Serik Sagitov, Chalmers Tekniska Högskola, April 23, 2014

Chapter 13. The analysis of categorical data

1 Fisher's exact test

Population proportions for categorical data

	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: $\pi_{11} = \pi_{12}$, $\pi_{21} = \pi_{22}$ using two 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	$n_{.1}$	$n_{.2}$	$n_{..}$

Use n_{11} as a test statistic. Conditionally on $n_{1.}$ the null distribution is hylergeometric $n_{11} \sim \text{Hg}(N, n, p)$ with parameters $N = n_{..}, n = n_{.1}, Np = n_{1.}, Nq = n_{2.}$

$$\text{P}(n_{11} = k) = \frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}}, \quad \max(0, n - Nq) \le k \le \min(n, Np).$$

Example: sex bias in promotion
Data: 48 copies of the same file with 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. Observed data

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

Reject H_0 for large n_{11} using the null distribution $\text{P}(n_{11} = k) = \frac{\binom{35}{k}\binom{13}{24-k}}{\binom{48}{24}}$, $11 \le k \le 24$. Since $\text{P}(n_{11} \le 14) = \text{P}(n_{11} \ge 21) = 0.025$ we find a one-sided $P = 0.025$, and a two-sided $P = 0.05$. Significant evidence of sex bias, reject the null hypothesis.

2 χ^2-test of homogeneity

Population proportions: IJ parameters with $J(I-1)$ independent parameters

	Population 1	Population 2	...	Population 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

Null hypothesis of homogeneity meaning that 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

	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	$n_{.1}$	$n_{.2}$	...	$n_{.J}$	$n_{..}$

J independent multinomial distributions $(n_{1j}, \ldots, n_{Ij}) \sim \text{Mn}(n_{.j}; \pi_{1j}, \ldots, \pi_{Ij})$, $j = 1, \ldots, J$.
Under the H_0 the MLE of π_{ij} are the pooled sample proportion $\hat{\pi}_{ij} = n_{i\cdot}/n_{..}$. These yield the expected cell counts $\hat{E}_{ij} = n_{\cdot j} \cdot \hat{\pi}_{ij} = n_{i\cdot}n_{\cdot j}/n_{..}$ and the χ^2-test statistic formula

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

Reject H_0 for large values of X^2 using the approximate null distribution $X^2 \overset{a}{\sim} \chi^2_{\text{df}}$ with $\text{df} = (I - 1)(J - 1)$, wich is obtained as $\text{df} = J(I - 1) - (I - 1) = (I - 1)(J - 1)$.

df = no. independent counts – no. independent parameters estimated from the data

Example: 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

The observed test statistic is $X^2 = 27.24$. With df = 4 it is larger than $\chi^2_{4,0.005} = 14.86$. Conclusion: reject H_0 at 0.5% significance level. Cautious people are more favorable to small cars.

3 Chi-square test of independence

One population cross-classified with respect to two classifications A, B with numbers of classes I, J. IJ population proportions with $IJ - 1$ of them independent.

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

Null hypothesis of independence H_0: $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for all pairs (i,j) to be tested against H_1: $\pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j}$ for at least one pair (i,j) (dependence). Data: a cross-classified sample

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

A multinomial distribution in the matrix form $\|n_{ij}\| \sim \mathrm{Mn}(n_{\cdot\cdot}; \|\pi_{ij}\|)$. Under H_0 the MLE of π_{ij} are $\hat{\pi}_{ij} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \cdot \frac{n_{\cdot j}}{n_{\cdot\cdot}}$ implying the same expected cell counts as before $\hat{E}_{ij} = n_{\cdot\cdot} \cdot \hat{\pi}_{ij} = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ with the same df $= (IJ-1) - ((I-1)+(J-1)) = (I-1)(J-1)$.
Conclusion: the same χ^2 test procedure for homogeneity test and for the independence test.

Homogeneity: $\mathrm{P}(A=i|B=j) = \mathrm{P}(A=i)$ for all (i,j) is equivalent to independence: $\mathrm{P}(A=i, B=j) = \mathrm{P}(A=i)\mathrm{P}(B=j)$ for all (i,j)

Ex¡mple: marital status and educational level
A 2×2 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 the marital status and the education level. Observed $X^2 = 16.01$. With df = 1 we can use the normal distribution table, since $Y \sim \chi_1^2$ is equivalent to $\sqrt{Y} \sim \mathrm{N}(0,1)$ so that

$$\mathrm{P}(Y > z_{\alpha/2}^2) = \mathrm{P}(\sqrt{Y} > z_{\alpha/2}) + \mathrm{P}(-\sqrt{Y} < -z_{\alpha/2}) = 2\mathrm{P}(\sqrt{Y} > z_{\alpha/2}) = \alpha.$$

As $\sqrt{16.01} = 4.001$ is more than 3 standard deviations, we conclude that a P-value is less that 0.1% and we reject the null hypothesis of independence.

4 Matched-pairs designs

Example: Hodgkin's disease and tonsillectomy
Test H_0: "tonsillectomy has no influence on disease onset" using a 2×2 cross-classification:
D = **D**iseased (affected), $\bar{D}$ = unaffected
X = e**X**posed (tonsillectomy), $\bar{X}$ = non-exposed
Three sampling designs: simple random sampling, a prospective study (X-sample and $\bar{X}$-sample), a retrospective study (D-sample and $\bar{D}$-sample).
Since the disease is rare, incidence of Hodgkin's disease is 2 in 10 000, one usually gets something like

random sampling: results in counts like

	X	$\bar{X}$
D	0	0
$\bar{D}$	0	n

prospective case-control study: results in counts like

	X	$\bar{X}$
D	0	0
$\bar{D}$	n_1	n_2

retrospective case-control study: results in counts like

	X	$\bar{X}$
D	n_{11}	n_{12}
$\bar{D}$	n_{21}	n_{22}

Two retrospective case-control study datasets

VGD-1971

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

and

JJ-1972

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

resulted in two chi-square tests of homogeneity $X^2_{\text{VGD}} = 14.29$, $X^2_{\text{JJ}} = 1.53$, df = 1. They give two strikingly different P-values:

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

The JJ-data should not be analyzed using the chi-square test of homogeneity. The JJ-data is based on a matched-pairs design and violates the assumption of independent samples:
$n = 85$ sibling $(D, \bar{D})$-pairs, same sex, close age.
A proper summary of the single bivariate sample distinguishes among four classes of $(D, \bar{D})$-pairs: (X, X), $(X, \bar{X})$, $(\bar{X}, X)$, $(\bar{X}, \bar{X})$:

	$X\|\bar{D}$	$\bar{X}\|\bar{D}$	Total
$X\|D$	$n_{11} = 26$	$n_{12} = 15$	41
$\bar{X}\|D$	$n_{21} = 7$	$n_{22} = 37$	44
Total	33	52	85

Notice that this contingency table contains more information than the previous one.

McNemar's test

A model for the data: 2×2 cross-classified population

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

The relevant null hypothesis is H_0: $\pi_{1.} = \pi_{.1}$ or equivalently H_0: $\pi_{12} = \pi_{21}$.
The MLEs of the population frequencies under the null hypothesis:

$$\hat{\pi}_{11} = \frac{n_{11}}{n}, \quad \hat{\pi}_{22} = \frac{n_{22}}{n}, \quad \hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12} + n_{21}}{2n}$$

results in the test statistic

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - n\hat{\pi}_{ij})^2}{n\hat{\pi}_{ij}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

whose approximate null distribution is χ^2_1 with df $= 4 - 1 - 2$. Reject the H_0 for large values of X^2.

Example: Hodgkin. The JJ-data gives $X^2_{\text{McNemar}} = 2.91$ and a P-value = 0.09 smaller than 0.215. Too few informative (off-diagonal) observations.

5 Odds ratios

Odds and probability of a random event A: $\text{odds}(A) := \frac{\text{P}(A)}{\text{P}(\bar{A})}$ and $\text{P}(A) = \frac{\text{odds}(A)}{1+\text{odds}(A)}$. Notice that $\text{odds}(A) \approx \text{P}(A)$ for small $\text{P}(A)$.
Conditional odds: $\text{odds}(A|B) := \text{P}(A|B)/\text{P}(\bar{A}|B) = \text{P}(AB)/\text{P}(\bar{A}B)$. Odds ratio for a pair of events

$$\Delta_{AB} := \frac{\text{odds}(A|B)}{\text{odds}(A|\bar{B})} = \frac{\text{P}(AB)\text{P}(\bar{A}\bar{B})}{\text{P}(\bar{A}B)\text{P}(A\bar{B})}, \quad \Delta_{AB} = \Delta_{BA}, \quad \Delta_{A\bar{B}} = \frac{1}{\Delta_{AB}}$$

is a measure of dependence between the two random events
if $\Delta_{AB} = 1$, then events A and B are independent,
if $\Delta_{AB} > 1$, then $\text{P}(A|B) > \text{P}(A|\bar{B})$ so that B increases probability of A, in particular, $\Delta_{AA} = \infty$,
if $\Delta_{AB} < 1$, then $\text{P}(A|B) < \text{P}(A|\bar{B})$ so that B decreases probability of A, in particular, $\Delta_{A\bar{A}} = 0$.

Retrospective case-control studies
Conditional probabilities and observed counts

	X	$\bar{X}$	Total
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	$\bar{X}$	Total
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 eXposition to a certain factor on the onset of the Disease in question. Estimated odds ratio

$$\hat{\Delta}_{DX} = \frac{(n_{00}/n_{0.})(n_{11}/n_{1.})}{(n_{01}/n_{0.})(n_{10}/n_{1.})} = \frac{n_{00}n_{11}}{n_{01}n_{10}}$$

Example: Hodgkin.
VGD-1971 study: $\hat{\Delta}_{DX} = \frac{67\cdot 64}{43\cdot 34} = 2.93$. Conclusion: tonsillectomy increases the odds for Hodgkin's onset by factor 2.93.
JJ-1972 study: $\hat{\Delta}_{DX} = \frac{41\cdot 52}{33\cdot 44} = 1.47$.