Contingency Table: χ^2 Test

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Outline

- χ^2 Test for equivalence of two binomial proportions
- χ^2 Test for independence, 2×2 tables
- χ^2 Test for multiple binomial proportions
- χ^2 Test for independence, $r \times c$ tables
- χ^2 Test for goodness of fit

χ^2 Test

- An alternative approach to testing equality of proportions uses the χ^2 statistic

$$\sum \frac{(\text{Observed-Expected})^2}{\text{Expected}}$$

- "Observed" are the observed counts
- "Expected" are the expected counts under the null hypothesis
- The sum is overall four cells
- This statistic follows a χ^2 distribution with 1 df
- The χ^2 statistic is exactly the square of the difference in proportions **Score statistic**.

Example

Treat	Side Effects	None	Total
X	44	56	100
Y	77	43	120
Total	121	99	220

- $H_0 : p_1 = p_2$

- The χ^2 statistic is $\sum \frac{(O-E)^2}{E}$
- $O_{11} = 44$, $E_{11} = \frac{121}{220} \times 100 = 55$
- $O_{21} = 77$, $E_{21} = \frac{121}{220} \times 120 = 66$
- $O_{12} = 56$, $E_{12} = \frac{99}{220} \times 100 = 45$
- $O_{22} = 43$, $E_{22} = \frac{99}{220} \times 120 = 54$

$$\chi^2 = \frac{(44-55)^2}{55} + \frac{(77-66)^2}{666} + \frac{(56-45)^2}{45} + \frac{(43-54)^2}{54}$$

Which turns out to be 8.96. Compare to a χ^2 with one degree of freedom (reject for large values).

```
pchisq(8.96, 1, lower.tail = FALSE)
#result is 0.002
```

R code

```
dat <- matrix(c(44, 77, 56, 43), 2)
chisq.test(dat)
chisq.test(dat, correct = FALSE)
```

Notation Reminder

$n_{11} = X$ $n_{21} = Y$	$n_{12} = n_1 - X$ $n_{22} = n_2 - Y$	$n_1 = n_{1+}$ $n_2 = n_{2+}$
n_{+1}	n_{+2}	

Notes

- Reject if the statistic is too large
- Alternative is two sided
- Do not divide α by 2
- A small χ^2 statistic implies little difference between the observed values and those expected under H_0
- The χ^2 statistic and approach generalized to other kinds of tests and larger contingency tables
- Alternative computational form for the χ^2 statistic

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{+1}n_{+2}n_{1+}n_{2+}}$$

Notes

- Notice that the statistic

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{+1}n_{+2}n_{1+}n_{2+}}$$

  does not change if you transpose the rows and the columns of the table
- Surprisingly, the χ^2 statistic can be used
  - the rows are fixed (binomial)
  - the columns are fixed (binomial)
  - the total sample size is fixed (multinomial)
  - non are fixed (Poisson)
- For a given set of data, any of these assumptions results in the same value for the statistic

Testing independence

- Maternal age versus birthweight[1]
- Cross-sectional sample, only the total sample size is fixed
- H_0: MA is independent of BW
- H_a: MA is not independent of BW

	Birthweight		
Mat. Age	< 2500g	$\geq 2,500$ g	Total
$<$ 20 year	20	80	100
$\geq 20 year$	30	270	300
Total	50	350	400

1

[1] From Agresti Categorical Data Analysis 2nd

Continued

- Under H_0 (est) $P(\text{MA} < 20) = \frac{100}{400} = .25$
- Under H_0 (est) $P(\text{BW} < 2500) = \frac{50}{400} = .125$
- Under H_0 (est)

$$P(\text{MA} < 20 \text{ and } \text{BW} < 2,500) = .25 \times .125$$

- Therefore
  - $E_{11} = \frac{100}{400} \times \frac{50}{400} \times 400 = 12.5$
  - $E_{12} = \frac{100}{400} \times \frac{350}{400} \times 400 = 87.5$
  - $E_{21} = \frac{300}{400} \times \frac{50}{400} \times 400 = 37.5$
  - $E_{22} = \frac{300}{400} \times \frac{350}{400} \times 400 = 262.5$
  - $\chi^2 = \frac{(20-12.5)^2}{12.5} + \frac{(80-87.5)^2}{87.5} \frac{(30-37.5)^2}{37.5} + \frac{(270-262.5)^2}{262.5} = 6.86$
- Compare to critical value
  `qchisq(.95, 1)=3.84`
- Or calculate P-value
  `pchisq(6.86, 1, lower.tail = F)=.009`

χ^2 testingcontinued

Group	Alcohol Use High	Low	Total
Clergy	32	268	300
Educators	51	199	250
Executives	67	233	300
Retailers	83	267	350
Total	233	967	1,200

2

[2] From Agresti Categorical Data Analysis 2nd

- Interest lies in testing whether or not the proportions of high alcohol use is the same in the four occupations
- $H_0 : p_1 = p_2 = p_3 = p4 = p$
- H_a: at least two of the $p_j s$ are unequal
- $O_{11} = 32, E_{11} = 300 \times \frac{233}{1200}$
- $O_{12} = 268, E_{11} = 300 \times \frac{967}{1200}$
- ...
- χ^2 statistic $\sum \frac{(O-E)^2}{E} = 20.59$
- $df = (Rows - 1) \times (Columns - 1) = 3$
- p-value pchisq(20.59, 3, lower.tail=F) ≈ 0

Word distributions

	Book			
Word	1	2	3	Total
a	147	186	101	434
an	25	26	11	62
this	32	39	15	86
that	94	105	37	236
with	59	74	28	161
without	18	10	10	38
Total	375	440	202	1017

3

[3] From Rice Mathematical Statistics and Data Analysis 2nd

Example: Word distributions

- H_0 : The probabilities of each word are the same for every book
- H_a : At least two are different
- $O_{11} = 147, E_{11} = 375 \times \frac{434}{1017}$
- $O_{12} = 186, E_{12} = 440 \times \frac{434}{1017}$
- ..
- $\sum \frac{(O-E)^2}{E} = 12.27$
- df=(6-1)(3-1) =10

Testing independence

	Wife's Rating			
Husband	N	F	V	A	Total
N	7	7	2	3	19
F	2	8	3	7	20
V	1	5	4	9	19
A	2	8	9	14	33
Total	12	28	18	33	91

N=never, F=fairly often, V=very often, A=almost always
[4]

[4] From Agresti Categorical Data Analysis 2nd

Independence cont'd

- H_0 : H and W ratings are independent
- H_a : not independent
- $P(H = N \& W = A) = P(H = N)P(W = A)$
- stat=$\sum \frac{(O-E)^2}{E}$
- $O_{11} = 7, E_{11} = 91 \times \frac{19}{91} \times \frac{12}{91} = 2.51$
- $O_{12} = 186, E_{12} = 440 \times \frac{434}{1017}$
- $E_{ij} = n_{i+}n_{+j}/n$
- df=(Rows-1)(Columns-1)

Independence cont'd

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

- $\sum \frac{(O-E)^2}{E} = 16.96$
- df=(4-1)(4-1)=9
- p-value=0.049
- Cell counts might be too small to use large sample approximation

Notes

- χ^2 result requires large cell counts
- *df* is always $(Rows - 1)(Columns - 1)$
- Generalization of Fishers exact test can be used or continuity corrections can be employed

Exact Permutation Test

- Reconstruct the individual data W:
  NNNNNNNFFFFFFFVVAAANNFFFFFFFF ...
  H: NNNNNNNNNNNNNNNNNNFFFFFFFF...
- Permute either the W or H row
- Recalculate the contingency table
- Calculate the χ^2 statistic for each permutation
- Percentage of time it is larger than the observed value is an exact *p*-value

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

χ^2 goodness of fit

Results from R's random number generation (RNGs)

	[0, 0.25)	[0.25, 0.5)	[0.5, 0.75)	[0.75, 1)	Total
Count	254	235	267	244	1000
Ture p	0.25	0.25	0.25	0.25	1

- $H_0 : p_1 = 0.25, p_2 = 0.25, p_3 = 0.25, p_4 = 0.25$
- H_a : any $p_i \neq$ it's hypothesized value

Continued

- $O_1 = 254, E_1 = 1000 \times 0.25 = 250$
- $O_2 = 235, E_1 = 1000 \times 0.25 = 250$
- $O_3 = 267, E_1 = 1000 \times 0.25 = 250$
- $O_4 = 244, E_1 = 1000 \times 0.25 = 250$
- $\sum \frac{(O-E)^2}{E} = 2.264$
- df=3
- p-value=0.51

Notes on GOF

- Test of whether or not observed counts equal theoretical values
- Test statistic is $\sum \frac{(O-E)^2}{E}$
- TS follows χ^2 distribution for large n
- df if the number of cells -1
- Undirected alternative is problematic
- Especially useful for testing RNGs