Nonparametric Statistics Notes

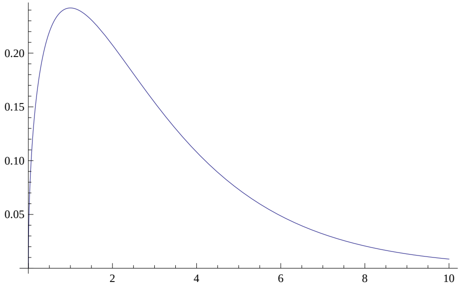
Chapter 4: Contingency Tables

Jesse Crawford

Department of Mathematics
Tarleton State University

Definition

- Let $Z_1, \ldots, Z_k$ be IID $N(0,1)$ random variables.
- $Y = Z_1^2 + \cdots + Z_k^2$ has a *chi-squared* distribution with k degrees of freedom.
- $Y \sim \chi^2(k)$



Outline

Testing for Differences in Probabilities (2×2 case)

	Class 1	Class 2	Total
Population 1	O_{11}	O_{12}	n_1
Population 2	O_{21}	O_{22}	n_2
Total	C_1	C_2	$N = n_1 + n_2$

- Assumptions:
  - The random samples are statistically independent.
  - $p_1 = P(\text{Class 1})$ in Population 1
  - $p_2 = P(\text{Class 1})$ in Population 2
  - Row totals are fixed. Column totals are random.
- Testing problems:
  - $\mathrm{H}_0 : p_1 = p_2$ vs. $\mathrm{H}_1 : p_1 \neq p_2$ (Two-tailed)
  - $\mathrm{H}_0 : p_1 \geq p_2$ vs. $\mathrm{H}_1 : p_1 < p_2$ (Lower-tailed)
  - $\mathrm{H}_0 : p_1 \leq p_2$ vs. $\mathrm{H}_1 : p_1 > p_2$ (Upper-tailed)

Testing for Differences in Probabilities (2×2 case)

	Class 1	Class 2	Total
Population 1	O_{11}	O_{12}	n_1
Population 2	O_{21}	O_{22}	n_2
Total	C_1	C_2	$N = n_1 + n_2$

- Test statistic:

$$T = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$$

- Null distribution: $T \approx N(0, 1)$
- p-values:

$$2 \cdot \min[P(Z \leq t_{\text{obs}}), P(Z \geq t_{\text{obs}})] \quad \text{(Two-tailed)}$$

$$P(Z \leq t_{\text{obs}}) \quad \text{(Lower-tailed)}$$

$$P(Z \geq t_{\text{obs}}) \quad \text{(Upper-tailed)}$$

Testing for Differences in Probabilities (2 $\times$ 2 case)

	Class 1	Class 2	Total
Population 1	O_{11}	O_{12}	n_1
Population 2	O_{21}	O_{22}	n_2
Total	C_1	C_2	$N = n_1 + n_2$

- Expected cell frequencies under H_0:

$$E_{ij} = \frac{n_i C_j}{N}$$

- Chi-squared Statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \left[\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} \right] - N$$

Testing for Differences in Probabilities (2×2 case)

	Class 1	Class 2	Total
Population 1	O_{11}	O_{12}	n_1
Population 2	O_{21}	O_{22}	n_2
Total	C_1	C_2	$N = n_1 + n_2$

- Chi-squared Statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \left[\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} \right] - N$$

- Null distribution: $\chi^2 \approx \chi^2(1)$ (Degrees of freedom $= 1$)
- p-value: $P(\chi^2 \geq \chi^2_{\text{obs}})$ (Two-tailed test only)

When is the Chi-squared Distribution a Good Approximation?

- Cochran's Criterion: The approximation may be poor if
  - Any E_{ij} is less than 1, or
  - more than 20% of the E_{ij}'s are less than 5
- Conover's Criterion: The approximation may be poor if
  - Any E_{ij} is less than 0.5, or
  - more than 50% of the E_{ij}'s are less than 1

Testing for Differences in Probabilities ($r \times c$ case)

	Class 1	Class 2	$\cdots$	Class c	Total
Population 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Population 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Population r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
Total	C_1	C_2	$\cdots$	C_c	N

- Assumptions:
  - The random samples are statistically independent.
  - $p_{ij} = P(\text{Class } j)$ in Population i
  - Row totals are fixed. Column totals are random.
- Two-tailed Testing problem:

$$\text{H}_0 : \text{All probabilities in the same column are equal to each other}$$
$$(p_{1j} = p_{2j} = \cdots = p_{rj}, \text{ for all } j)$$

Testing for Differences in Probabilities ($r \times c$ case)

	Class 1	Class 2	$\cdots$	Class c	Total
Population 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Population 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Population r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
Total	C_1	C_2	$\cdots$	C_c	N

- $E_{ij} = \frac{n_i C_j}{N}$
- Chi-squared Statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \left[\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} \right] - N$$

- Null distribution: $\chi^2 \approx \chi^2[(r-1)(c-1)]$
- p-value: $P(\chi^2 \geq \chi^2_{\text{obs}})$ (Two-tailed test only)

Example

- Website visitors were shown three different website layouts.
- 100 were shown layout 1
- 50 were shown layout 2
- 200 were shown layout 3
- Time spent browsing was also recorded.

	$T \leq 5$	$5 < T \leq 10$	$10 \leq T < 15$	$15 \leq T$
Layout 1	55	27	11	7
Layout 2	16	23	6	5
Layout 3	40	71	22	17

- Test the null hypothesis that the probability distribution of time spent browsing is the same for the different layouts.
- Note: Row totals are fixed, and column totals are random.

Testing for Independence ($r \times c$ case)

	Column 1	Column 2	$\cdots$	Column c	Total
Row 1	O_{11}	O_{12}	$\cdots$	O_{1c}	R_1
Row 2	O_{21}	O_{22}	$\cdots$	O_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	R_r
Total	C_1	C_2	$\cdots$	C_c	N

- Assumptions:
  - Random sample of N observations.
  - Each observation is a member of exactly one of the r rows and one of the c columns.
  - Both the row and column totals are random.
- Two-tailed Testing problem:

$$\mathrm{H}_0 : P(\text{row } i, \text{column } j) = P(\text{row } i) \cdot P(\text{column } j), \text{ for all } i, j.$$

- Testing procedure is the same as the previous test.

Chi-squared Test with Fixed Marginal Totals.

	Column 1	Column 2	$\cdots$	Column c	Total
Row 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Row 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
Total	c_1	c_2	$\cdots$	c_c	N

- Assumptions:
  - Both the row and column totals are fixed.
  - The data were randomly selected from all contingency tables with those row and column totals.
- Chi-squared testing procedure is the same as the previous tests.
  - May perform poorly because row and column totals are both fixed.
  - Need alternative methods.

Chi-squared Test with Fixed Marginal Totals.

	Column 1	Column 2	$\cdots$	Column c	Total
Row 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Row 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
Total	c_1	c_2	$\cdots$	c_c	N

Alternatives to chi-squared test:

- 2×2 case:
  - Fisher's exact test.
  - Uses hypergeometric distribution to calculate exact p-value.
  - `fisher.test(A)`
- $r \times c$ case:
  - Simulate p-value.
  - `chisq.test(A,simulate.p.value=TRUE)`

Outline

The Median Test

- Setting: Several independent random samples.
- Testing problem:

$$\begin{aligned} H_0 &: \text{All populations have the same median vs.} \\ H_1 &: \text{At least two have different medians.} \end{aligned}$$

- Testing procedure:
  - Grand Median $=$ Median of all samples combined.

Sample	1	2	$\cdots$	c	Totals
$>$ Grand Median	O_{11}	O_{12}	$\cdots$	O_{1c}	a
$\leq$ Grand Median	O_{21}	O_{22}	$\cdots$	O_{2c}	b
Totals	n_1	n_2	$\cdots$	n_c	N

  - Perform a chi-squared test.

Sample	1	2	$\cdots$	c	Totals
$>$ Grand Median	O_{11}	O_{12}	$\cdots$	O_{1c}	a
$\leq$ Grand Median	O_{21}	O_{22}	$\cdots$	O_{2c}	b
Totals	n_1	n_2	$\cdots$	n_c	N

Example

- Corn yields for four different methods of growing corn:
- Method 1: 83, 89, 89, 90, 91, 91, 92, 94, 96
- Method 2: 81, 83, 83, 84, 84, 88, 89, 90, 91, 91
- Method 3: 91, 93, 94, 95, 96, 100, 101
- Method 4: 77, 78, 79, 80, 81, 81, 81, 82
- Test whether the medians for these different methods are equal.

Outline

Cramer's Contingency Coefficient

- Let $T = \chi^2$ be the chi-squared statistic from an $r \times c$ contingency table.
- $N =$ number of total observations in table.
- Let $q = \min(r, c)$
- The largest possible value of T is $N(q-1)$

Definition

$$R_1 = \frac{T}{N(q-1)}$$

$$\text{Cramer's Contingency Coefficient} = \sqrt{R_1}$$

Definition

$$R_1 = \frac{T}{N(q-1)}$$

$$\text{Cramer's Contingency Coefficient} = \sqrt{R_1}$$

Interpretation of Cramer's Coefficient

$$0 \leq \text{Cramer's Contingency Coefficient} \leq 1.$$

- A value of 1 suggests complete dependence.
- A value of 0 suggests complete independence.
- The *p*-value of a chi-squared test of independence is a more reliable measure.

Outline

Chi-squared Goodness-of-Fit Tests

- One random sample.
- Each observation is either in Class 1, Class 2, . . ., or Class c.

Class	1	2	$\cdots$	c	Total
Observed Frequencies	O_1	O_2	$\cdots$	O_c	N

- $p_j = P(\text{Class } j)$
- $p_j^\star = P(\text{Class } j)$, under the null hypothesis
- $E_j = p_j^\star N$

$$\chi^2 = \left[\sum_{j=1}^{c} \frac{O_j^2}{E_j}\right] - N$$

- χ^2 has a chi-squared distribution.
- Degrees of freedom $= \dim(\mathrm{H}_1) - \dim(\mathrm{H}_0)$

Outline

Example

- 12 basketball games
- 3 basketball fans make predictions
- $1 =$ correct prediction
- $0 =$ incorrect prediction

Game	Fan 1	Fan 2	Fan 3	Totals
1	1	1	1	3
2	1	1	1	3
3	0	1	0	1
4	1	1	0	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
11	1	1	1	3
12	1	1	1	3
Totals	8	10	7	25

- Is there a statistically significant difference in the accuracy of the three fans predictions?

Cochran's Q-Test for Related Observations

	Treatments				
Subjects	1	2	$\cdots$	c	Row Totals
1	X_{11}	X_{12}	$\cdots$	X_{1c}	R_1
2	X_{21}	X_{22}	$\cdots$	X_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	X_{r1}	X_{r2}	$\cdots$	X_{rc}	R_r
Column Totals	C_1	C_2	$\cdots$	C_c	N

- Subjects are a large random sample from the population.
- X_{ij} is either 1 or 0.
- $p_{ij} = P(X_{ij} = 1)$
- Testing problem:

$$\mathrm{H}_0 : \text{For each row } i,\ p_{i1} = p_{i2} = \cdots = p_{ic}$$

H_0 : For every subject, all treatments are equally effective for that subject.

Cochran's Q-Test for Related Observations

	Treatments				
Subjects	1	2	$\cdots$	c	Row Totals
1	X_{11}	X_{12}	$\cdots$	X_{1c}	R_1
2	X_{21}	X_{22}	$\cdots$	X_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	X_{r1}	X_{r2}	$\cdots$	X_{rc}	R_r
Column Totals	C_1	C_2	$\cdots$	C_c	N

- $p_{ij} = P(X_{ij} = 1)$
- H_0 : For each row i, $p_{i1} = p_{i2} = \cdots = p_{ic}$

$$Q = c(c-1)\frac{\sum_{j=1}^{c}(C_j - \frac{N}{c})^2}{\sum_{i=1}^{r} R_i(c - R_i)}$$

- Null distribution: $T \sim \chi^2(c-1)$.

Cochran's Q-Test with Two Treatments

- If there are only two treatments, Cochran's Q-test is equivalent to the McNemar test.

```r
library(RVAideMemoire)


game.prediction=c(1,1,1,
                  1,1,1,
                  0,1,0,
                  1,1,0,
                  0,0,0,
                  1,1,1,
                  1,1,1,
                  1,1,0,
                  0,0,1,
                  0,1,0,
                  1,1,1,
                  1,1,1)

fan=gl(3,1,36,labels=1:3)
block=gl(12,3,labels=c(1:12))


cbind(game.prediction,fan,block)

cochran.qtest(game.prediction~fan|block)
```

```
> cbind(game.prediction,fan,block)
      game.prediction fan block
 [1,]               1   1     1
 [2,]               1   2     1
 [3,]               1   3     1
 [4,]               1   1     2
 [5,]               1   2     2
 [6,]               1   3     2
 [7,]               0   1     3
 [8,]               1   2     3
 [9,]               0   3     3
[10,]               1   1     4
[11,]               1   2     4
[12,]               0   3     4
[13,]               0   1     5
[14,]               0   2     5
[15,]               0   3     5
[16,]               1   1     6
[17,]               1   2     6
[18,]               1   3     6
[19,]               1   1     7
[20,]               1   2     7
[21,]               1   3     7
[22,]               1   1     8
[23,]               1   2     8
[24,]               0   3     8
[25,]               0   1     9
[26,]               0   2     9
[27,]               1   3     9
[28,]               0   1    10
[29,]               1   2    10
[30,]               0   3    10
[31,]               1   1    11
[32,]               1   2    11
[33,]               1   3    11
[34,]               1   1    12
[35,]               1   2    12
[36,]               1   3    12
```