Inference for Counts in Two-Way Tables

- data is often presented as counts in two-way tables
- in some cases the data consists of random samples from each of several subpopulations, and the goal is to assess whether the possible outcomes have the same distribution within each subpopulation
- in other cases the data consists of a single sample from a population where each sampled item is cross classified according to two categorical variables
- in general there are r rows and c columns and the counts are $X_{ij}, i = 1, \ldots, r, j = 1, \ldots, c$

	Column Variable			
Row Variable	1	...	c	Total
1	X_{11}	...	X_{1c}	$X_{1.}$
...	...	...	...	...
r	X_{r1}	...	X_{rc}	$X_{r.}$
Total	$X_{.1}$	...	$X_{.c}$	$X_{..}$

Test of Homogeneity

Example: Consider the following table, in which the cells in 5 samples were classified as to type

	Cell type						
Sample	P	Q	R	S	T	U	Total
1	3	24	18	13	32	10	100
2	7	25	16	11	27	14	100
3	5	28	14	10	31	12	100
4	7	22	19	11	26	15	100
5	17	19	13	8	25	18	100

- we'll call the values X_{ij}, for $i = 1, ..., r$ and $j = 1, ..., c$
- in this case the row totals $r_i = X_{i\cdot}$ were fixed by the study design
- we can assume each row follows a multinomial distribution, with number of trials r_i (all 100 in this example) and probabilities p_{ij}, $j = 1, \ldots, c$

- note that the probabilities sum to 1 in each row, $\sum_{j=1}^{r} p_{ij} = 1$
- the question of interest is whether the distribution of cell types is the same across all samples, that is, whether the multinomial probabilities are the same in each row
- so the null hypothesis is

$$H_0 : p_{ij} = p_{kj} = p_{.j}$$

for all i, j, k or, more simply

$$H_0 : \quad the\ distribution\ of\ cell\ types\ is\ the\ same\ in\ each\ sample$$

- the alternative hypothesis is

$$H_a : p_{ij} \neq p_{kj}$$

for some i, j, k, or

$$H_a : \quad the\ distribution\ of\ cell\ types\ is\ not\ the\ same\ in\ each\ sample$$

- we can also say

$$H_0: \quad the\ distributions\ of\ cell\ types\ are\ homogeneous$$
$$H_a: \quad the\ distributions\ of\ cell\ types\ are\ not\ homogenous$$

Test for Association

Example: 327 people were randomly selected and their hair colour and handedness was determined. The results are

	Hair Colour			
Handedness	Red	Brown	Other	Total
Left	12	54	37	103
Right	29	75	66	170
Ambidex.	7	27	20	54
Total	48	156	123	327

- here there is a single sample, and the cells of the multinomial distribution form a two-way table
- the hypotheses of interest are

$$H_0: \quad there\ is\ no\ association\ between$$

$$H_a : \quad \begin{array}{r} handedness\ and\ hair\ colour \\ there\ is\ an\ association\ between \\ handedness\ and\ hair\ colour \end{array}$$

- the cells of the table follow a multinomial distribution with probabilities p_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$
- as with all multinomials, the probabilities must sum to one,

$$\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$$

- the null hypothesis of no association is

$$H_0 : p_{ij} = p_{i.} p_{.j}$$

  for all i, j
- here, $p_{i.}$ is the probability of being in the ith row, and $p_{.j}$ is the probability of being in the jth column
- recall that the joint probability of independent events is the product of the

marginal probabilities, so the null hypothesis assumes independence between the row and column variables

- the alternative hypothesis is

$$H_a : p_{ij} \neq p_{i.}p_{.j}$$

for some i, j

Test for Homogeneity or Association

- the analysis of the data is the same for tests of homogeneity and test of independence!
- this is fortunate, because the distinction between the two is not always clear
- even if the data are collected as a single sample, the null hypothesis of no association can be phrased as homogeneity of *conditional* distributions across the rows or columns
  - for example, with the handedness-hair colour example we can ask whether the distribution of hair colour is the same

given the person is right handed, left handed or ambidexterous people

- the test statistic is the goodness of fit statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - e_{ij})^2}{e_{ij}}$$

- the test statistic compares the observed counts to those expected assuming H_0 is true in each cell of the table

- the expected counts are given by

$$\begin{aligned} e_{ij} &= \frac{X_{i.} X_{.j}}{X_{..}} \\ &= \frac{row\ i\ sum \times column\ j\ sum}{overall\ sum} \end{aligned}$$

- X^2 has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom

- for the approximation to be valid, all expected counts should be greater than 5

- we also assume that the data consist of (a) random sample(s)
- the P value is $P(\chi^2_{(r-1)(c-1)} > X^2)$, and as usual smaller P values give stronger evidence against H_0
- in general
  - observed and expected close: little evidence of association
  - observed and expected far apart: strong evidence of association
- the expected counts are calculated as,

$$e_{ij} = \frac{row\ sum \times column\ sum}{overall\ sum} = \frac{X_{i.}X_{.j}}{X_{..}}$$

  where the terms have different meaning depending on the hypothesis
- for *heterogeneity* the row (or column) totals are considered fixed, and the expected counts are

$$e_{ij} = \frac{r_i X_{.j}}{n} = r_i \hat{p}_{.j}$$

where $n = X_{..} = \sum r_i$ is the overall sum

- here $\hat{p}_{.j} = X_{.j}/n$ is the pooled estimate of the common value of $p_{.j}$ under the null hypothesis, and multiplying by $r_i = X_{i.}$ gives the mean or expected value
- for *association*, when we calculate expected counts using

$$e_{ij} = \frac{X_{i.}X_{.j}}{n}$$

we are using the formula $n\hat{p}_{ij}$ under the null hypothesis, where $\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j}$, and $\hat{p}_{i.} = X_{i.}/n$ and $\hat{p}_{.j} = X_{.j}/n$

- so

$$\begin{aligned} e_{ij} &= n\hat{p}_{ij} = n\hat{p}_{i.}\hat{p}_{.j} \\ &= n\frac{X_{i.}}{n}\frac{X_{.j}}{n} = \frac{X_{i.}X_{.j}}{n} \end{aligned}$$

Examples

Cell types:

- here the null hypothesis is that the distributions of cell types are the same for each sample
- the data can be analyzed in Minitab, as follows

```
MTB > chis c2-c7

Chi-Square Test: C2, C3, C4, C5, C6, C7

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected
counts

        C2     C3     C4     C5     C6     C7  Total
1        3     24     18     13     32     10    100
      7.80  23.60  16.00  10.60  28.20  13.80
     2.954  0.007  0.250  0.543  0.512  1.046

2        7     25     16     11     27     14    100
      7.80  23.60  16.00  10.60  28.20  13.80
     0.082  0.083  0.000  0.015  0.051  0.003

3        5     28     14     10     31     12    100
      7.80  23.60  16.00  10.60  28.20  13.80
```

```
     1.005   0.820   0.250   0.034   0.278   0.235

4        7      22      19      11      26      15     100
      7.80   23.60   16.00   10.60   28.20   13.80
     0.082   0.108   0.563   0.015   0.172   0.104

5       17      19      13       8      25      18     100
      7.80   23.60   16.00   10.60   28.20   13.80
    10.851   0.897   0.563   0.638   0.363   1.278

Total   39     118      80      53     141      69     500

Chi-Sq = 23.802, DF = 20, P-Value = 0.251
```

- we have no evidence of a difference in distribution of cell types in the different samples

Hair colour/handedness:

- the observed counts are

	Hair Colour			
Handedness	Red	Brown	Other	Total
Left	12	54	37	103
Right	29	75	66	170
Ambidex.	7	27	20	54
Total	48	156	123	327

- calculating proportions relative to a row or column total, or to the overall total often reveals interesting features of the data
- for example, division by the sample size gives

	Hair Colour			
Handedness	Red	Brown	Other	Total
Left	.04	.17	.11	.32
Right	.09	.23	.20	.52
Ambidex.	.02	.08	.06	.16
Total	.15	.48	.37	1.00

- for this we can see that most people have brown hair and are right handed
- or we could calculate proportions relative to the row or column totals
- using the row total gives

	Hair Colour			
Handedness	Red	Brown	Other	Total
Left	.12	.52	.36	1.00
Right	.17	.44	.39	1.00
Ambidex.	.13	.50	.37	1.00
Total	.15	.48	.37	1.00

- the three distributions of hair colour appear to be similar for left, right and ambidexterous people
- the expected counts are

	Hair Colour			
Handedness	Red	Brown	Other	Total
Left	15.12	49.14	38.74	103
Right	24.95	81.10	63.94	170
Ambidex.	7.93	25.76	20.31	54
Total	48	156	123	327

- for example $e_{11} = 103(48)/327 = 15.12$
- note that the expected counts add up to the same row and column totals as the observed counts
- note also that the expected counts are not rounded

- the contributions to X^2 are

	Hair Colour			
Handedness	Red	Brown	Other	Total
Left	.644	.481	.078	
Right	.656	.459	.066	
Ambidex.	.108	.060	.005	
Total				

- the total test statistic is $X^2 = 2.557$ on $2 \times 2 = 4$ degrees of freedom
- the P value is .63 indicating that there is no evidence against the null hypothesis of no association between handedness and hair colour
- the calcuations can be done in Minitab using the *chisquare* command

```
MTB > print c1-c3
 ROW     C1     C2     C3
   1     12     54     37
   2     29     75     66
   3      7     27     20
DATA> chisquare c1-c3
Expected counts are printed below observed
```

counts

	C1	C2	C3	Total
1	12	54	37	103
	15.12	49.14	38.74	
2	29	75	66	170
	24.95	81.10	63.94	
3	7	27	20	54
	7.93	25.76	20.31	
Total	48	156	123	327

ChiSq = 0.644 + 0.481 + 0.078 +
0.656 + 0.459 + 0.066 +
0.108 + 0.060 + 0.005 = 2.557

df = 4