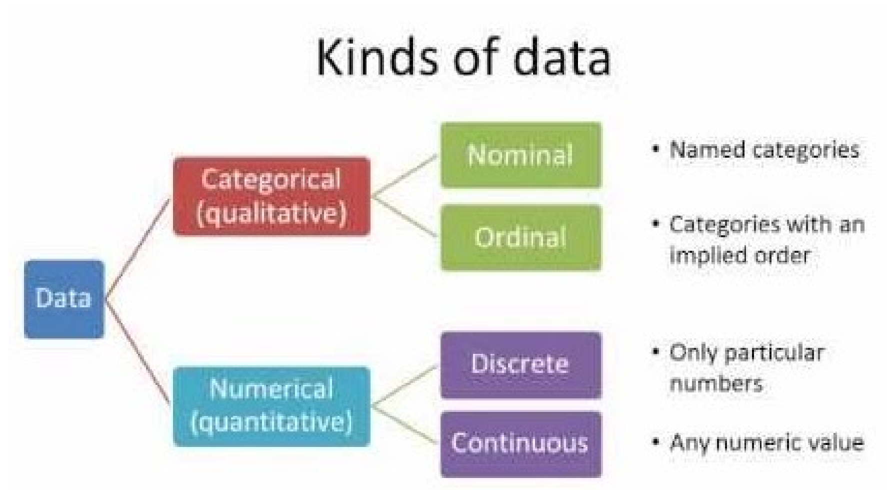
Slides 15: Categorical data tests

- Small sample test for proportion
- Multinomial models for categorical data
- Chi-squared test of homogeneity
- Chi-squared test of independence
- Fisher's exact test
- Matched-pairs design
- McNemar's test
- Odds ratios



Binomial model for the data value $X \sim \mathrm{Bin}(n, p)$. To test H_0: $p = p_0$ for a small n, use the exact null distribution $X \sim \mathrm{Bin}(n, p_0)$.

Example: extrasensory perception

A person is asked to guess the suits of 20 cards. The number of cards guessed correctly $X \sim \mathrm{Bin}(20, p)$.

For $H_0 : p = 0.25$ and $H_1 : p > 0.25$ the p-value is computed using Bin(20,0.25) distribution

x_{obs}	8	9	10	11
$\mathrm{P}(X \geq x)$	.101	.041	.014	0.004

Suppose $x_{\text{obs}} = 9$, then we reject H_0 at $\alpha = 5\%$ significance level.
In this case, the power function of the test is

p	0.27	0.30	0.40	0.5	0.60	0.70
$\mathrm{P}(X \geq 9)$	0.064	0.113	0.404	0.748	0.934	0.995

Question. What possible guessing strategy lies behind the one-sided alternative $H_1 : p < 0.25$?

Data: $n = 23480$ suicides in US, 1970. Is there a seasonal variation?

Multinomial model: the observed counts $\sim \mathrm{Mn}(n, p_1, \ldots, p_{12})$.

Month	O_j	Days	p_j^o	$E_j = np_j^o$	$O_j - E_j$
Jan	1867	31	0.085	1994	−127
Feb	1789	28	0.077	1801	−12
Mar	1944	31	0.085	1994	−50
Apr	2094	30	0.082	1930	164
May	2097	31	0.085	1994	103
Jun	1981	30	0.082	1930	51
Jul	1887	31	0.085	1994	-107
Aug	2024	31	0.085	1994	30
Sep	1928	30	0.082	1930	-2
Oct	2032	31	0.085	1994	38
Nov	1978	30	0.082	1930	48
Dec	1859	31	0.085	1994	-135

Simple H_0 of no seasonal effect: $H_0 : (p_1, \ldots, p_{12}) = (p_1^o, \ldots, p_{12}^o)$

The χ^2-test statistic: $\mathrm{X}^2 = \sum_j \frac{(O_j - E_j)^2}{E_j} = 47.4.$

Reject H_0 for very small p-value: 1-pchisq(47.4,df=11)= 0.000002.

Question. Which months give the largest deviations from H_0?

Example: marital status and educational level.

One sample of size $n = 1436$ is drawn from a population of married women: 2×2 contingency table , $I = J = 2$. Observed (expected) counts

	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

produce the chi-squared test statistic

$$\mathrm{X}^2 = \sum \frac{(\mathrm{obs}-\mathrm{exp})^2}{\mathrm{exp}} = 16.01$$

Since $Z \sim \mathrm{N}(0,1)$ is equivalent to $Z^2 \sim \chi^2_1$, we get under H_0

$$\mathrm{P}(\mathrm{X}^2 > 16.01) \approx \mathrm{P}(|Z| > 4.001) = 2(1 - \Phi(4.001)) = 0.00006.$$

Reject the null hypothesis of independence. College-educated women, once they marry, are less likely to divorce.

Question. How are the expected counts computed? Why $\mathrm{df} = 1$?

Consider a cross-classification for a pair of categorical factors A and B.

Factor A has I levels and factor B has J levels.

The joint distribution of a single cross-classification event (left table) and the conditional distributions (right table).

	b_1	b_2	...	b_J	Total
a_1	π_{11}	π_{12}	...	π_{1J}	$\pi_{1\cdot}$
a_2	π_{21}	π_{22}	...	π_{2J}	$\pi_{2\cdot}$
...	...	...	...	...	...
a_I	π_{I1}	π_{I2}	...	π_{IJ}	$\pi_{I\cdot}$
Total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	...	$\pi_{\cdot J}$	1

b_1	b_2	...	b_J
$\pi_{1\|1}$	$\pi_{1\|2}$	...	$\pi_{1\|J}$
$\pi_{2\|1}$	$\pi_{2\|2}$	...	$\pi_{2\|J}$
...	...	...	...
$\pi_{I\|1}$	$\pi_{I\|2}$	...	$\pi_{I\|J}$
1	1	...	1

$$\pi_{ij} = \mathrm{P}(A = a_i, B = b_j), \quad \pi_{i|j} = \mathrm{P}(A = a_i | B = b_j) = \frac{\pi_{ij}}{\pi_{\cdot j}}$$

Hypothesis of independence $\quad H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad$ for all pairs (i, j)

Hypothesis of homegeniety $\quad H_0 : \pi_{i|j} = \pi_i \quad$ for all pairs (i, j)

Question. Can you prove that these two are equivalent?

Consider a table of $I \times J$ observed counts obtained from J independent samples taken from J population distributions:

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

This model is described by J multinomial distributions

$$(N_{1j}, \dots, N_{Ij}) \sim \mathrm{Mn}(n_{\cdot j}; \pi_{1|j}, \dots, \pi_{I|j}), \quad j = 1, \dots, J.$$

The total df $= J(I-1)$ for J independent samples of size I.

Under the hypothesis of homogeneity $\qquad H_0 : \pi_{i|j} = \pi_i$ for all (i, j)

MLE π_i is the pooled sample proportion $\qquad \hat{\pi}_i = \frac{n_{i\cdot}}{n_{\cdot\cdot}}$

Expected cell counts $\qquad E_{ij} = n_{\cdot j} \cdot \hat{\pi}_i = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

$$\mathrm{X}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}} \text{ with } \mathrm{X}^2 \overset{H_0}{\approx} \chi^2_{\mathrm{df}}, \text{ and df} = (I-1)(J-1)$$

Example: small cars and personality

A car company studies how customers' attitude toward small cars relates to different personality types.

The next table summarises the 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-squared test statistic is

$$\mathrm{X}^2 = 27.24 \quad \text{with df} = (3-1)\cdot(3-1) = 4.$$

After comparing $\boldsymbol{\chi}^2$ with the table value $\chi^2_4(0.005) = 14.86$, 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.

Data: observed counts for a single cross-classifying sample

	b_1	b_2	$\dots$	b_J	Total
a_1	n_{11}	n_{12}	$\dots$	n_{1J}	$n_{1.}$
a_2	n_{21}	n_{22}	$\dots$	n_{2J}	$n_{2.}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
a_I	n_{I1}	n_{I2}	$\dots$	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.J}$	$n_{..}$

whose joint distribution is multinomial

$$(N_{11}, \dots, N_{IJ}) \sim \text{Mn}(n_{..}; \pi_{11}, \dots, \pi_{IJ})$$

MLEs of $\pi_{i.}$ and $\pi_{.j}$: $\hat{\pi}_{i.} = \frac{n_{i.}}{n_{..}}$ and $\hat{\pi}_{.j} = \frac{n_{.j}}{n_{..}}$

Under the hypothesis of independence $\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}^2}$

we get the same expected cell counts as before $E_{ij} = n_{..}\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$

with the same X^2 and the same approximate null distribution.

Test of independence $\quad \text{df} = (IJ - 1) - (I - 1 + J - 1) = (I - 1)(J - 1)$

Test of homogeneity $\quad \text{df} = J(I - 1) - (I - 1) = (I - 1)(J - 1)$

Fisher's exact test deals with the null hypothesis $H_0 : p_1 = p_2$, when the sample sizes are not sufficiently large for applying normal approximations for the binomial distributions. We summarise binary data of two independent samples in a 2×2 table of sample counts

	Sample 1	Sample 2	Total
Number of successes	x	y	$Np = x + y$
Number of failures	$n - x$	$m - y$	$Nq = n + m - x - y$
Sample sizes	n	m	$N = n + m$

Fisher's idea:use X as a test statistic conditionally on the total number of successes $x + y$. The null distribution of X is hypergeometric

$$X \sim \mathrm{Hg}(N, n, p)$$

with $N = n + m$ being interpreted as the number of balls in an urn and $Np = x + y$ as the number of black balls, meaning success as an outcome.

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

This distribution determines the rejection region of the test.

Data were collected after 48 copies of the same file with 24 files labeled as "male" and the other 24 labeled as "female" were sent to 48 experts.

	Male	Female	Total
Promote	21	14	35
Hold file	3	10	13
Total	24	24	48

p = probability to promote file. We wish to test

$$H_0 : p_1 = p_2 \text{ (no gender bias)}, \quad H_1 : p_1 > p_2 \text{ males are favoured.}$$

Reject H_0 in favour of H_1 for large values of x under the null distribution

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

This is a symmetric distribution with

$$\mathrm{P}(X \le 14) = \mathrm{P}(X \ge 21) = 0.025.$$

so that a one-sided p-value = 0.025, and a two-sided p-value = 0.05. We conclude that there is a significant evidence of sex bias, and reject H_0.

Hodgkin disease which very low incidence of 2 in 10 000. To test a possible influence of tonsillectomy on the onset of Hodgkin's disease, researchers use cross-classification data of the form

	X	X^c
D	n_{11}	n_{12}
D^c	n_{21}	n_{22}

where the four counts distinguish among sampled individual who are

either D = affected (have the **D**isease) or D^c = unaffected,
either X = e**X**posed (had tonsillectomy) or X^c = non-exposed.

Three possible sampling designs:

(1) simple random sampling: would give $n_{11} = n_{12} = n_{21} = 0$
(2) prospective study: would give $n_{11} = n_{12} = 0$
(3) retrospective study: take a affected-sample and a control unaffected-sample, then find who had been exposed in the past

Two retrospective case-control studies had produced opposite results of the chi-squared test of homogeneity.

Study A	X	X^c
D	67	34
D^c	43	64

Study B	X	X^c
D	41	44
D^c	33	52

Study A (Vianna, Greenwald, Davis, 1971) gave $\mathrm{X}_\mathrm{A}^2 = 14.29$ and the p-value was found to be very small

$$\mathrm{P}(\mathrm{X}_\mathrm{A}^2 \geq 14.29) \approx 2(1 - \Phi(\sqrt{14.29})) = 0.0002.$$

Study B (Johnson and Johnson, 1972) gave $\mathrm{X}_\mathrm{B}^2 = 1.53$ and the p-value was strikingly different

$$\mathrm{P}(\mathrm{X}_\mathrm{B}^2 \geq 1.53) \approx 2(1 - \Phi(\sqrt{1.53})) = 0.215.$$

It turned out that the study B was based on a design violating the assumption of the chi-squared test of homogeneity.

In study B, the data consisted of $m = 85$ 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, X^c), (X^c, X), (X^c, X^c)

	unaffected X	unaffected X^c	Total
affected X	$m_{11} = 26$	$m_{12} = 15$	41
affected X^c	$m_{21} = 7$	$m_{22} = 37$	44
Total	33	52	85

Notice that this contingency table contains more information than the previous one. An appropriate test in this setting is McNemar's test (see below). For the data of study B, the McNemar's test statistic is

$$\mathrm{X}^2 = \frac{(m_{12} - m_{21})^2}{m_{12} + m_{21}} = 2.91,$$

giving the p-value of $$\mathrm{P}(\mathrm{X}^2 \geq 2.91) \approx 2(1 - \Phi(\sqrt{2.91})) = 0.09$$

The correct p-value is much smaller than that of 0.215 computed using the test of homogeneity. Since there are very few informative, only $m_{12} + m_{21} = 22$, observations, more data is required.

Consider data of size m obtained by matched-pairs design from

	unaffected X	unaffected X^c	Total
affected X	p_{11}	p_{12}	$p_{1.}$
affected X^c	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

The null hypothesis is not the hypothesis of independence but rather

H_0: $p_{1.} = p_{.1}$, or equivalently, H_0: $p_{12} = p_{21} = p$ for an unspecified p

MLEs for the population frequencies under the null hypothesis are

$$\hat{p}_{11} = \frac{m_{11}}{m}, \quad \hat{p}_{22} = \frac{m_{22}}{m}, \quad \hat{p}_{12} = \hat{p}_{21} = \hat{p} = \frac{m_{12} + m_{21}}{2m}.$$

These yield the McNemar test statistic of the form

$$\mathrm{X}^2 = \sum_i \sum_j \frac{(m_{ij} - m\hat{p}_{ij})^2}{m\hat{p}_{ij}} = \frac{(m_{12} - m_{21})^2}{m_{12} + m_{21}},$$

whose approximate null distribution is χ_1^2. Here df $= 4 - 1 - 2 = 1$ because 2 independent parameters are estimated from the data.

Odds and probability of a random event A:

$$\text{odds}(A) = \frac{\mathrm{P}(A)}{\mathrm{P}(\bar{A})} \quad \text{and} \quad \mathrm{P}(A) = \frac{\text{odds}(A)}{1+\text{odds}(A)}$$

notice that for small $\mathrm{P}(A)$: $\text{odds}(A) \approx \mathrm{P}(A)$. Conditional odds

$$\text{odds}(A|B) = \frac{\mathrm{P}(A|B)}{\mathrm{P}(A^c|B)} = \frac{\mathrm{P}(AB)}{\mathrm{P}(A^cB)}.$$

Odds ratio for a pair of events defined by

$$\Delta_{AB} = \frac{\text{odds}(A|B)}{\text{odds}(A|B^c)} = \frac{\mathrm{P}(AB)\mathrm{P}(A^cB^c)}{\mathrm{P}(A^cB)\mathrm{P}(AB^c)},$$

has the properties

$$\Delta_{AB} = \Delta_{BA}, \quad \Delta_{AB^c} = \frac{1}{\Delta_{AB}}$$

and gives a measure of dependence between a pair of random events :

if $\Delta_{AB} = 1$, then events A and B are independent,
if $\Delta_{AB} > 1$, then $\mathrm{P}(A|B) > \mathrm{P}(A|B^c)$ so that B favors A,
if $\Delta_{AB} < 1$, then $\mathrm{P}(A|B) < \mathrm{P}(A|B^c)$ so that B hinders A.

Return to conditional probabilities and observed counts

	X	X^c	Total
D	$\mathrm{P}(X\|D)$	$\mathrm{P}(X^c\|D)$	1
D^c	$\mathrm{P}(X\|D^c)$	$\mathrm{P}(X^c\|D^c)$	1

	X	X^c	Total
D	n_{11}	n_{12}	$n_{1\cdot}$
D^c	n_{21}	n_{22}	$n_{2\cdot}$

The corresponding odds ratio

$$\Delta_{DX} = \frac{\mathrm{P}(X|D)\mathrm{P}(X^c|D^c)}{\mathrm{P}(X^c|D)\mathrm{P}(X|D^c)} = \frac{\mathrm{odds}(D|X)}{\mathrm{odds}(D|X^c)}$$

quantifies the influence of eXposition to a certain factor on the onset of the Disease in question. Estimated odds ratio

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

Example: Study A for Hodgkin's disease gives the odds ratio

$$\widehat{\Delta}_{DX} = \frac{67 \cdot 64}{43 \cdot 34} = 2.93$$

tonsillectomy increases the odds for Hodgkin's onset by factor 2.93.