Chi squared analysis

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn

- Contingency tables
- Chi-squared test for independence

Consider a study of vitamin D defiency

Status	Men	Women
Defiency	20	65
Insufficiency	65	120
Normal	115	115

Question of interest:

Is vitamin D status independent of sex ?

Survey of education and ethnicity

Education	Asian	Caucasian	Hispanics
Primary	31	120	84
Secondary	305	536	311
Tertiary	274	484	165
Total	610	1240	560

Question of interest:

Is there an association between educational levels and ethnicity?

Chi squared test

- Also known as Pearson's Chi squared test
- Purpose: to test for independence between factors
- Applicable to 2x2 or r x c tables (r = number of rows and c = number of columns)

Independence

- Null hypothesis: independence
- Independent = there is NO association
- If 2 factors are independent, then there is no association between the 2 factors

Vitamin D and sex

Status	Men	Women
Defiency	20 (0.100)	65 (0.217)
Insufficiency	65 (0.325)	120 (0.400)
Normal	115 (0.575)	115 (0.383)
Total	200 (1.000)	300 (1.000)

- **Women had higher prevalence of vitamin D defiency than men**
- **Is the difference statistically significant?**

Vitamin D and sex

Status	**Men**	**Women**	**Total**
Defiency	20 (0.100)	65 (0.217)	**85 (0.17)**
Insufficiency	65 (0.325)	120 (0.400)	**185 (0.37)**
Normal	115 (0.575)	115 (0.383)	**230 (0.46)**
Total	200 (1.000)	300 (1.000)	**500 (1.00)**

If there sex and vitamin D status are independent, what would we expect ?

Under the assumption of independence

Status	Men	Women	Total (average)
Defiency			0.17
Insufficiency			0.37
Normal			0.46
Total	200	300	1.00

We would expect the proportion (probability) of vitamin D status for men is the same as for women

Average = expected probability

Under the assumption of independence

Expected values

Status	Men	Women	Total (average)
Defiency	0.17 x 200 = **34**	0.17 x 300 = **51**	**0.17**
Insufficiency	0.37 x 200 = **74**	0.37 x 300 = **111**	**0.37**
Normal	0.46 x 200 = **92**	0.46 x 300 = **138**	**0.46**
Total	200	300	**1.00**

Compared with observed values

Observed and expected values

Status	Men	Women
Defiency	20 **(34)**	65 **(51)**
Insufficiency	65 **(74)**	120 **(111)**
Normal	115 **(92)**	115 **(138)**
Total	200	300

How do we assess the differences between observed and expected values

Answer: Chi-squared statistic

Chi-squared statistic

Observed (O) and expected values (E)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Chi-squared statistic

Observed (O) and expected values (E)

Status	Men	Women
Defiency	20 **(34)**	65 **(51)**
Insufficiency	65 **(74)**	120 **(111)**
Normal	115 **(92)**	115 **(138)**
Total	200	300

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(20 - 34)^2}{34} + \frac{(65 - 74)^2}{74} + \frac{(115 - 92)^2}{92}$$

$$+ \frac{(65 - 51)^2}{51} + \frac{(120 - 111)^2}{111} + \frac{(115 - 138)^2}{138} = 21.01$$

R codes

```
dat = matrix(c(20,65,115,65,120,115), 3)
chisq.test(dat)
```

```
Pearson's Chi-squared test
data:  dat
X-squared = 21.0155, df = 2, p-value =
2.732e-05
```

Chi-squared statistic

- For a contingency tables with r rows and c columns, the chi-squared statistic is distributed with (r – 1)*(c – 1) degrees of freedom (df)
- For 3 rows and 2 columns, the chi-squared statistic is distributed with 2 df
- Under the assumption of independence, chi-squared statistic with 2 df should be (expected) 5.99
- R code: `qchisq(.95, df=2)`

Chi-squared statistic

- Under the assumption of independence, chi-squared statistic with 2 df should be (expected) 5.99
- But we observed the chi squared statistic of 21.01
- P(chi-squared $\geq$ 21.01 | independence) = 0.0000273
- R code:

```
pchisq(21.01, 2, lower.tail = FALSE)
```

Vitamin D and sex

- We conclude that there was a SIGNIFICANT association between sex and vitamin D defiency
- In other words, the distribution of vitamin D status is *significantly* dependent on sex

Data on education and ethnicity

Education	Asian	Caucasian	Hispanics
Primary	31 (0.051)	120 (0.097)	84 (0.150)
Secondary	305 (0.500)	536 (0.432)	311 (0.555)
Tertiary	274 (0.449)	484 (0.390)	165 (0.295)
Total	610 (1.000)	1240 (1.000)	560 (1.000)

- **There seems different between groups in terms of educational levels**
- **Are the differences significant ?**

R analysis

Education	Asian	Caucasian	Hispanics
Primary	31	120	84
Secondary	305	536	311
Tertiary	274	484	165
Total	610	1240	560

```
dat = matrix(c(31, 305, 274, 120, 536, 484,
84, 311, 165), 3)
chisq.test(dat)
```

R analysis

```
dat = matrix(c(31, 305, 274, 120, 536, 484,
84, 311, 165), 3)

chisq.test(dat)
```

```
Pearson's Chi-squared test
data:  dat
X-squared = 54.9432, df = 4, p-value = 3.339e-11
```

Data on education and ethnicity

Education	Asian	Caucasian	Hispanics
Primary	31 (0.051)	120 (0.097)	84 (0.150)
Secondary	305 (0.500)	536 (0.432)	311 (0.555)
Tertiary	274 (0.449)	484 (0.390)	165 (0.295)
Total	610 (1.000)	1240 (1.000)	560 (1.000)

- **Are the differences significant ?**
- **YES**

When cell counts are small

Consider the following data …

Husband's rating	Wife's rating			
	N	F	V	A
N	7	7	2	3
F	2	8	3	7
V	1	5	4	9
A	2	8	9	14
Total	12	28	18	33

N=never, F=fairly often, V=very often, A=almost always

Sparse data, not quite appropriate for the usual chi squared test (problem of large sample approximation)

Chi squared test

```
x = matrix(c(7,7,2,3,
             2,8,3,7,
             1,5,4,9,
             2,8,9,14),4)
chisq.test(x)

Pearson's Chi-squared test
data:  x
X-squared = 16.9552, df = 9, p-value =
0.04942
Warning message:
In chisq.test(x) : Chi-squared approximation
may be incorrect
```

Exact permutation test

- Reconstruct the individual data

```
W:NNNNNNNFFFFFFFVVAAANNFFFFFFFF ...

H:NNNNNNNNNNNNNNNNNNNNFFFFFFFFF ...
```

- Permute either the W or H row
- Recalculate the contingency table
- Calculate the χ^2 statistic for each permutation
- Percentage of times it is larger than the observed value is an exact P-value

R code

```
x = matrix(c(7,7,2,3,
             2,8,3,7,
             1,5,4,9,
             2,8,9,14),4)

chisq.test(x, simulate.p.value = TRUE)
```

Results

```
Pearson's Chi-squared test with simulated p-
value (based on 2000 replicates)
data:  x
X-squared = 16.9552, df = NA, p-value =
0.05297
```

Fisher's exact test

Chemical toxicant and 10 mice

	Tumour	None
Treated	4	1
Control	2	3

- H0 : p1 = p2 = p
- Can't use Z or χ^2 because sample size is small
- Don't have a specific value for p

Fisher's exact test

- Under the null hypothesis every permutation is equally likely
- Observed data

```
    Treatment : T T T T T C C C C C

Tumor :TTTTNTTNNN
```

- Permuted

```
    Treatment : T C C T C T T C T C

Tumor : NTTNNTTTNT
```

- Fisher's exact test uses this null distribution to test the hypothesis that $p_1 = p_2$

Hyper-geometric distribution

- X number of tumors for the treated
- Y number of tumors for the controls
- H0 : $p_1 = p_2 = p$
- Under H0

$X \sim \text{Binom}(n1, p)$

$Y \sim \text{Binom}(n2, p)$

$X+Y \sim \text{Binom}(n1+n2, p)$

Hyper-geometric distribution

$$P(X = x \mid X + Y = z) = \frac{\begin{pmatrix} n_1 \\ x \end{pmatrix} \begin{pmatrix} n_2 \\ z - x \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ z \end{pmatrix}}$$

This is the hypergeometric pmf

R codes

```
dat = matrix(c(4, 1, 2, 3), 2)

fisher.test(dat, alternative = "greater")
```

R codes

```
Fisher's Exact Test for Count

Data data: dat

p-value = 0.2619

alt hypoth: true odds ratio is greater than 1 95
percent confidence interval: 0.3152217 Inf sample
estimates: odds ratio 4.918388
```