Serik Sagitov, Chalmers and GU, February 1, 2016

Chapter 13. The analysis of categorical data

Categorical data appear in the form of a contingency table containing the sample counts for various combinations of categories. Here the statistical models are based on the multinomial distribution.

Joint probabilities $\pi_{ij} = \mathrm{P}(A = i, B = j)$, marginal probabilities $\pi_{i\cdot} = \mathrm{P}(A = i)$, $\pi_{\cdot j} = \mathrm{P}(B = j)$, conditional probabilities $\pi_{i|j} = \mathrm{P}(A = i|B = j) = \frac{\pi_{ij}}{\pi_{\cdot j}}$.

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

	B_1	B_2	$\ldots$	B_J
A_1	$\pi_{1\|1}$	$\pi_{1\|2}$	$\ldots$	$\pi_{1\|J}$
A_2	$\pi_{2\|1}$	$\pi_{2\|2}$	$\ldots$	$\pi_{2\|J}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_I	$\pi_{I\|1}$	$\pi_{I\|2}$	$\ldots$	$\pi_{I\|J}$
Total	1	1	$\ldots$	1

The left table corresponds to a single population distribution for a cross-classification $A \times B$.
The null hypothesis of independence states no relationship between the two factors A and B
H_0: $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for all pairs (i, j) is a nested model with $I - 1 + J - 1$ degrees of freedom.
The right table describes J population distributions for a common classification A.
The null hypothesis of homogeneity states the equality of J population distributions
H_0: $\pi_{i|j} = \pi_i$ for all pairs (i, j) is a nested model with $I - 1$ degrees of freedom.

The hypothesis of homogeneity is equivalent to the hypothesis of independence.

1 Fisher's exact test

Consider two populations distinguishing between two categories. Then the null hypothesis of homogeneity has the form H_0: $\pi_{1|1} = \pi_{1|2}$. Data is given by two independent samples summarised as a 2×2 table of sample counts

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

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

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

Example (gender bias)
Data: 48 copies of the same file with 24 files labeled as "male" and the other 24 labeled as "female". Two possible outcomes: promote or hold file.

	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	$n_{.1} = 24$	$n_{.2} = 24$	$n_{..} = 48$

We wish to test H_0: $\pi_{1|1} = \pi_{1|2}$, no gender bias, against H_1: $\pi_{1|1} > \pi_{1|2}$, males are favoured. Fisher's test would reject H_0 in favour of the one-sided alternative H_1 for large values of $K = n_{11}$ having the null distribution

$$\mathrm{P}(K = k) = \frac{\binom{35}{k}\binom{13}{24-k}}{\binom{48}{24}} = \frac{\binom{35}{35-k}\binom{13}{k-11}}{\binom{48}{24}}, \quad 11 \le k \le 24.$$

This is a symmetric distribution with $\mathrm{P}(K \le 14) = \mathrm{P}(K \ge 21) = 0.025$ so that a one-sided $P = 0.025$, and a two-sided $P = 0.05$.

2 Chi-square test of homogeneity

J independent samples taken from J distributions. The table of IJ observed counts:

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

Multinomial distributions $(n_{1j}, \dots, n_{Ij}) \sim \mathrm{Mn}(n_{\cdot j}; \pi_{1|j}, \dots, \pi_{I|j})$, $j = 1, \dots, J$.
Under the hypothesis of homogeneity $H_0 : \pi_{i|j} = \pi_i$, the maximum likelihood estimates of π_i are the pooled sample proportion $\hat{\pi}_i = n_{i\cdot}/n_{\cdot\cdot}$, $i = 1, \dots, I$. Usinf these estimates we compute the expected cell counts $\hat{E}_{ij} = n_{\cdot j} \cdot \hat{\pi}_i = n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ and the chi-square test statistic becomes

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

Reject H_0 for large values of X^2 using the approximate null distribution $X^2 \overset{a}{\sim} \chi^2_{\mathrm{df}}$ with

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

Example (small cars and personality)
Attitude toward small cars for different personality types. The table of observed (expected) counts:

	Cautious	Middle-of-the-road	Explorer	Total
Favourable	79(61.6)	58(62.2)	49(62.2)	186
Neutral	10(8.9)	8(9.0)	9(9.0)	27
Unfavourable	10(28.5)	34(28.8)	42(28.8)	86
Total	99	100	100	299

The chi-square test statistic is $X^2 = 27.24$, and $\mathrm{df} = (3-1) \cdot (3-1) = 4$. After comparing X^2 with $\chi^2_{4,0.005} = 14.86$, we reject the hypothesis of homogeneity at 0.5% significance level. Persons who saw themselves as cautious conservatives are more likely to express a favourable opinion of small cars.

3 Chi-square test of independence

Data: a single cross-classifying sample is summarised in terms of the observed counts, whose joint distribution is multinomial $(n_{ij}) \sim \text{Mn}(n_{\cdot\cdot}; (\pi_{ij}))$.

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

The maximum likelihood estimates of $\pi_{i\cdot}$ and $\pi_{\cdot j}$ are $\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}}$ and $\hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}$. Therefore, under the hypothesis of independence $\hat{\pi}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}^2}$ implying the same expected cell counts as before
$\hat{E}_{ij} = n_{\cdot\cdot}\hat{\pi}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$ with the same df $= (IJ-1)-(I-1+J-1) = (I-1)(J-1)$.

The same chi-square test rejection rule for the homogeneity test and independence test.

Example (marital status and educational level)
A sample is drawn from a population of married women. Each observation is placed in a 2×2 contingency table depending on woman's educational level and her marital status.

	Married only once	Married more than once	Total
College	550 (523.8)	61(87.2)	611
No college	681(707.2)	144(117.8)	825
Total	1231	205	1436

The observed chi-square test statistic is $X^2 = 16.01$. With df = 1 we can use the normal distribution table, since $Z^2 \sim \chi_1^2$ is equivalent to $Z \sim \text{N}(0,1)$. Thus

$$\text{P}(X^2 > 16.01) \approx \text{P}(|Z| > 4.001) = 2(1 - \Phi(4.001)).$$

We see that a P-value is less that 0.1%, and we reject the null hypothesis of independence.
College-educated women, once they do marry, are much less likely to divorce.

4 Matched-pairs designs

Example (Hodgkin's disease)
To test H_0: tonsillectomy has no influence on the onset of Hodgkin's disease, researchers use cross-classification data of the form

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

where the counts distinguish among sampled individual who are either D = affected (have the **D**isease) or $\bar{D}$ = unaffected, and either X = e**X**posed (had tonsillectomy) or $\bar{X}$ = non-exposed

Three possible sampling designs:
simple random sampling,
prospective study: take an X-sample and a control $\bar{X}$-sample, then watch who gets affected,
retrospective study: take a D-sample and a control $\bar{D}$-sample, then find who had been exposed.

Since the Hodgkin disease is rare, the incidence of 2 in 10 000, random samples would give counts like $\begin{pmatrix} 0 & 0 \\ 0 & n \end{pmatrix}$, while prospective case-control studies usually would give $\begin{pmatrix} 0 & 0 \\ n_1 & n_2 \end{pmatrix}$.

Two retrospective case-control studies

Study A: Vianna, Greenwald, Davis (1971), and study B: Johnson and Johnson (1972)

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

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

resulted in two chi-square tests of homogeneity $X_A^2 = 14.29$, $X_B^2 = 1.53$, df $= 1$. They give two strikingly different P-values:

$$P(X_A^2 \geq 14.29) \approx 2(1 - \Phi(\sqrt{14.29})) = 0.0002, \qquad P(X_B^2 \geq 1.53) \approx 2(1 - \Phi(\sqrt{1.53})) = 0.215.$$

The study B was based on a matched-pairs design violating the assumption of the chi-square test of homogeneity. The sample consisted of $n = 85$ sibling pairs having same sex and close age: one of the siblings was affected the other not.
A proper summary of the study B sample distinguishes among four groups of sibling pairs: (X, X), $(X, \bar{X})$, $(\bar{X}, X)$, $(\bar{X}, \bar{X})$

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

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

McNemar's test

Consider data obtained by matched-pairs design for the population distribution

	unaffected X	unaffected $\bar{X}$	Total
affected X	π_{11}	π_{12}	$\pi_{1.}$
affected $\bar{X}$	π_{21}	π_{22}	$\pi_{2.}$
$\pi_{.1}$	$\pi_{.2}$	1	

The relevant null hypothesis is not the hypothesis of independence but rather
H_0: $\pi_{1.} = \pi_{.1}$ or equivalently H_0: $\pi_{12} = \pi_{21} = \pi$ for an unspecified π.
The maximum likelihood estimates for the population frequencies under the null hypothesis

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

yield a new chi-square test statistic

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

whose approximate null distribution is χ_1^2. Reject the H_0 for large values of X^2_{McNemar}.

Example (Hodgkin's disease)
The data of study B give $X^2_{\text{McNemar}} = 2.91$ and a P-value $= 0.09$ which is much smaller than that of 0.215 computed using the test of homogeneity. Too few informative, only $n_{12}+n_{21} = 22$, observations.

5 Odds ratios

Odds and probability of a random event A: $\quad \text{odds}(A) = \frac{\text{P}(A)}{\text{P}(\bar{A})}$ 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 for A given B: $\quad \text{odds}(A|B) = \frac{\text{P}(A|B)}{\text{P}(\bar{A}|B)} = \frac{\text{P}(AB)}{\text{P}(\bar{A}B)}$.

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, $\Delta_{A\bar{A}} = 0$.

Odds ratios for case-control studies
Return to 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_{11}	n_{12}	$n_{1\cdot}$
$\bar{D}$	n_{21}	n_{22}	$n_{2\cdot}$

The corresponding 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

$$\hat{\Delta}_{DX} = \frac{(n_{11}/n_{1\cdot})(n_{22}/n_{2\cdot})}{(n_{12}/n_{1\cdot})(n_{21}/n_{2\cdot})} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Example (Hodgkin's disease)
Study A gives the odds ratio $\hat{\Delta}_{DX} = \frac{67\cdot 64}{43\cdot 34} = 2.93$.
Conclusion: tonsillectomy increases the odds for Hodgkin's onset by factor 2.93.
Study B gives the odds ratio $\hat{\Delta}_{DX} = \frac{41\cdot 52}{33\cdot 44} = 1.47$.