Statistical tests for binary and categorical data

06.03.2024

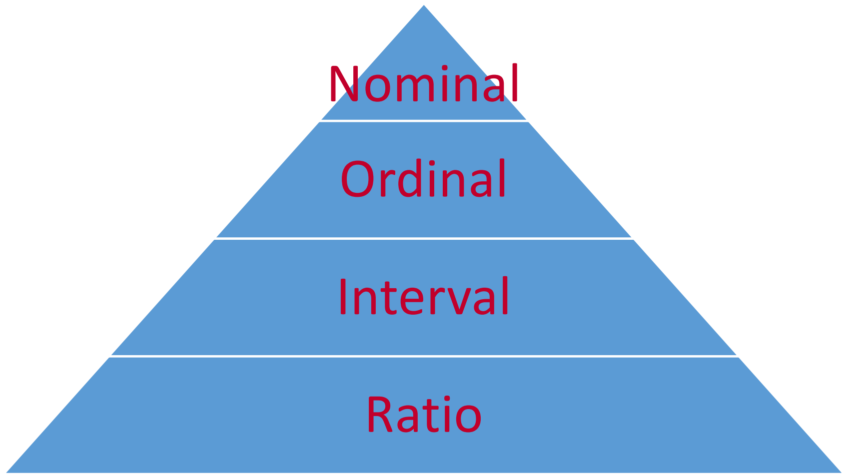
Chung Shing Rex Ha
Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI)
University Medical Center
rexha@uni-mainz.de

Agenda

- Binary and categorical data
- Test for one populations
- Test for two independent populations
- Test for two dependent (paired) populations

Binary and categorical data

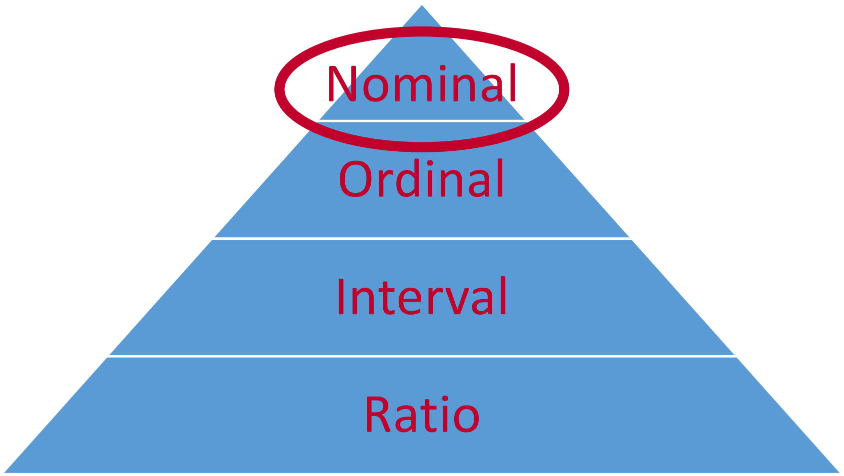
- Four levels of measurements



Binary and categorical data

- Four levels of measurement

- Nominal scale
  - Binary values (two classes)
  - Categorical values (more than two classes)



- Since they are discrete data, the assumption of normal distribution is not appropriate.
  - t-test is not feasible

Agenda

- Binary and categorical data
- Test for one populations
- Test for two independent populations
- Test for two dependent (paired) populations

Binary data - One population

- Back to the previous case study, suppose the pharmaceutical company want to know if 60% of the drug A users are over 65 years old
- First, we need to create a 1 x 2 contingency table

	Age <= 65	Age > 65	Total (n)
Drug A users	14	36	50

Binary data - One population

- Back to the previous case study, suppose the pharmaceutical company want to know if 60% of the drug A users are over 65 years old
- First, we need to create a 1 x 2 contingency table

	Age <= 65	Age > 65	Total (n)
Drug A users	14	36	50

- H_0: The proportion of drug A users over 65 years old is equal to 60%

$$\pi = \pi_0 = 0.6$$

  H_1: The proportion of drug A users over 65 years old is not equal to 60%

$$\pi \neq \pi_0 = 0.6$$

Binary data - One population

- Under the null hypothesis of independence, the cells of the tables follows follows binomial distribution

$$k \sim Binomial(n, p)$$

- Use exact binomial test to calculate the exact p-value
- For large samples, Wald test statistic

$$Z_w = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})/N}} \qquad Z_w \sim N(0, 1)$$

- $(1 - \alpha)\times 100\%$ confident interval (Wald interval):

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1 - \hat{\pi})/N}$$

Binary data - One population

- R code (using exact binomial test)

```
binom.test(x=36, n=50, p=0.6, alternative="two.sided")

     Exact binomial test

data:  36 and 50
number of successes = 36, number of trials = 50, p-value =
0.1113
alternative hypothesis: true probability of success is not
equal to 0.6
95 percent confidence interval:
 0.5750946 0.8376894
sample estimates:
probability of success
                  0.72
```

	Age <= 65	Age > 65	Total (n)
Drug A users	14	36	50

Agenda

- Binary and categorical data
- Test for one populations
- Test for two independent populations
- Test for two dependent (paired) populations

Binary data - Two independent populations

- Now the company wants to know if there is any association between age group and drug allocation in the study
- We create a 2 x 2 contingency table

	Age <= 65	Age > 65	Row Total
Drug A	14	36	50
Drug B	25	25	50
Column Total	39	61	100

Binary data - Two independent populations

- Now the company wants to know if there is any association between age group and drug allocation in the study
- We create a 2 x 2 contingency table

	Age <= 65	Age > 65	Row Total
Drug A	14	36	50
Drug B	25	25	50
Column Total	39	61	100

- H_0: The drug allocation is independent of the age groups
  $\Leftrightarrow$ The odds ratio is equal to 1, i.e., $\mathrm{OR} = 1$
  H_1: The drug allocation is not independent of the age groups
  $\Leftrightarrow$ The odds ratio is not equal to 1, i.e., $\mathrm{OR} \neq 1$

Binary data - Two independent populations

- Under the null hypothesis of independence, the cells of the tables follows hypergeometric distribution.
- We use Fisher's exact test to obtain the exact p-value
- R code

```
fisher.test(x=matrix(c(14, 36, 25, 25), nrow=2, ncol=2,
byrow=TRUE), alternative="two.sided")

     Fisher's Exact Test for Count Data


data:  matrix(c(14, 36, 25, 25), nrow = 2, ncol = 2, byrow =
TRUE)
p-value = 0.0397
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1552801 0.9613924
sample estimates:
odds ratio
 0.3926958
```

	Age <= 65	Age > 65	Row Total
Drug A	14	36	50
Drug B	25	25	50
Column Total	39	61	100

Categorical data - Two independent populations

- What if the age group is divided into three classes (below 30, between 30 and 65 and above 65)?
- We have a 2 x 3 contingency table instead

	Age < 30	$30 \leq$ Age $\leq$ 65	Age > 65	Row Total
Drug A	5	9	36	50
Drug B	10	15	25	50
Column Total	15	24	61	100

- H_0: The drug allocation is independent of the age groups

  H_1: The drug allocation is not independent of the age groups

Categorical data - Two independent populations

- Pearson's chi-squared test can be used to test the independence of two variables by comparing the observed frequencies O_i and the expected frequencies E_i
- Under the null hypothesis, values occur in each cell are uniformly distributed, i.e., with equal frequency: $E_i = \frac{Total\ sample\ size\ (N)}{number\ of\ cells\ (n)}$
- Test statistics $\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$
- χ^2 is asymptotically χ^2-distributed with degree of freedom (r - 1) x (c - 1)
- Application of Pearson's chi-squared test can be extended to R x C contingency table

Categorical data - Two independent populations

- R code

	Age < 30	30 ≤ Age ≤ 65	Age > 65	Row Total
Drug A	5	9	36	50
Drug B	10	15	25	50
Column Total	15	24	61	100

```
chisq.test(matrix(c(5, 9, 36, 10, 15, 25), nrow=2, ncol=3,
byrow=TRUE))

    Pearson's Chi-squared test


data:  matrix(c(5, 9, 36, 10, 15, 25), nrow = 2, ncol = 3,
byrow = TRUE)

X-squared = 5.1503, df = 2, p-value = 0.07614
```

H_0 is no longer rejected!

Agenda

- Binary and categorical data
- Test for one populations
- Test for two independent populations
- Test for two dependent (paired) populations

Two dependent populations

- For paired populations in case of 2 x 2 contingency table, we can use McNemar's test
- $H_0: p_b = p_c$

  $H_1: p_b \neq p_c$
- McNemar’s test statistic $\chi^2 = \frac{(b-c)^2}{b+c}$

	Test 2 positive	Test 2 negative	Row Total
Test 1 positive	a	b	a + b
Test 1 negative	c	d	c + d
Column Total	a + c	b + d	a + b + c + d

- For larger contingency tables, we can use Bowker test
- In R, McNemar’s test and Bowker test can be use via mcnemar.test()