Lecture 35 Chi-squared Test For Association/Independence

BIO210 Biostatistics

Xi Chen

Fall, 2024

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Clinical Infectious Diseases

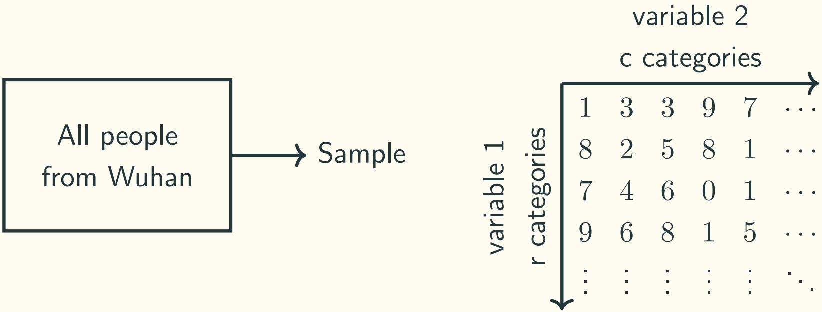
BRIEF REPORT

Relationship Between the ABO Blood Group and the Coronavirus Disease 2019 (COVID-19) Susceptibility

Jiao Zhao,[1,a] Yan Yang,[2,a] Hanping Huang,[3,a] Dong Li,[4,a] Dongfeng Gu,[1] Xiangfeng Lu,[5] Zheng Zhang,[2] Lei Liu,[2] Ting Liu,[3] Yukun Liu,[6] Yunjiao He,[1] Bin Sun,[1] Meilan Wei,[1] Guangyu Yang,[7,b] Xinghuan Wang,[8,b] Li Zhang,[3,b] Xiaoyang Zhou,[4,b] Mingzhao Xing,[1,b] and Peng George Wang[1,b]

[1]School of Medicine, The Southern University of Science and Technology, Shenzhen,

ABO Blood Types And COVID-19



$r \times c$ contingency table

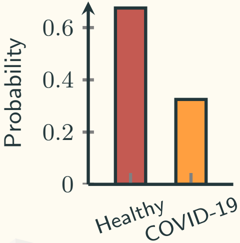
	A	**B**	**AB**	**O**
Healthy	1,188	920	336	1,250
COVID-19	670	469	178	458

Contingency Table

Extended $r \times c$ contingency table

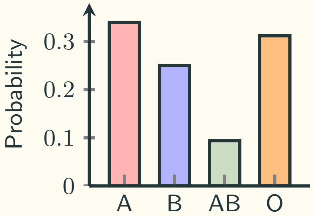
Y \ X	A	B	AB	O	Total
Healthy	1,188	920	336	1,250	3,694
COVID-19	670	469	178	458	1,775
Total	1,858	1,389	514	1,708	5,469

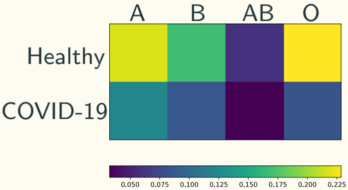
Marginal Distribution

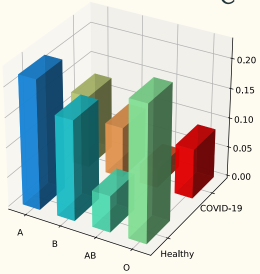


Joint Distribution

Marginal Distribution







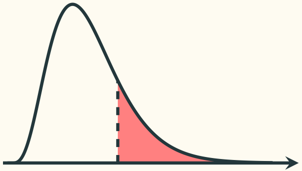
Chi-squared Test For Association/Independence

Question: Is there any association/relation between ABO blood groups and COVID-19 susceptibility ?

$$\begin{cases} H_0: & \text{No association/No relation} \\ H_1: & \text{There is an association/They are related} \end{cases}$$

$$\Updownarrow$$

$$\begin{cases} H_0: & \chi^2 = \sum_{\text{cells}} \frac{(O_i - E_i)^2}{E_i} = 0 \\ H_1: & \chi^2 \neq 0 \Rightarrow \chi^2 > 0 \end{cases}$$



Constructing The Expected Values In The Contingency Table

Observed:

	A	B	AB	O	Total
Healthy	1,188	920	336	1,250	3,694
COVID-19	670	469	178	458	1,775
Total	1,858	1,389	514	1,708	5,469

Expected (if H_0 were true):

	A	B	AB	O	Total
Healthy	$1858 \times \frac{3694}{5469}$	$1389 \times \frac{3694}{5469}$	$514 \times \frac{3694}{5469}$	$1708 \times \frac{3694}{5469}$	**3,694**
COVID-19	$1858 \times \frac{1775}{5469}$	$1389 \times \frac{1775}{5469}$	$514 \times \frac{1775}{5469}$	$1708 \times \frac{1775}{5469}$	**1,775**
Total	**1,858**	**1,389**	**514**	**1,708**	**5,469**

Constructing The Expected Values In The Contingency Table

Observed:

	A	B	AB	O	Total
Healthy	1,188	920	336	1,250	3,694
COVID-19	670	469	178	458	1,775
Total	1,858	1,389	514	1,708	5,469

Expected (if H_0 were true):

	A	B	AB	O	Total
Healthy	$3694 \times \frac{1858}{5469}$	$3694 \times \frac{1389}{5469}$	$3694 \times \frac{514}{5469}$	$3694 \times \frac{1708}{5469}$	**3,694**
COVID-19	$1775 \times \frac{1858}{5469}$	$1775 \times \frac{1389}{5469}$	$1775 \times \frac{514}{5469}$	$1775 \times \frac{1708}{5469}$	**1,775**
Total	**1,858**	**1,389**	**514**	**1,708**	**5,469**

Contingency Table

	A	B	AB	O	Total
Healthy	1,188	920	336	1,250	3,694
COVID-19	670	469	178	458	1,775
Total	1,858	1,389	514	1,708	5,469

v.s.

	Healthy	COVID-19	Total
A	1,188	670	1,858
B	920	469	1,389
AB	336	178	514
O	1,250	458	1,708
Total	3,694	1,775	5,469

- Equivalent
- Test statistics are exactly the same
- p-values are exactly the same

Chi-squared Tests p-value Calculation

Observed:

	A	B	AB	O	Total
Healthy	1,188	920	336	1,250	3,694
COVID-19	670	469	178	458	1,775
Total	1,858	1,389	514	1,708	5,469

Expected:

	A	B	AB	O	Total
Healthy	1254.97	938.19	347.18	1153.66	3,694
COVID-19	603.03	450.81	166.82	554.34	1,775
Total	1,858	1,389	514	1,708	5,469

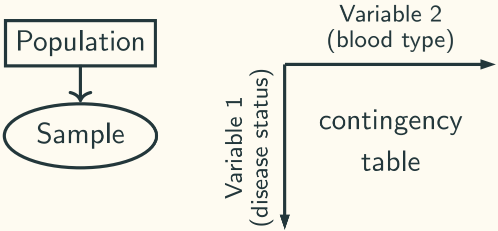
$$\chi^2 = \sum_{\text{cells}} \frac{(O_i - E_i)^2}{E_i} = 38.00$$

$$df = (r-1)(c-1) = 3$$

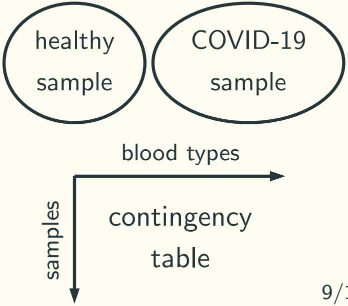
$$p = \mathbb{P}\left(\chi^2_3 \geqslant 38.00\right) = 2.82 \times 10^{-8}$$

Chi-squared Tests For Homogeneity vs Association/Independence

- Association/Independence:
  - H_0: no association between variables 1 & 2
  - H_1: association between variables 1 & 2



- Homogeneity:
  - H_0: from the same population/have the same distribution
  - H_1: from different populations/have different distributions



- The test statistics and p-values are exactly the same.
- The way of drawing samples and formulating hypotheses are different.
- Sometimes extremely similar or even indistinguishable.

Assumptions When Using Chi-squared Test

- Randomness, independence
- Because we used normal approximation for the binomial, we need large sample size: $np \geqslant 10$ and $nq \geqslant 10$. This means: all cells in the expected table should be at least 10.
- When normal approximation cannot be used: Fisher's exact test.

ABO Blood Types & COVID-19

Clinical Infectious Diseases

BRIEF REPORT

Relationship Between the ABO Blood Group and the Coronavirus Disease 2019 (COVID-19) Susceptibility

Jiao Zhao,[1,a] Yan Yang,[2,a] Hanping Huang,[3,a] Dong Li,[4,a] Dongfeng Gu,[1] Xiangfeng Lu,[5] Zheng Zhang,[2] Lei Liu,[2] Ting Liu,[3] Yukun Liu,[6] Yunjiao He,[1] Bin Sun,[1] Meilan Wei,[1] Guangyu Yang,[7,b] Xinghuan Wang,[8,b] Li Zhang,[3,b] Xiaoyang Zhou,[4,b] Mingzhao Xing,[1,b] and Peng George Wang[1,b]

[1]School of Medicine, The Southern University of Science and Technology, Shenzhen,

	A	B	AB	O	Total
Healthy	1,188	920	336	1,250	3,694
COVID-19	670	469	178	458	1,775
Total	1,858	1,389	514	1,708	5,469

$$\chi^2 = \sum_{\text{cells}} \frac{(O_i - E_i)^2}{E_i} = 38\,, \;\; p = \mathbb{P}\left(\chi^2_3 \geqslant 38\right) < 0.05$$

Conclusion: we reject H_0, which means the data suggest there is some relationship between ABO blood types and COVID-19 susceptibility.

What's next?: We can do *post hoc* tests.

Post hoc Tests

To correct for multiple testing: how many tests are we doing?

Rule of thumb: Define your question and decide the tests in advance.

1. Which blood types have association with COVID-19 ?

One category vs all the rest.

	A	**Non-A**	**Total**
Healthy	1,188	2,506	3,694
COVID-19	670	1,105	1,775
Total	1,858	3,611	5,469

	B	**Non-B**	**Total**
Healthy	920	2,774	3,694
COVID-19	469	1,306	1,775
Total	1,389	4,080	5,469

	AB	**Non-AB**	**Total**
Healthy	336	3,358	3,694
COVID-19	178	1,597	1,775
Total	514	4,955	5,469

	O	**Non-O**	**Total**
Healthy	1,250	2,444	3,694
COVID-19	458	1,317	1,775
Total	1,708	3,761	5,469

2. I don't know what I'm looking for, so I'm going to perform tests among all possible pairs:

- A **vs** non-A
- B **vs** non-B
- AB **vs** non-AB
- O **vs** non-O
- A & B **vs** AB & O
- A & O **vs** B & AB
- A & AB **vs** B & O
- B & AB **vs** A & O
- B & O **vs** A & AB
- ... ...

Post hoc Tests

One category vs all the rest

	A vs non-A	B vs non-B	AB vs non-AB	O vs non-O
χ^2	16.679	1.457	1.224	36.047
p	4.427×10^{-5}	0.227	0.268	1.926×10^{-9}

From the paper $\chi^2 = \sum_{\text{cells}} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$, Yates correction (Frank Yates)

	A vs non-A	B vs non-B	AB vs non-AB	O vs non-O
χ^2	16.431	1.378	1.117	35.674
p	5.045×10^{-5}	0.240	0.291	2.333×10^{-9}
OR	1.279	1.083	1.114	0.680
95% CI	[1.136, 1.440]	[0.952, 1.232]	[0.920, 1.349]	[0.599, 0.771]

Odds Ratio

	Exposed	Unexposed	Total
Disease	a	b	a+b
No disease	c	d	c+d
Total	a+c	b+d	n

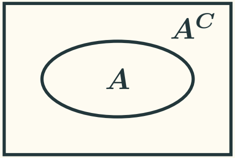
Odds ratio: $OR = \dfrac{P(\text{disease} \mid \text{exposed})/[1 - P(\text{disease} \mid \text{exposed})]}{P(\text{disease} \mid \text{unexposed})/[1 - P(\text{disease} \mid \text{unexposed})]}$

$$\hat{OR} = \frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Convenient to calculate, but confusing for understanding.

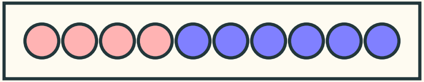
Probability vs Odds

Sample space Ω

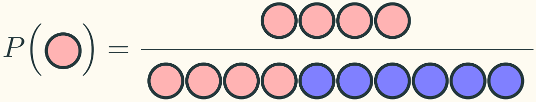


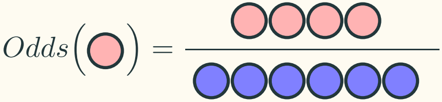
Probability: $P(A) = \dfrac{\text{area of } A}{(\text{area of } A) + (\text{area of } A^C)}$

Odds: a measurement in favour of an event, $\dfrac{P(A)}{P(A^C)} = \dfrac{P(A)}{1 - P(A)}$



Randomly choose a ball from the box:





Odds Ratio (OR)

	Category X	Category Y	Total
EOI	a	b	a+b
The rest	c	d	c+d
Total	a+c	b+d	n

Risk (Probability): Risk_{EOI}, Risk_{EOI} under X is $\frac{a}{a+c}$, Risk_{EOI} under Y is $\frac{b}{b+d}$

Relative risk (ratio of probability): $RR = \frac{\text{Risk}_{\text{EOI}} \text{ under X}}{\text{Risk}_{\text{EOI}} \text{ under Y}}$

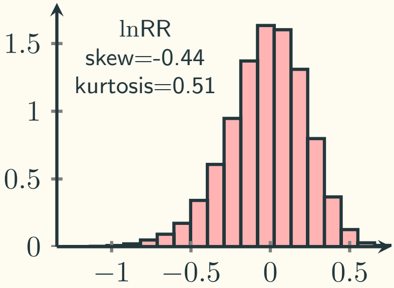
Odds (ratio of probability): Odds_{EOI}, Odds_{EOI} under X is $\frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$, Odds_{EOI} under Y is $\frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$

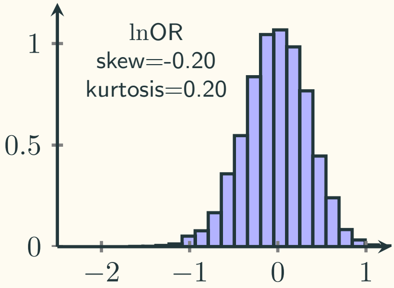
Odds ratio (ratio of ratio of probability): $OR = \frac{\text{Odds}_{\text{EOI}} \text{ under X}}{\text{Odds}_{\text{EOI}} \text{ under Y}} = \frac{a/c}{b/d} = \frac{ad}{bc}$

Sampling Distribution of $\ln$RR & $\ln$OR

	Category X	Category Y	Total
EOI	a	b	a+b
The rest	c	d	c+d
Total	a+c	b+d	n

10,000 simulations under the null hypothesis and keep records of RR and OR:





Sampling Distribution of $\ln$OR

	Category X	Category Y	Total
EOI	a	b	a+b
The rest	c	d	c+d
Total	a+c	b+d	n

- $\ln \hat{\text{OR}} \sim \mathcal{N}\left(0, \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)$
- 95% CI: $\ln \hat{\text{OR}} \pm Z_{0.025}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$ or $\ln \hat{\text{OR}} \pm 1.96\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
- 95% CI with continuity correction: $\ln \hat{\text{OR}} \pm 1.96\sqrt{\frac{1}{a+0.5} + \frac{1}{b+0.5} + \frac{1}{c+0.5} + \frac{1}{d+0.5}}$

Reproduce The Result

One category vs the rest *post hoc* tests (with continuity correction):

	A vs non-A		B vs non-B		AB vs non-AB		O vs non-O	
Healthy	1,188	2,506	920	2,774	336	3,358	1,250	2,444
COVID-19	670	1,105	469	1,306	178	1,597	458	1,317
χ^2	16.431		1.378		1.117		35.674	
p	5.045×10^{-5}		0.240		0.291		2.333×10^{-9}	
OR	0.782		0.924		0.898		1.471	
95% CI	[0.695, 0.880]		[0.812, 1.051]		[0.741, 1.087]		[1.296, 1.667]	

Results from the paper:

	A vs non-A	B vs non-B	AB vs non-AB	O vs non-O
OR	1.279	1.083	1.114	0.680
95% CI	[1.136, 1.440]	[0.952, 1.232]	[0.920, 1.349]	[0.599, 0.771]