AMS 572 Data Analysis I
Inference for one-way and two-way count data

Pei-Fen Kuan

Applied Math and Stats, Stony Brook University

Inference for one-way count data

Review: Multinomial Distribution

Multinomial experiment:

1. The experiment consists of n identical and independent trials.
2. The outcome of each trial falls into one of k class.
3. The probability that the outcome of a single trial will fall in a particular class, say class i is p_i, $(i = 1, \ldots, k)$ and remains the same from trial to trial

$$p_1 + p_2 + \ldots + p_k = 1$$

4. The random variable X_i is equal to the number of trials in which the outcome falls in class i, $i = 1, \ldots, k$. Note that

$$X_1 + X_2 + \ldots + X_k = n$$

Review: Multinomial Distribution

The joint p.m.f. of $X_1, X_2, \ldots, X_k$ is given by

$$\begin{aligned} p(n_1, n_2, \ldots, n_k) &= P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) \\ &= \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k} \end{aligned}$$

where $\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{k} n_i = n$.

The random variables $X_1, X_2, \ldots, X_k$ are said to have a multinomial distribution with parameters n and $p_1, p_2, \ldots, p_k$.

Inference on $k > 2$ proportions

- Also known as inference for one way count data
- Suppose $H_0 : p_1 = p_1^0, p_2 = p_2^0, \ldots, p_k = p_k^0$ vs H_a : At least one $p_i \neq p_i^0$
- Under H_0, the expected count of category i is $e_i = np_i^0$, $i = 1, \ldots, k$
- Test statistic

$$W_0 = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i}$$

  where x_i is the observed number of observations in category i and $e_i = np_i^0$
- Under H_0, W_0 converges to χ^2_{k-1}
- At significance level α, reject H_0 if $W_0 > \chi^2_{k-1,\alpha,U}$

χ^2 test for inference on $k > 2$ proportions

- The null distribution of W_0 is approximately χ^2_{k-1} when sample size is large
- Rule of thumb: all $e_i \geq 1$ and no more than $1/5$ of the $e_i < 5$.
- What happens if we have many cells with small counts?

χ^2 test for inference on $k > 2$ proportions

- Note that $H_0 : p_1 = p_1^0, p_2 = p_2^0, \ldots, p_k = p_k^0$ means we are testing how well a particular model fits the data
- Thus, this test is also known as the χ^2 goodness of fit test
- Rejecting the null implies the model does not provide an adequate fit to the data

Example: Gregor Mendel (1822-1884) was an Austrian monk whose genetic theory is one of the greatest scientific discovery of all time. In his famous experiment with garden peas, he proposed a genetic model that would explain inheritance. In particular, he studied how the shape (smooth or wrinkled) and color (yellow or green) of pea seeds are transmitted through generations. His model shows that the second generation of peas from a certain ancestry should have the following distribution. Conduct at test at $\alpha = 0.05$ if his theoretical probabilities are correct.

	wrinkled-green	wrinkled-yellow	smooth-green	smooth-yellow
Theoretical probabilities	$p_1 = \frac{1}{16}$	$p_2 = \frac{3}{16}$	$p_3 = \frac{3}{16}$	$p_4 = \frac{9}{16}$
Observed counts	31	102	108	315

Solution

	wrinkled-green	wrinkled-yellow	smooth-green	smooth-yellow
Theoretical probabilities	$p_1 = \frac{1}{16}$	$p_2 = \frac{3}{16}$	$p_3 = \frac{3}{16}$	$p_4 = \frac{9}{16}$
Observed count out of 556	X_1=31	X_2=102	X_3=108	X_4=315
Expected counts	e_1 =34.75	e_2=104.25	e_3=104.25	e_4=312.75

$H_0 : p_1 = \frac{1}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{9}{16}$ $H_a : H_0$ not true

T.S $W_0 = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} \overset{H_0}{\dot{\sim}} \chi^2_{k-1}$

$\approx 0.604 < \chi^2_{3,0.05,upper}$=7.815

$\therefore$At significance level 0.05, we cannot reject H_0

SAS Code

```
DATA GENE;
INPUT COLOR $ NUMBER;
DATALINES;
YELLOWSMOOTH  315
YELLOWWRINKLE 102
GREENSMOOTH   108
GREENWRINKLE  31
;

PROC FREQ DATA=GENE ORDER=DATA;
WEIGHT NUMBER;
TITLE3 'GOODNESS OF FIT ANALYSIS';
TABLES COLOR / CHISQ NOCUM TESTP=(0.5625 0.1875 0.1875 0.0625);
RUN;
```

SAS Output

```
                 GOODNESS OF FIT ANALYSIS

                    The FREQ Procedure

                                                    Test
COLOR         Frequency        Percent          Percent

YELLOWSM           315          56.65            56.25
YELLOWWR           102          18.35            18.75
GREENSMO           108          19.42            18.75
GREENWRI            31           5.58             6.25


                     Chi-Square Test
                for Specified Proportions

                  Chi-Square     0.6043
                  DF                  3
                  Pr > ChiSq     0.8954

                    Sample Size = 556
```

R Code and Output

```
> obs <- c(315,102,108,31)
> null.prob <- c(9/16,3/16,3/16,1/16)
> chisq.test(obs,p=null.prob)

Chi-squared test for given probabilities

data:  obs
X-squared = 0.6043, df = 3, p-value = 0.8954
```

Inference for two-way count data

Contingency Tables

- Two-way $(r \times c)$ contingency table:

$$
\begin{array}{c|cccc}
 & & & j & \\
\text{i} & 1 & 2 & \cdots & c \\
\hline
1 & n_{11} & n_{12} & \cdots & n_{1c} \\
2 & n_{21} & n_{22} & \cdots & n_{2c} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
r & n_{r1} & n_{r2} & \cdots & n_{rc}
\end{array}
$$

- Notation:

$$
n_{i\cdot} = \sum_{j=1}^{c} n_{ij} \qquad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}
$$

Contingency Tables

- Two scenarios where $r \times c$ table arise
  1. Sample from a **single** population and measure two characteristics, say X and Y

     $$\Pr[X = i, Y = j] = p_{ij} \ ; \ \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$$

     In this case, the total sample size is fixed
  2. Each row corresponds to a sample from a different population

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

     In this case, the row totals are fixed

Contingency Table: Example

- A survey of physicians asked about the size of community in which they were reared and the size of the community in which they practice

	Practice				
Reared	<5k	5-49k	50-99k	100k+	Total
<5k	40	38	32	37	147
5-49k	26	42	35	33	136
50-99k	24	26	34	31	115
100k+	30	39	53	60	182
	120	145	154	161	580

Contingency Table: Example

- A case-control study was conducted to investigate the relationship between age at first birth and breast cancer

	Age at 1st birth					
	<20	20-24	25-29	30-34	≥ 35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10245
	1742	5638	3904	1555	626	13465

Contingency Tables

- Physician's example H_0: size of place of practice is independent of size of place of rearing

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$$

- Test of independence

$$\Pr[X = i, Y = j] = \Pr[X = i] \Pr[Y = j]$$

for $i = 1, \ldots, r$; $j = 1, \ldots, c$

Contingency Tables

- Breast cancer example H_0 : distribution of age at 1st birth is the same for cases and controls

$$H_0 : p_{ij} = p_{i'j}; j = 1, 2, \ldots, c$$

- Test of homogeneity/association

Test of Independence or No Association

- Under either H_0, the estimated expected frequency in the (i, j) cell is

$$E_{ij} = \frac{n_{i.}n_{.j}}{N}$$

- Consider breast cancer example
  - If H_0 is true, would expect the proportion of women < 20 to be

$$\frac{n_{11} + n_{21}}{N} = \frac{n_{.1}}{N}$$

  - There are $n_{1.}$ cases, so we would expect

$$E_{11} = n_{1.}\frac{n_{.1}}{N} = \frac{n_{1.}n_{.1}}{N}$$

    cases to be < 20 years old

Test of Independence or Association

- Under H_0, the expected frequency in the (i, j) cell is

$$E_{ij} = \frac{n_{i.}n_{.j}}{N}$$

- Let

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

i.e.

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N}$$

Test of Independence

- Under H_0,
$$X^2 \sim \chi^2_{(r-1)(c-1)}$$
- Physician's Example:
$$(r-1)(c-1) = 3 \times 3 = 9$$

$$\text{Rejection Region} = C_{0.05} = \{X^2 : X^2 > \chi^2_{0.05,9,U} = 16.92\}$$

Physician's Example

- Expected values

	Practice				
Reared	<5k	5-49k	50-99k	100k+	Total
<5k	30.4	36.8	39.0	40.8	147
5-49k	28.1	34.0	36.1	37.8	136
50-99k	23.8	28.8	30.5	31.9	115
100k+	37.7	45.5	48.3	50.5	182
	120	145	154	161	580

Physician's Example

- Calculate test statistic

$$X^2 = \frac{(40-30.4)^2}{30.4} + \frac{(38-36.8)^2}{36.8} + \cdots + \frac{(60-50.5)^2}{50.5} = 12.76$$

- Do not reject H_0.
- There is insufficient evidence to conclude that place of practice and place of rearing are dependent; the data are consistent with the null hypothesis that place of practice and place of rearing are independent

SAS Code

```
data physician;
input reared $ practice $ count;
datalines;
less5k less5k 40
5to49k less5k 26
50to99k less5k 24
100k less5k 30
...
100k 50to99k 53
less5k 100k 37
5to49k 100k 33
50to99k 100k 31
100k 100k 60
run;

proc freq data=physician;
tables reared*practice/chisq;
weight count;
run;
```

SAS Output

```
                  GOODNESS OF FIT ANALYSIS

                     The FREQ Procedure

                 Table of reared by practice

reared     practice

Frequency
Percent
Row Pct
Col Pct  100k    50to99k 5to49k  less5k    Total

100k          60      53      39      30     182
           10.34    9.14    6.72    5.17   31.38
           32.97   29.12   21.43   16.48
           37.27   34.42   26.90   25.00

50to99k       31      34      26      24     115
            5.34    5.86    4.48    4.14   19.83
           26.96   29.57   22.61   20.87
           19.25   22.08   17.93   20.00

5to49k        33      35      42      26     136
            5.69    6.03    7.24    4.48   23.45
           24.26   25.74   30.88   19.12
           20.50   22.73   28.97   21.67

less5k        37      32      38      40     147
            6.38    5.52    6.55    6.90   25.34
           25.17   21.77   25.85   27.21
           22.98   20.78   26.21   33.33

Total        161        154        145       120         580
```

SAS Output

Statistics for Table of reared by practice

Statistic	DF	Value	Prob
Chi-Square	9	12.7632	0.1736
Likelihood Ratio Chi-Square	9	12.5264	0.1852
Mantel-Haenszel Chi-Square	1	8.0532	0.0045
Phi Coefficient		0.1483	
Contingency Coefficient		0.1467	
Cramer's V		0.0856	

Sample Size = 580

R Code and Output

```
> physician <- matrix(c(40,38,32,37,26,42,35,33,
24,26,34,31,30,39,53,60),nrow = 4,byrow=T,
dimnames = list(Practice = c("v5k","v5to49k","v50to99k","v100k")
Reared = c("v5k","v5to49k","v50to99k","v100k")))

> physician
          Reared
Practice    v5k v5to49k v50to99k v100k
  v5k        40      38       32    37
  v5to49k    26      42       35    33
  v50to99k   24      26       34    31
  v100k      30      39       53    60

> chisq.test(physician)

Pearson's Chi-squared test

data:  physician
X-squared = 12.7632, df = 9, p-value = 0.1736
```

Breast Cancer Example

- Underlying probabilities

	Age at 1st birth					
	<20	20-24	25-29	30-34	≥ 35	Total
Case	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	1
Control	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	1

Breast Cancer Example

- Null hypothesis

$$H_0 : p_{1j} = p_{2j} \text{ for } j = 1, 2, 3, 4, 5$$

- Can use same statistic

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Breast Cancer Example

- Expected frequencies

	Age at 1st birth					
	<20	20-24	25-29	30-34	≥ 35	Total
Case	416.6	1348.3	933.6	371.9	149.7	3220
Control	1325.4	4289.7	2970.4	1183.1	476.3	10245
	1742	5638	3904	1555	626	13465

Breast Cancer Example

- Test statistic

$$X^2 = \frac{(320 - 416.6)^2}{416.6} + \cdots + \frac{(406 - 476.3)^2}{476.3} = 130.3$$

- Rejection region

$$C_{0.05} = \{X^2 : X^2 > \chi^2_{0.05,4,U} = 9.49\}$$

- Reject H_0
- The age distributions are not the same

Asymptotic Approximation

- Note the χ^2 distribution for X^2 is an approximation
- The approximation works well for if $E_{ij} \geq 5$ for all i, j

Read Chapter 10