

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 $n_1, n_2, \dots, n_k$ be the counts of the numbers of experimental units in the k categories, where $n_1 + n_2 + \dots + n_k = n$.

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

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

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

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

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

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

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

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

We use the test statistic

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

We sometimes write $\hat{n}_i = np_i^{(0)}$. If H_0 is true, $X^2 \approx \chi_{k-1}^2$.

- 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	$\dots$	c
Factor A	1	n_{11}	n_{12}	$\dots$	n_{1c}
	2	n_{21}	n_{22}	$\dots$	n_{2c}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	r	n_{r1}	n_{r2}	$\dots$	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

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

where

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

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

$$X^2 \approx \chi_{(r-1)(c-1)}^2$$

EXAMPLE 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

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

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 $\chi_{k-1}^2 \equiv \chi_3^2$ distribution. From McClave and Sincich tables,

$$\chi_{0.05,3}^2 = 7.815 < X^2$$

so H_0 is **rejected**.

EXAMPLE 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				$n_{i.}$
		Black	Brunette	Red	Blonde	
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

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

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

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

In fact, on complete calculation, we find that

$$X^2 = 138.2898.$$

We compare this with the $\chi^2_{(r-1)(c-1)} \equiv \chi^2_9$ distribution. From McClave and Sincich,

$$\chi^2_{0.05,9} = 16.919 < X^2$$

so H_0 is **rejected**

```
> #DNA data Chisq Test
> DNA.data<-c(38514,24631,25685,38249)
> DNA.fitted<-rep(1/4,4)*sum(DNA.data)
> (Xsq<-sum((DNA.data-DNA.fitted)^2/DNA.fitted))

+ [1] 5522.6

> (pval<-1-pchisq(Xsq,df=3))

+ [1] 0

> #DNA data Chisq Test: alternative null
> DNA.data<-c(38514,24631,25685,38249)
> DNA.fitted<-c(0.3,0.2,0.2,0.3)*sum(DNA.data)
>
> (Xsq<-sum((DNA.data-DNA.fitted)^2/DNA.fitted))

+ [1] 31.4923

> (pval<-1-pchisq(Xsq,df=3))

+ [1] 6.69513e-07

> #####
> #Hair & Eye colour
> xtabs(count~eye+hair,data=HairEyes)

+      hair
+ eye    Black Brown Red  Blond
+ Brown    68   119  26     7
+ Blue     20    84  17    94
+ Hazel    15    54  14    10
+ Green     5    29  14    16
```

```

> chisq.test(xtabs(count~eye+hair,data=HairEyes))

+
+ Pearson's Chi-squared test
+
+ data:  xtabs(count ~ eye + hair, data = HairEyes)
+ X-squared = 138.3, df = 9, p-value <2e-16

> #Direct calculation
> HairEyes.table<-xtabs(count~eye+hair,data=HairEyes)
> HE.rowtots<-apply(HairEyes.table,1,sum)
> HE.col tots<-apply(HairEyes.table,2,sum)
>
> #Form the fitted values from the row and column totals
> HE.fitted<-outer(HE.rowtots,HE.col tots,"*")/sum(HairEyes.table)
> Xsq<-sum((HairEyes.table-HE.fitted)^2/HE.fitted)
> (pval<-1-pchisq(Xsq,df=9))

+ [1] 0

```