Serik Sagitov, Chalmers Tekniska Högskola, May 28, 2005

Chapter 13. The analysis of categorical data

1. Fisher's exact test

Population proportions

	Population 1	Population 2
Category 1	π_{11}	π_{12}
Category 2	π_{21}	π_{22}
Total	1	1

Test hypothesis of homogeneity H_0: $\pi_{11} = \pi_{12}$
using two (small) independent samples

Sample counts

	Population 1	Population 2	Total
Category 1	n_{11}	n_{12}	$n_{1.}$
Category 2	n_{21}	n_{22}	$n_{2.}$
Sample sizes	$n_{.1}$	$n_{.2}$	$n_{..}$

Conditional null distribution $n_{11} \sim \mathrm{Hg}(N, n, p)$

$N = n_{..}, n = n_{.1}, Np = n_{1.}, Nq = n_{2.}$

$\mathrm{P}(n_{11} = k) = \frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}}$

$\max(0, n - Nq) \le k \le \min(n, Np)$

Ex 1: sex bias in promotion

Data: 48 copies of the same file
24 labeled as "male" and other 24 labeled as "female"

Test H_0: $\pi_{11} = \pi_{12}$ no sex bias against
H_1: $\pi_{11} > \pi_{12}$ males are favored

	Male	Female	
Promote	$n_{11} = 21$	$n_{12} = 14$	$n_{1.} = 35$
Hold file	$n_{21} = 3$	$n_{22} = 10$	$n_{2.} = 13$
	$n_{.1} = 24$	$n_{.2} = 24$	$n_{..} = 48$

Reject H_0 for large n_{11}

null distribution $\mathrm{P}(n_{11} = k) = \frac{\binom{35}{k}\binom{13}{24-k}}{\binom{48}{24}}$, $11 \leq k \leq 24$

$\mathrm{P}(n_{11} \leq 14) = \mathrm{P}(n_{11} \geq 21) = 0.025$

Significant evidence of sex bias

one-sided $P = 0.025$, two-sided $P = 0.05$

2. χ^2-test of homogeneity

Population proportions

	Pop. 1	Pop. 2	...	Pop. J
Category 1	π_{11}	π_{12}	...	π_{1J}
Category 2	π_{21}	π_{22}	...	π_{2J}
...	...	...	...	...
Category I	π_{I1}	π_{I2}	...	π_{IJ}
Total	1	1	...	1

Homogeneity = all J distributions are equal

H_0: $(\pi_{11}, ..., \pi_{I1}) = (\pi_{12}, ..., \pi_{I2}) = ... = (\pi_{1J}, ..., \pi_{IJ})$

Test H_0 against H_1: $\pi_{ij} \neq \pi_{il}$ for some (i, j, l)

using sample counts in J independent samples

$(n_{1j}, \ldots, n_{Ij}) \sim \mathrm{Mn}(n_{.j}; \pi_{1j}, \ldots, \pi_{Ij})$, $j = 1, \ldots, J$

	Pop. 1	Pop. 2	...	Pop. J	Total
Category 1	n_{11}	n_{12}	...	n_{1J}	$n_{1.}$
Category 2	n_{21}	n_{22}	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...
Category I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I.}$
Sample sizes	$n_{.1}$	$n_{.2}$	...	$n_{.J}$	$n_{..}$

Under H_0 the MLE of π_{ij}

pooled sample proportion $\hat{\pi}_{ij} = n_{i.}/n_{..}$

expected cell counts $\hat{E}_{ij} = n_{.j} \cdot \hat{\pi}_{ij} = n_{i.}n_{.j}/n_{..}$

Reject H_0 for large

$$X^2 = \Sigma_{i=1}^{I} \Sigma_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{.j}/n_{..}}$$

Approximate null distribution

$X^2 \overset{a}{\sim} \chi^2_{\text{df}}$ with df $= (I-1)(J-1)$

df = no. counts – no. estimates $= (I-1)J - (I-1)$

Ex 2: small cars and personality

Attitude toward small cars for different personality types

	Cautious	Midroad	Explorer	Total
Favorable	79(61.6)	58(62.2)	49(62.2)	186
Neutral	10(8.9)	8(9.0)	9(9.0)	27
Unfavorable	10(28.5)	34(28.8)	42(28.8)	86
Total	99	100	100	299

$X^2 = 27.24$, df $= 4$, $\chi^2_{4,0.005} = 14.86$

Reject H_0 at 0.5% level

cautious people are more favourable to small cars

3. Chi-square test of independence

One population

two classifications A, B with numbers of classes I, J

Population proportions

Classes	B_1	B_2	$\ldots$	B_J	Total
A_1	π_{11}	π_{12}	$\ldots$	π_{1J}	$\pi_{1.}$
A_2	π_{21}	π_{22}	$\ldots$	π_{2J}	$\pi_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_I	π_{I1}	π_{I2}	$\ldots$	π_{IJ}	$\pi_{I.}$
Total	$\pi_{.1}$	$\pi_{.2}$	$\ldots$	$\pi_{.J}$	1

Null hypothesis of independence H_0: $\|\pi_{ij}\| = \|\pi_{i.}\pi_{.j}\|$

against H_1: $\|\pi_{ij}\| \neq \|\pi_{i.}\pi_{.j}\|$ (dependence)

using a cross-classified sample $\|n_{ij}\| \sim \mathrm{Mn}(n_{..}; \|\pi_{ij}\|)$

Classes	B_1	B_2	$\ldots$	B_J	Total
A_1	n_{11}	n_{12}	$\ldots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\ldots$	n_{2J}	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_I	n_{I1}	n_{I2}	$\ldots$	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.J}$	$n_{..}$

Under H_0 the MLE of π_{ij} are $\hat{\pi}_{ij} = \frac{n_{i.}}{n_{..}} \cdot \frac{n_{.j}}{n_{..}}$

expected cell counts $\hat{E}_{ij} = n_{..} \cdot \hat{\pi}_{ij} = n_{i.}n_{.j}/n_{..}$

$\mathrm{df} = (IJ - 1) - ((I-1) + (J-1)) = (I-1)(J-1)$

Apply the same test procedure as with homogeneity test

Homogeneity $\mathrm{P}(A = i|B = j) = \mathrm{P}(A = i)$ for all (i, j)
equality of conditional distributions = independence

Ex 3: marital status and educational level

Contingency table

Education	Married once	Married > once	Total
College	550 (523.8)	61(87.2)	611
No College	681(707.2)	144(117.8)	825
Total	1231	205	1436

H_0: no relationship between mar. status and ed. level

$X^2 = 16.01$, df $= 1$ can use normal distribution table

$\sqrt{16.01} = 4.001$, P $< 0.1\%$ reject H_0

4. Matched-pairs designs

Ex 4: Hodgkin's disease and tonsills

2×2 cross-classification

$D =$ **D**iseased (affected), $\bar{D} =$ unaffected

$X =$ e**X**posed (tonsillectomy), $\bar{X} =$ non-exposed

H_0: tonsillectomy has no influence on disease onset

Three sampling designs

simple random sampling

a prospective study (X-sample and $\bar{X}$-sample)

a retrospective study (D-sample and $\bar{D}$-sample)

Retrospective design catches affected subjects

first two designs bring mostly unaffected

incidence of Hodgkin's disease is 2 in 10 000

Two datasets

VGD-1971

	X	$\bar{X}$
D	67	34
$\bar{D}$	43	64

JJ-1972

	X	$\bar{X}$
D	41	44
$\bar{D}$	33	52

$X^2_{\text{VGD}} = 14.29$, $X^2_{\text{JJ}} = 1.53$, df $= 1$
$\text{P}(X^2_{\text{VGD}} \geq 14.29) \approx 2(1 - \Phi(\sqrt{14.29})) = 0.0002$
$\text{P}(X^2_{\text{JJ}} \geq 1.53) \approx 2(1 - \Phi(\sqrt{1.53})) = 0.215$

JJ-data violates the assumption of independent samples
$n = 85$ sibling $(D, \bar{D})$-pairs, same sex, close age
matched-pairs design

Four classes of sibling pairs

	exposed $\bar{D}$-sib	unexposed $\bar{D}$-sib	
exposed D-sibling	$n_{11} = 26$	$n_{12} = 15$	41
unexposed D-sibling	$n_{21} = 7$	$n_{22} = 37$	44
total	33	52	85

McNemar's test

2×2 cross-classified population

π_{11}	π_{12}	$\pi_{1.}$
π_{21}	π_{22}	$\pi_{2.}$
$\pi_{.1}$	$\pi_{.2}$	1

MLE: $\hat{\pi}_{11} = \frac{n_{11}}{n}$, $\hat{\pi}_{22} = \frac{n_{22}}{n}$, $\hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12}+n_{21}}{n}$
test statistic $X^2 = \sum \frac{(n_{ij} - n\hat{\pi}_{ij})^2}{n\hat{\pi}_{ij}} = \frac{(n_{12} - n_{21})^2}{n_{12}+n_{21}}$

Reject H_0: $\pi_{1.} = \pi_{.1}$ or H_0: $\pi_{12} = \pi_{21}$ for large X^2
approximate null distribution is χ^2_1, df $= 4 - 1 - 2$

Ex 4: Hodgkin: JJ-data $X^2_{\text{McNemar}} = 2.91$, P $= 0.09$

4. Odds ratios

Odds and probability of a random event A

$\text{odds}(A) = \text{P}(A)/\text{P}(\bar{A})$

$\text{odds}(A) \approx \text{P}(A)$ for small $\text{P}(A)$ $\quad \text{P}(A) = \frac{\text{odds}(A)}{1+\text{odds}(A)}$

$\text{odds}(A|B) = \text{P}(A|B)/\text{P}(\bar{A}|B) = \text{P}(AB)/\text{P}(\bar{A}B)$

Odds ratio for a pair of random events

$\Delta_{AB} = \frac{\text{odds}(A|B)}{\text{odds}(A|\bar{B})} = \frac{\text{P}(AB)\text{P}(\bar{A}\bar{B})}{\text{P}(\bar{A}B)\text{P}(A\bar{B})}$

Measure of dependence

if $\Delta_{AB} = 1$, events A and B are independent

if $\Delta_{AB} > 1$, $\text{P}(A|B) > \text{P}(A|\bar{B})$

if $\Delta_{AB} < 1$, $\text{P}(A|B) < \text{P}(A|\bar{B})$

$$\boxed{\Delta_{AB} = \Delta_{BA},\ \Delta_{A\bar{B}} = \frac{1}{\Delta_{AB}}}$$

Ex 4: Hodgkin's disease and tonsills

Conditional probabilities and observed counts

in a retrospective study like VGD-1971

	X	$\bar{X}$	
D	$\text{P}(X|D)$	$\text{P}(\bar{X}|D)$	1
$\bar{D}$	$\text{P}(X|\bar{D})$	$\text{P}(\bar{X}|\bar{D})$	1

	X	$\bar{X}$	
D	n_{00}	n_{01}	$n_{0.}$
$\bar{D}$	n_{10}	n_{11}	$n_{1.}$

Odds ratio $\Delta_{DX} = \frac{\text{P}(X|D)\text{P}(\bar{X}|\bar{D})}{\text{P}(\bar{X}|D)\text{P}(X|\bar{D})}$

measures the influence of tonsillectomy on

Hodgkin's disease

Estimated odds ratio $\hat{\Delta} = \frac{(n_{00}/n_{0.})(n_{11}/n_{1.})}{(n_{01}/n_{0.})(n_{10}/n_{1.})} = \frac{n_{00}n_{11}}{n_{01}n_{10}}$

VGD-data

$\hat{\Delta} = \frac{65\cdot 64}{43\cdot 34} = 2.93 \quad \text{odds}(D|X) = 2.93 \cdot \text{odds}(D|\bar{X})$