_1 = p_4 = 0.3 \quad p_2 = p_3 = 0.2$$

Second null hypothesis

Nucleotide

	Observed N	Expected N	Residual
A	38514	38123.7	390.3
C	24631	25415.8	-784.8
G	25685	25415.8	269.2
T	38249	38123.7	125.3
Total	127079		

Chi-squared Statistic = 31.492

Test Statistics

	Nucleotide
Chi-Square(a)	31.492
df	3
Asymp. Sig.	.000

p-value < 0.001

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 25415.8.

Chi-Squared test for the Hair and Eye colour count data

Use

Analyze → Descriptive Statistics → Crosstabs

pulldown menus.

For the test of

H_0 : Hair and Eye colour are assigned independently

Eye Colour * Hair Colour Crosstabulation

Count

		Hair Colour				Total
		Black	Brown	Red	Blond	
Eye Colour	Brown	68	119	26	7	220
	Blue	20	84	17	94	215
	Hazel	15	54	14	10	93
	Green	5	29	14	16	64
Total		108	286	71	127	592

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	138.290(a)	9	.000
Likelihood Ratio	146.444	9	.000
Linear-by-Linear Association	28.292	1	.000
N of Valid Cases	592		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 7.68.

p-value < 0.001

Chi-square statistic = 138.290

Note the comment returned by SPSS: The chi-squared test is not appropriate if any of the cells in the table have expected count less than 5 under the null hypothesis.

In this case, there is no problem as the cell counts are large enough.