MORU Tropical Health Network

Basic R programing

11-12 Jan 2021

Lecture 5 (10:45-12:00): Categorical data analysis

Dr. Palang Chotsiri

palang@tropmedres.ac

What is categorical data

- A categorical variable has a measurement scale consisting of a set of categories.

Gender

☐ Male

☐ Female

Smoking status

☐ Smoker

☐ Non-smoker

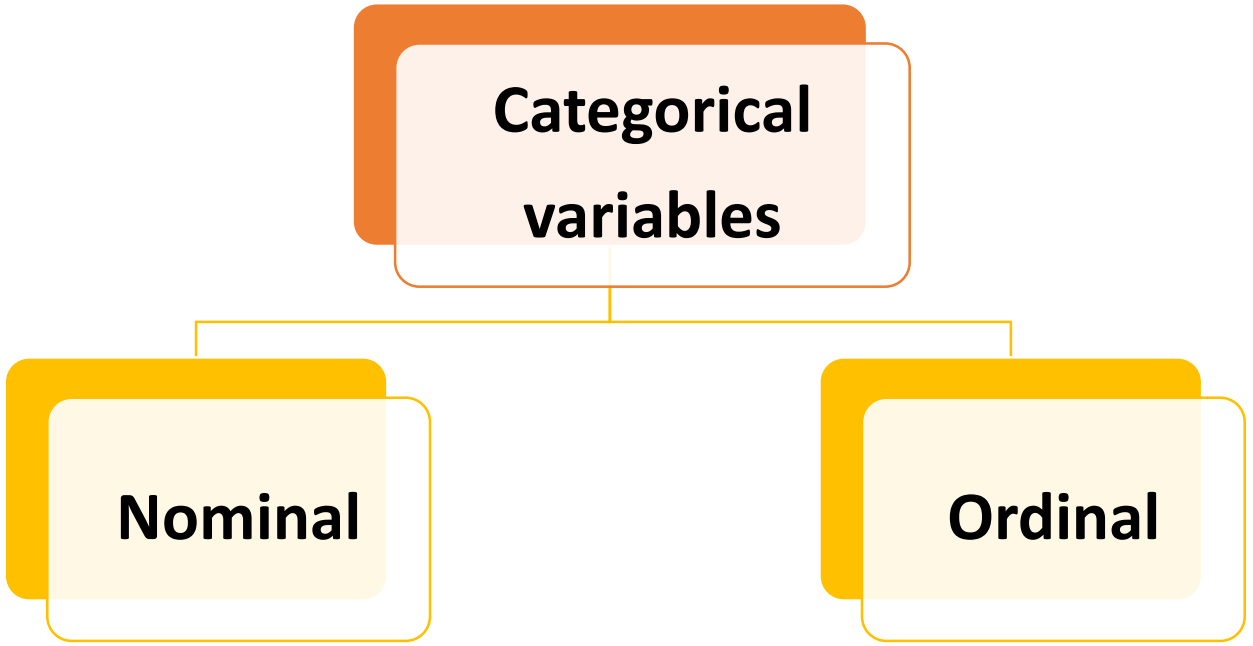
Treatment response

☐ Poor

☐ Fair

☐ Good

☐ Excellent



Gender

- ☐ Male
- ☐ Female

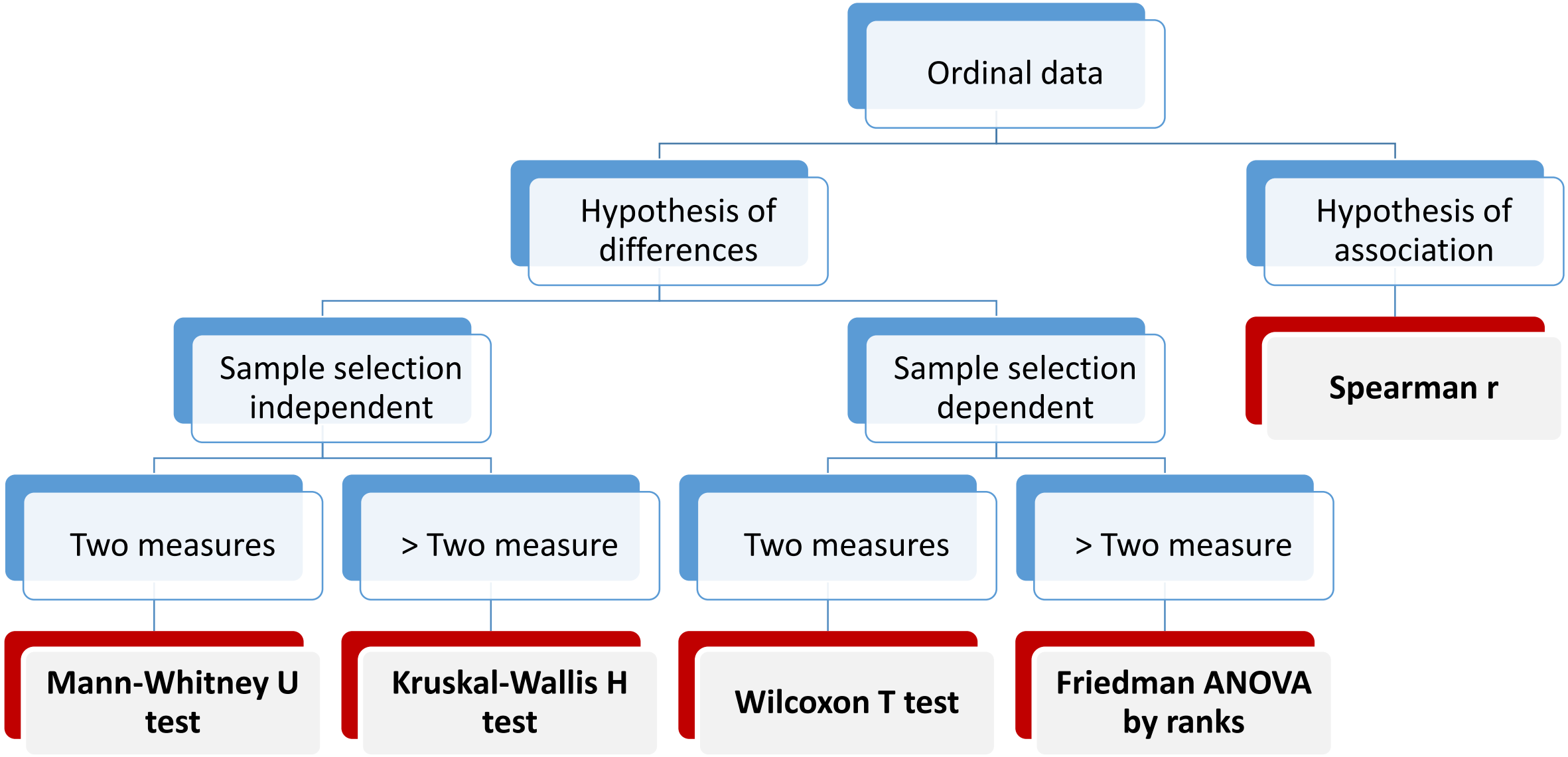
Smoking status

- ☐ Smoker
- ☐ Non-smoker

Treatment response

- ☐ Poor
- ☐ Fair
- ☐ Good
- ☐ Excellent

Ordinal variables



Statistical data analysis for nominal variables

One variable:

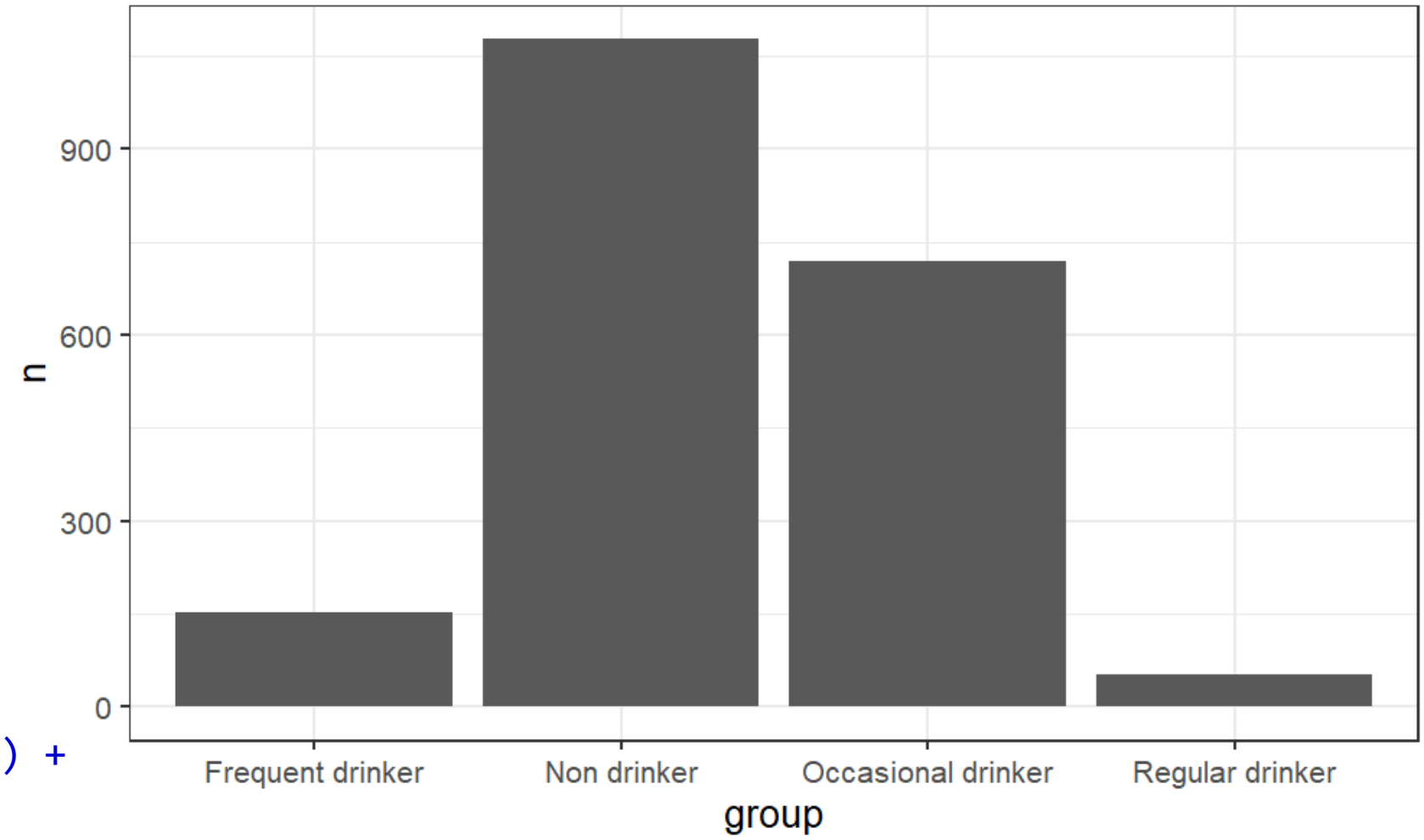
- Using table

Alcohol consumption	Non drinker	Occasional drinker	Frequent drinker	Regular drinker
n	1078	718	152	52

Statistical data analysis for nominal variables

One variable:

- Using table




```
ggplot(alcohol) +
  geom_col(aes(x=group,y=n)) +
  theme_bw()
```

How about two categorical variables

- Suppose now there are two categorical variables: X and Y, X has I categories & Y has J categories

- To study the relationship between X and Y ($X \Leftrightarrow Y$)

Contingency table

How about two categorical variables

Frequency count

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

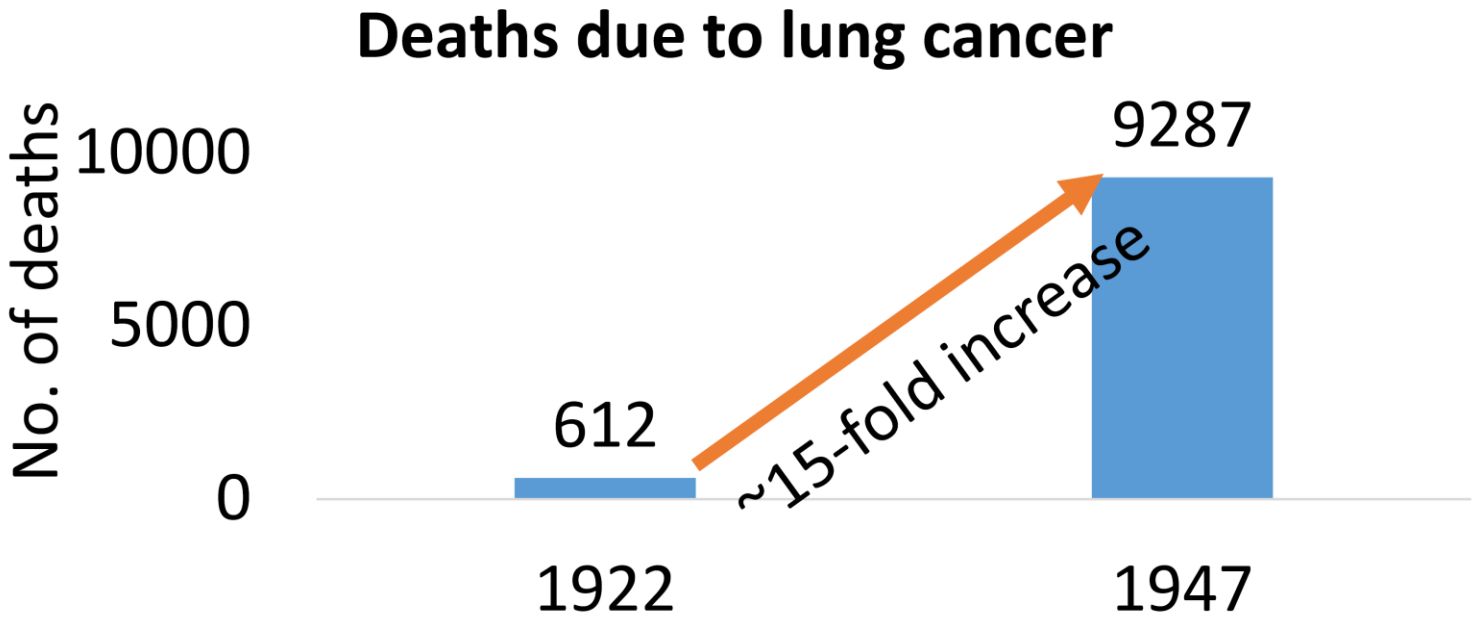
Marginal frequencies:
Row sum and Column sum

Total number of people

How about two categorical variables

- Examples:

- There was a dramatic increase in the no. of deaths due to lung cancer in England and Wales from 1922 to 1947.

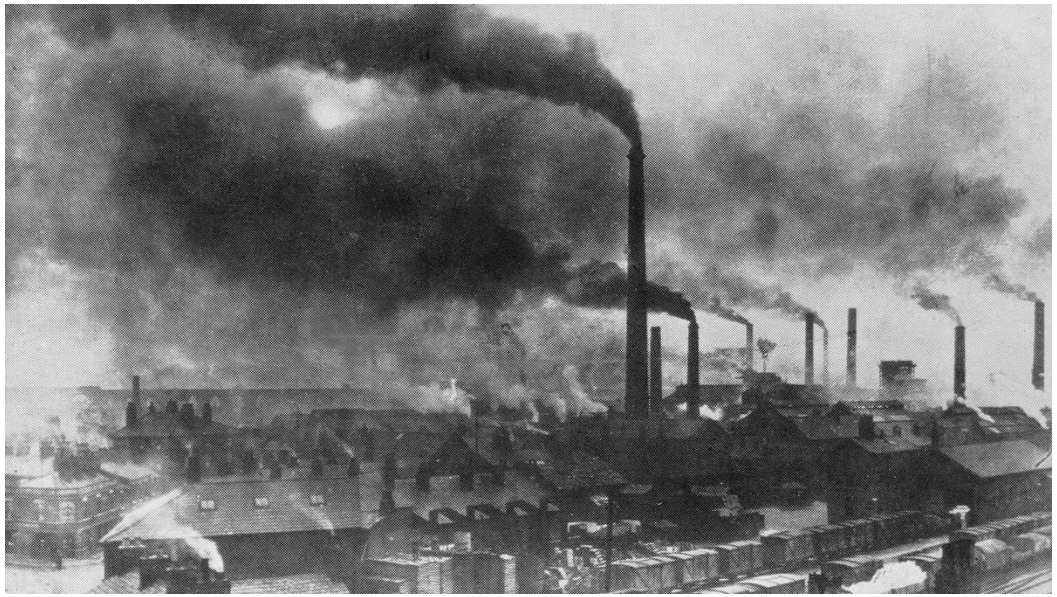


Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

How about two categorical variables

- Examples:

- The London Fog is one of the attractions of London in early 20th century. Many people believe that the main cause is a general atmospheric pollution.

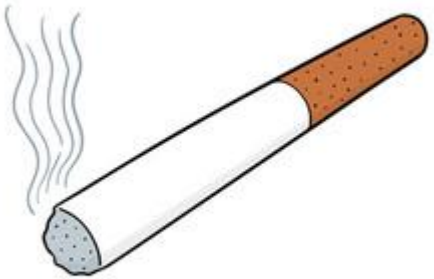


Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

How about two categorical variables

- Examples:

- However, two British researchers, Richard Doll and Bradford Hill, suspected **that the smoking of tobacco** might be the cause!
- To address the problem, they conducted a case-control study.



Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

How about two categorical variables

- Examples:

709 subjects with Lung cancer

688 are smokers

709 subjects with Lung cancer

650 are smokers

- All subjects were interviewed about their past smoking habits along many other potential risk factors in 20 hospitals in London.
- The cases and controls had similar demographic characteristics (age and gender).

Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

How about two categorical variables

- Examples:

- To study the relationship between lung cancer and smoking

Contingency table

- **The exposure X has two categories: smokers or non-smokers**
- **The outcome Y has two categories: lung cancer or no lung cancer**

→ We need 2×2 contingency table

Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

How about two categorical variables

- Examples:

- To study the relationship between lung cancer and smoking

Contingency table

- **The exposure X has two categories: smokers or non-smokers**
- **The outcome Y has two categories: lung cancer or no lung cancer**

→ We need 2×2 contingency table

Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

How about two categorical variables

- Examples:

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Marginal frequencies:
Row sum and Column sum

Total number of people

Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

How about two categorical variables

- Examples:

Exposure	Outcome		Total
	Yes	No	
Yes	688	650	1338
No	21	59	80
Total	709	709	1418

Marginal frequencies:
Row sum and Column sum

Total number of people

Doll, R. & Hill, A. B. Smoking and Carcinoma of the Lung. *Br. Med. J.***2,**739–748 (1950)

Joint, marginal and conditional probabilities

Gender	Heart disease		Total
	Yes	No	
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Joint probability (π_{ij}) : ex. $\Pr(Male, Heart\ disease)$

Joint, marginal and conditional probabilities

Gender	Heart disease		Total
	Yes	No	
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Marginal probability (π_i, π_j) : ex. $\Pr(Male)$

Joint, marginal and conditional probabilities

Gender	Heart disease		Total
	Yes	No	
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Marginal probability (π_i, π_j) : ex. $\Pr(Male)$

Joint, marginal and conditional probabilities

Gender	Heart disease		Total
	Yes	No	
Male	π_{11}	π_{12}	π_{1+}
Female	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Conditional probability : ex. $\Pr(Heart\ disease \mid Male)$

Example

	No wind	Some wind	Strong wind	Storm
No rain	0.1	0.2	0.05	0.01
Light rain	0.05	0.1	0.15	0.04
Heavy rain	0.05	0.1	0.1	0.05

Joint probability

- P(no wind, Light rain) = 0.05

Marginal probability

- P(no wind) = 0.1 + 0.05 + 0.05 = 0.2
- P(Light rain) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34

Conditional probability

- P(no wind | Light rain) = $\frac{0.05}{0.34}$ = 0.147

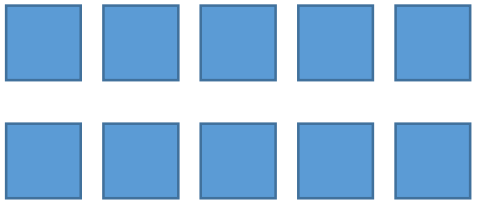
- P(Light rain | no wind) = $\frac{0.05}{0.2}$ = 0.25

Type of Observational Study

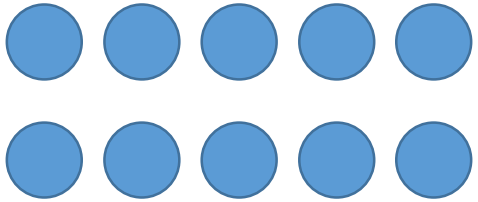
- Cohort study
- Case-control study
- Cross sectional study

Cohort study

Exposure = Yes



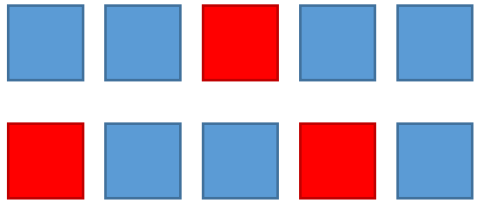
Exposure = No



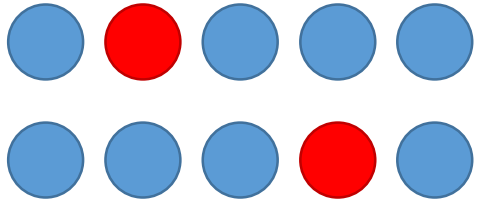
time

Cohort study

Exposure = Yes



Exposure = No



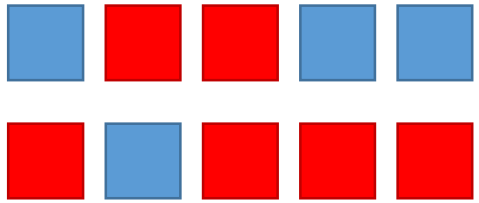
time

Cohort study

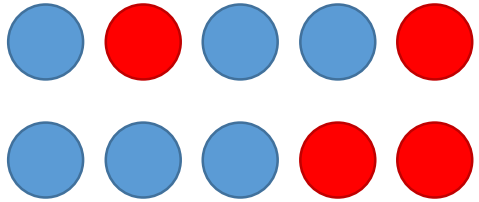
Exposure → Outcome

Incidence = no. of new case within a certain time period

Exposure = Yes



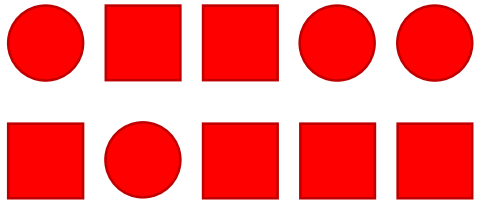
Exposure = No



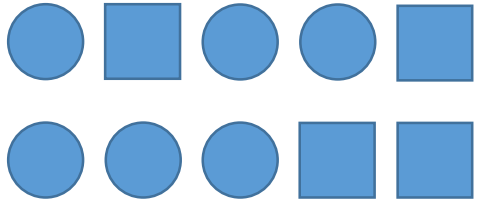
time

Case-Control Study

Outcome = Yes



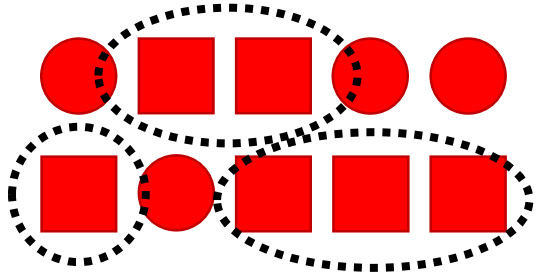
Outcome = No



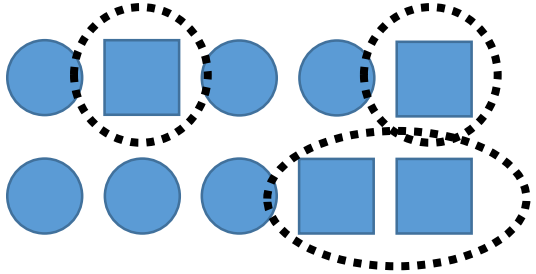
time

Case-Control Study

Outcome = Yes



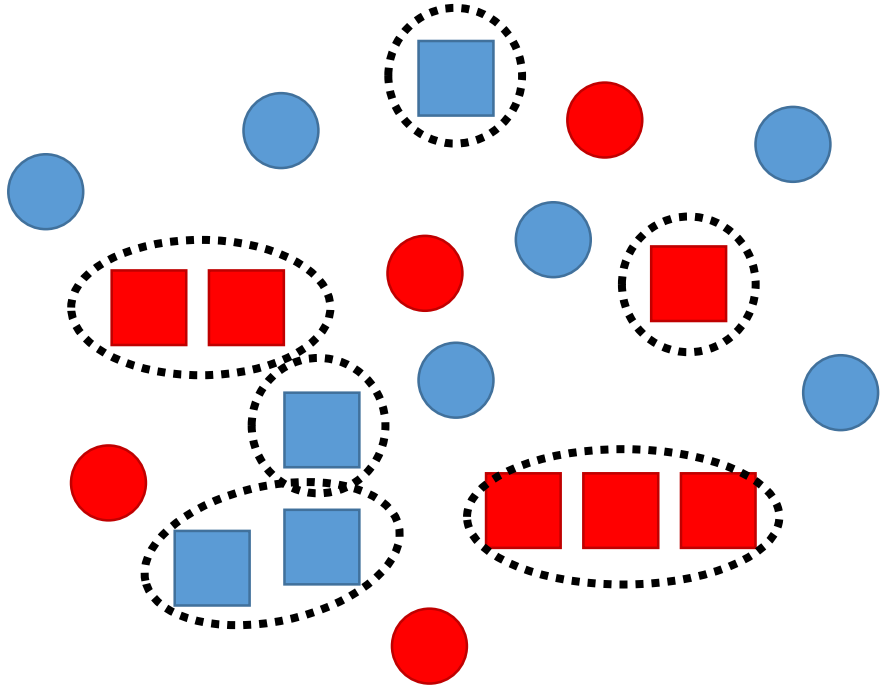
Outcome = No



Outcome → Exposure

time

Cross-Sectional Study



Exposure and outcome were selected at the same time!

Outcome ⇔ Exposure

Prevalence = no. of case/no. of people

time

Type of study	Type of association
Cohort study	Relative risk (RR)
Case-control study	Odds Ratio (OR)
Cross-sectional study	Relative risk (RR)

Relative Risk (RR)

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Risk a : Pr (having the outcome in the exposed group) $= \frac{a}{a+b}$

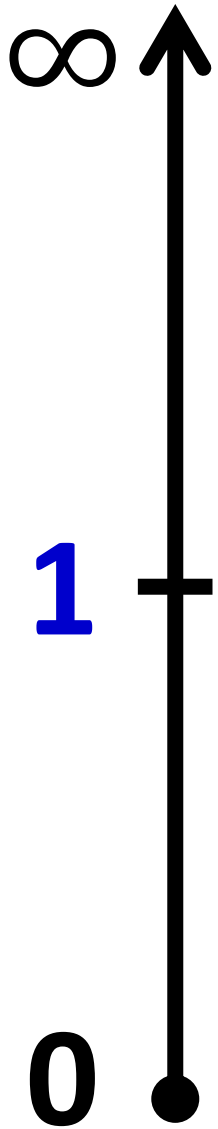
Risk c : Pr (having the outcome in the unexposed group) $= \frac{c}{c+d}$

Relative Risk (RR)

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$Relative\ Risk = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Relative Risk (RR)

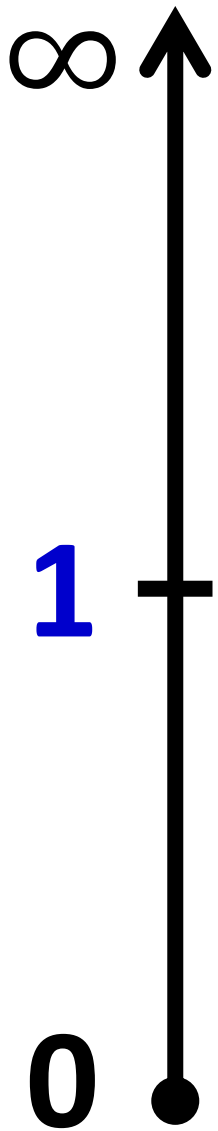


$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = 1$$

Risk between two groups are the same.

No association

Relative Risk (RR)

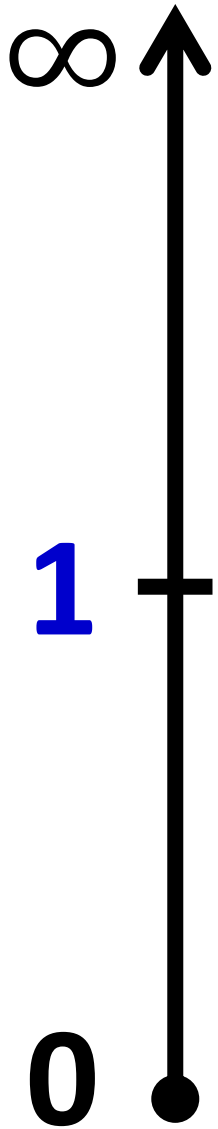


$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} > 1$$

Risk of exposed group > Risk of unexposed group

Being exposed → Higher risk

Relative Risk (RR)

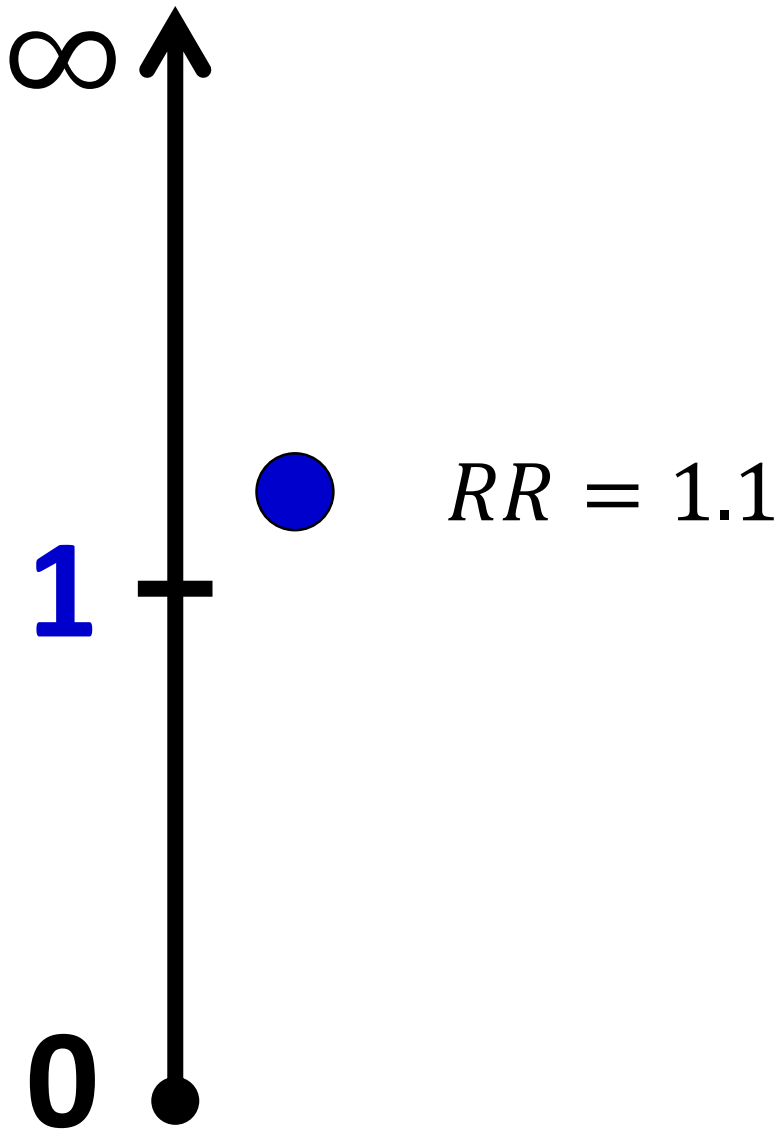


Risk of exposed group < Risk of unexposed group

Being exposed → Lower risk

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} < 1$$

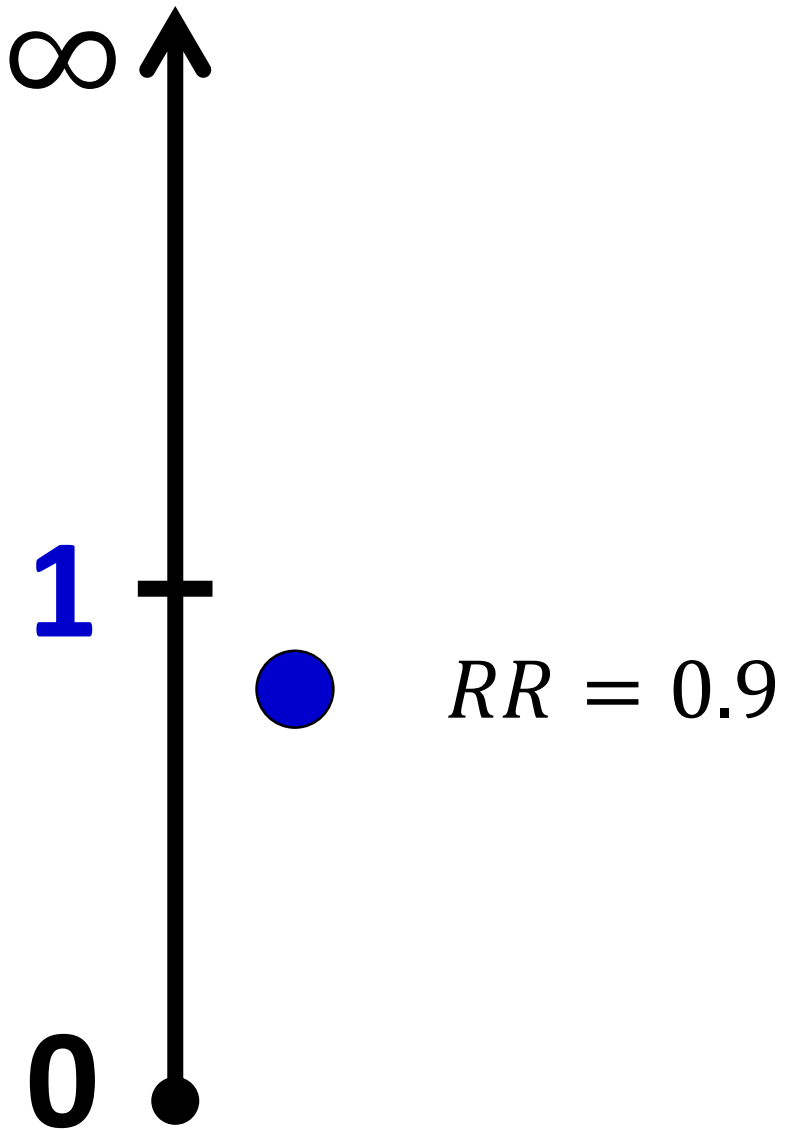
Relative Risk (RR)



$RR = 1.1$

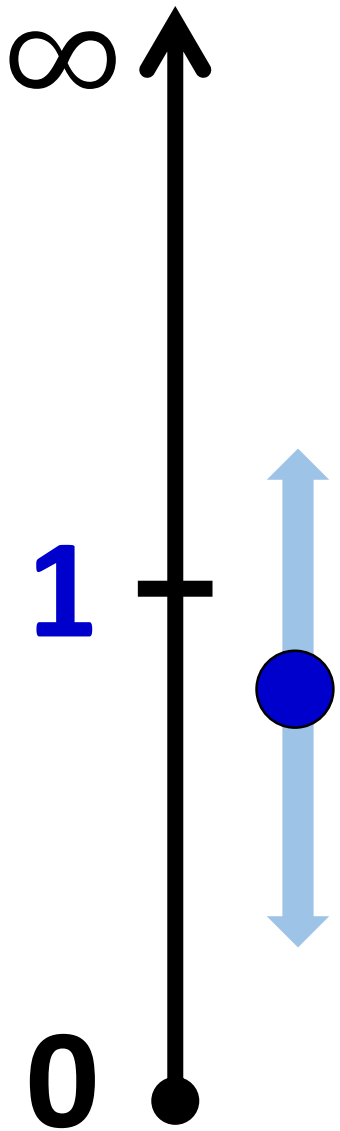
Is there any association?

Relative Risk (RR)



Is there any association?

Relative Risk (RR)

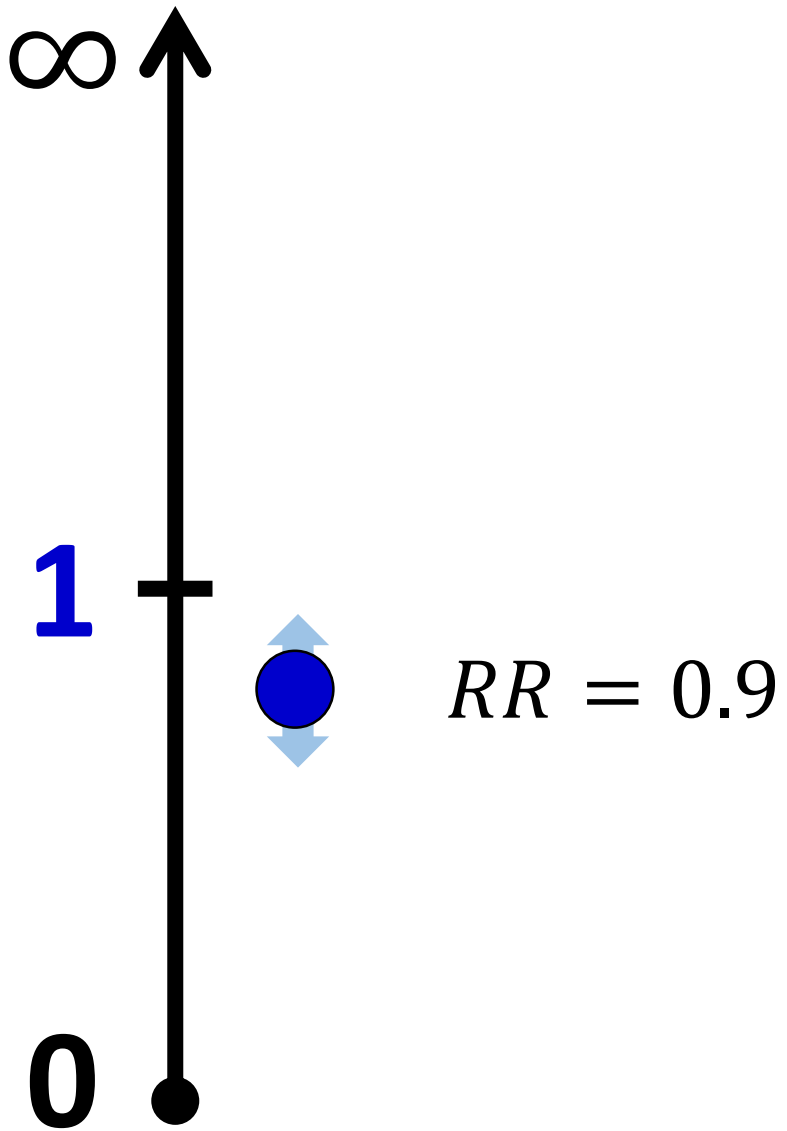


$RR = 0.9$

95% confidence interval

If 1 falls within 95% CI → No association

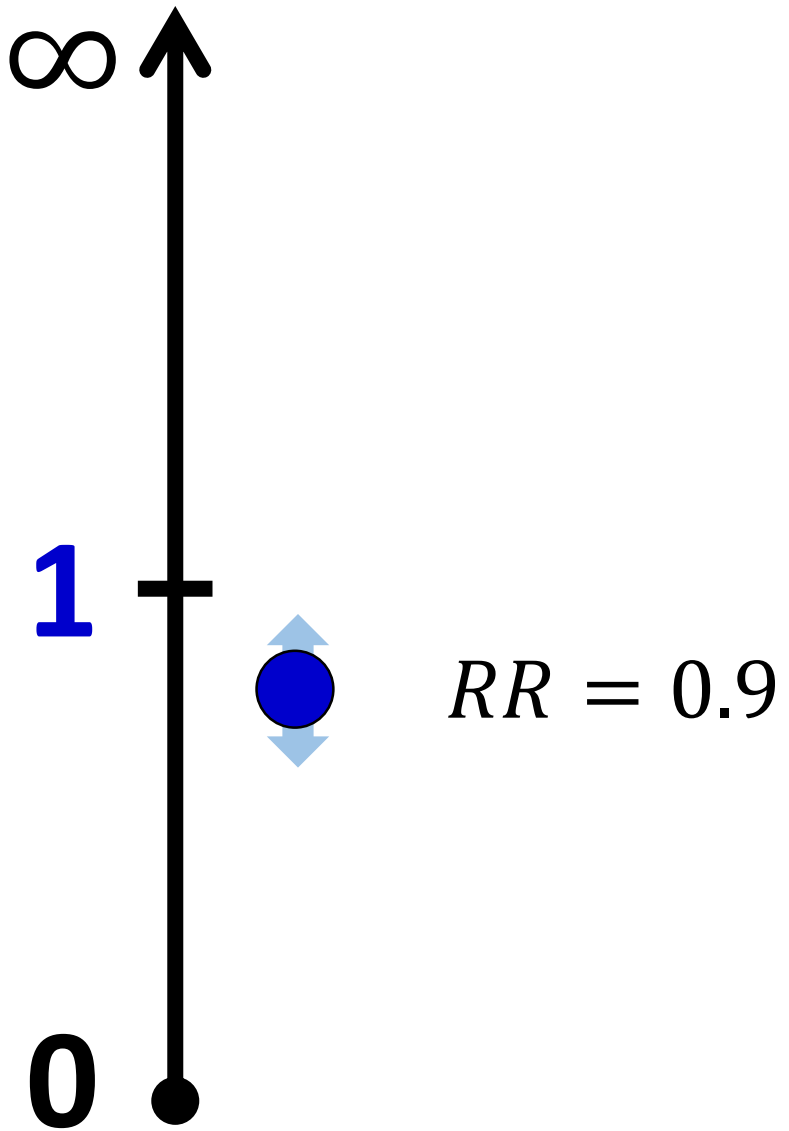
95% Confidence Interval



95% confidence interval

If not → there is an association

Relative Risk (RR)



$RR = 0.9$

~~$RR \pm 1.96\ SE$~~

95% CI of log(RR)

$$log\,(RR) \pm 1.96 \times SE\big(log(RR)\big)$$

95% CI of RR

$$e^{\left[log\,(RR) \pm 1.96 \times SE\left(log(RR)\right)\right]}$$

Vaccine study

ORIGINAL ARTICLE

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Fernando P. Polack, M.D., Stephen J. Thomas, M.D., Nicholas Kitchin, M.D., Judith Absalon, M.D., Alejandra Gurtman, M.D., Stephen Lockhart, D.M., John L. Perez, M.D., Gonzalo Pérez Marc, M.D., Edson D. Moreira, M.D., Cristiano Zerbini, M.D., Ruth Bailey, B.Sc., Kena A. Swanson, Ph.D., et al., for the C4591001 Clinical Trial Group*

A total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo.

Question:

- Please construct a contingency table.
- Calculate the risk of having COVID positive when subjects received vaccine injection
- Calculate the risk of having COVID positive when subjects did not received vaccine injection
- Calculate the Relative risk with 95% CI of the COVID vaccine.

- If the Vaccine efficacy (VE) is determined by (1 – RR), what is the COVID vaccine efficacy

- How do you discuss the results?

Function for calculating CI

```
> relative.risk = function(x,  conf.level=0.95)
> {
>   a = x[1,1]; b = x[1,2]; c = x[2,1]; d = x[2,2]
>   RR <- (a/(a+b)) / (c/(c+d))
>   ASE <- sqrt((b/(a*(a+b))) + (d/(c*(c+d))))
>   CI <- exp(log(RR) + c(-1,1) * qnorm(0.5*(1+conf.level)) *ASE)
>   list(estimator=RR,
>        ASE=ASE,
>        conf.interval=CI,
>        conf.level=conf.level)
> }
```

Odds Ratio (OR)

- **It is a ratio of Odds!!**

- **What is Odds?**

$$\text{Odds of an event} = \frac{\Pr(\text{event will occur})}{\Pr(\text{event will } \mathbf{not} \text{ occur})}$$

Odds Ratio (OR)

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Odds of exposed person develops a disease $= \frac{a}{b}$

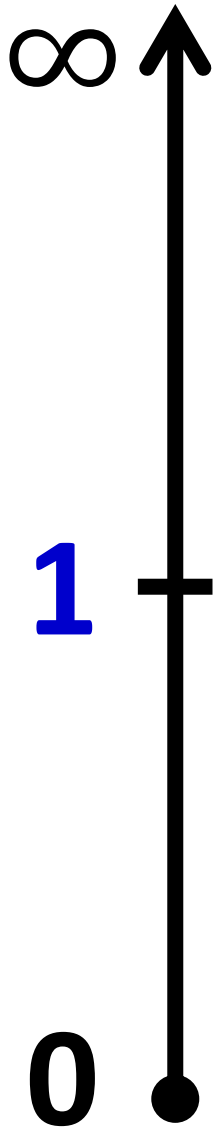
Odds of non-expose person develops a disease $= \frac{c}{d}$

Odds Ratio (OR)

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$\text{Odds Ratio} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c}$$

Relative Risk (RR)

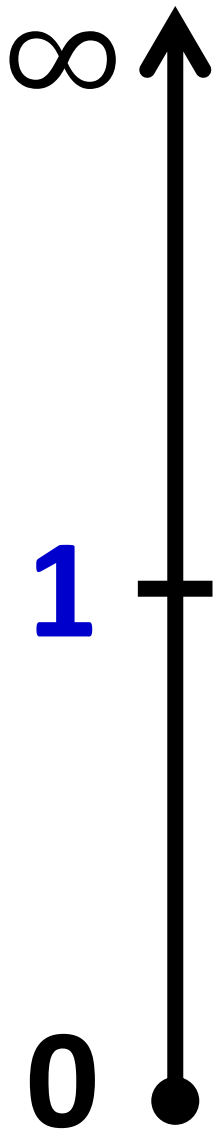


$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = 1$$

Risk between two groups are the same.

No association

Relative Risk (RR)

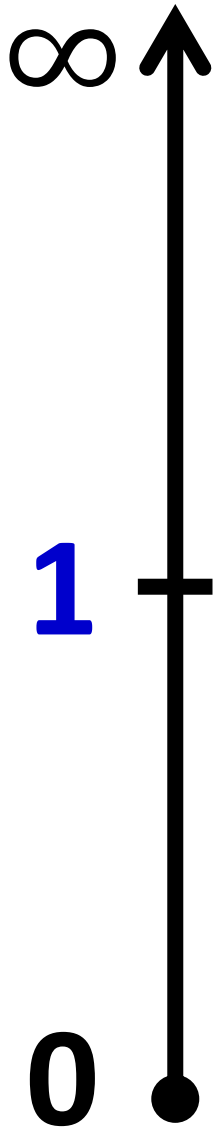


$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} > 1$$

Risk of exposed group > Risk of unexposed group

Being exposed → Higher risk

Relative Risk (RR)

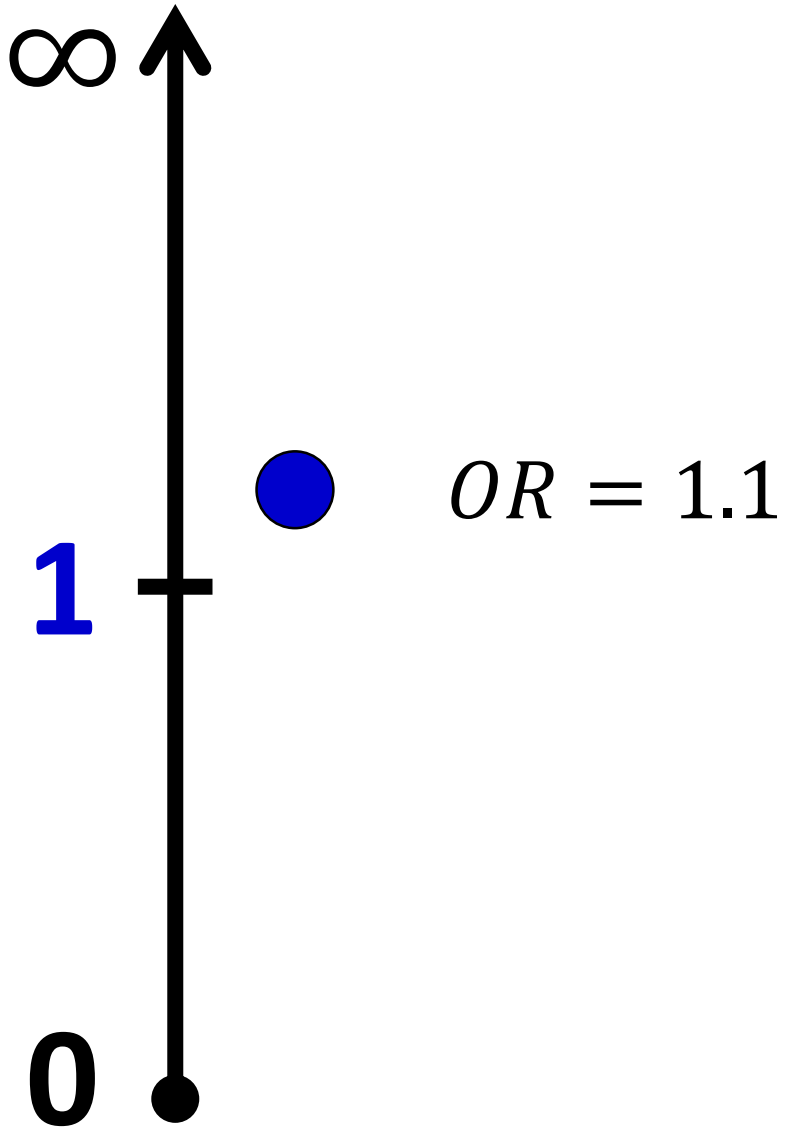


Risk of exposed group < Risk of unexposed group

Being exposed → Lower risk

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} < 1$$

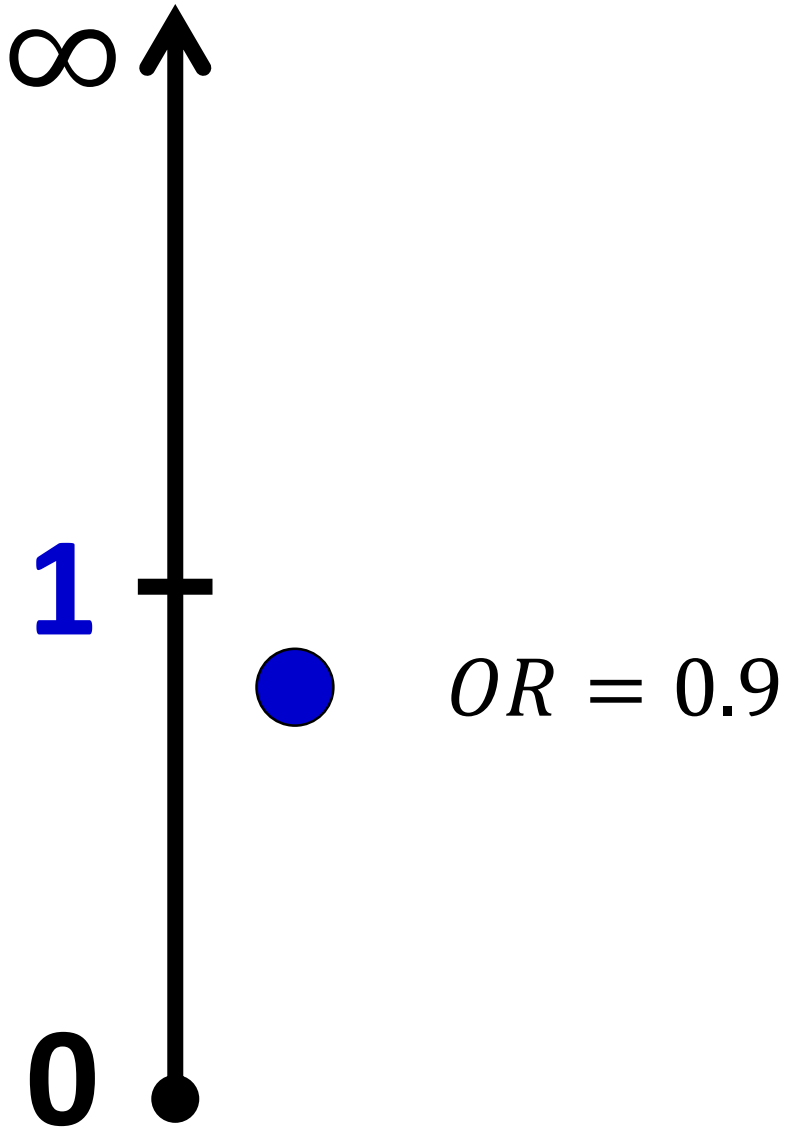
Relative Risk (RR)



$OR = 1.1$

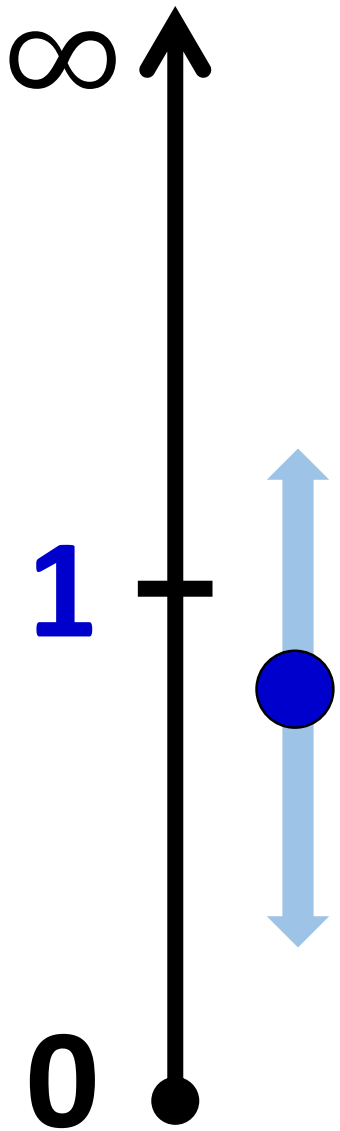
Is there any association?

Relative Risk (RR)



Is there any association?

Relative Risk (RR)

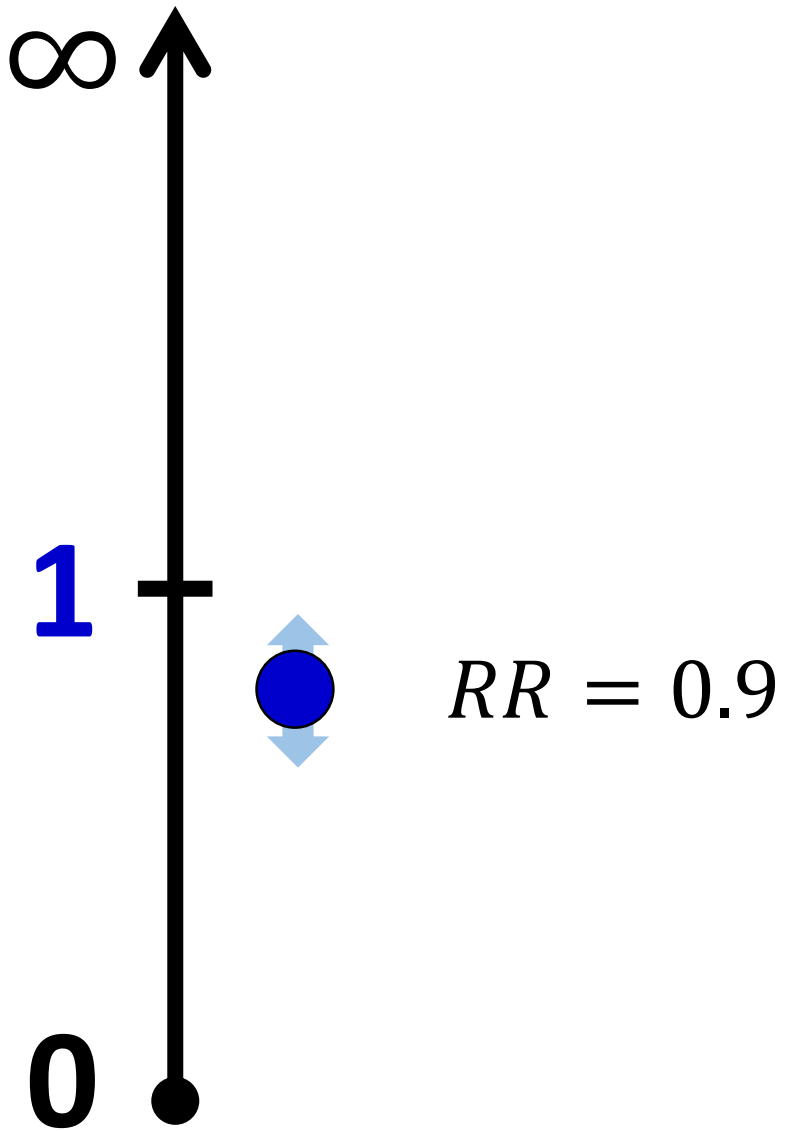


$OR = 0.9$

95% confidence interval

If 1 falls within 95% CI → No association

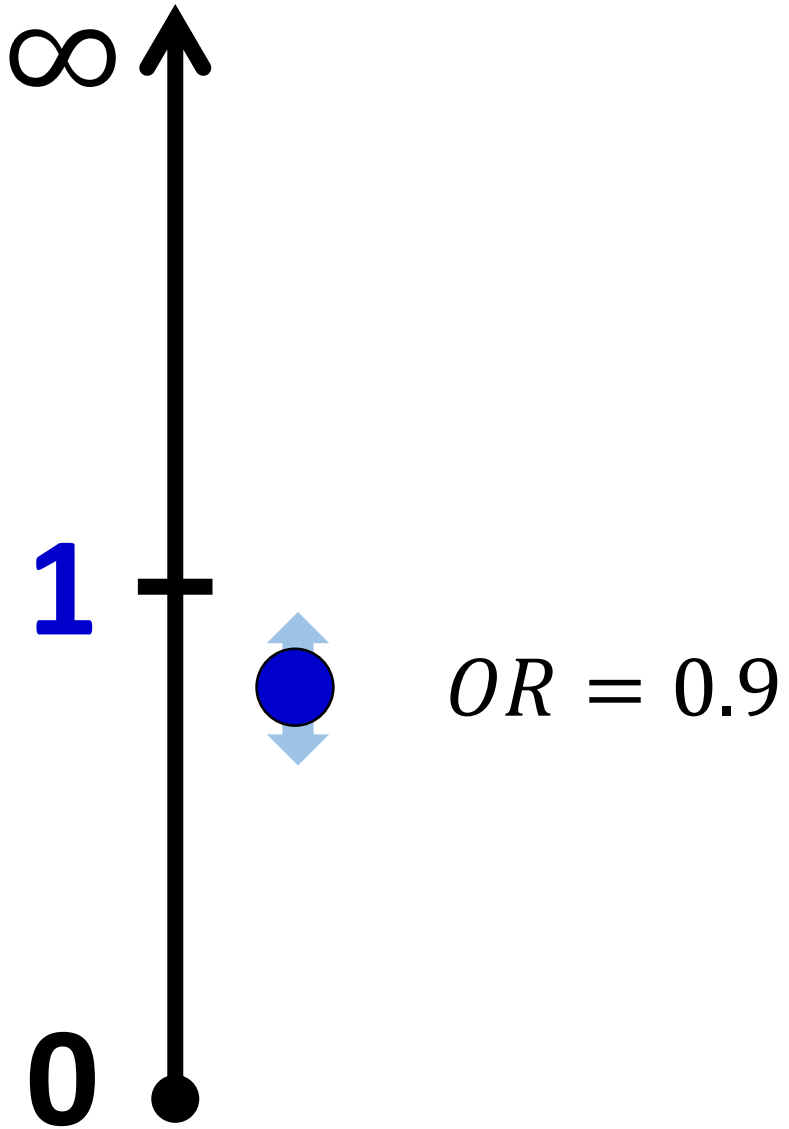
95% Confidence Interval



95% confidence interval

If not → there is an association

Relative Risk (RR)



~~$OR \pm 1.96\ SE$~~

95% CI of log(OR)

$$log\,(OR) \pm 1.96 \times SE\big(log(OR)\big)$$

95% CI of RR

$$e^{[log\,(OR) \pm 1.96 \times SE(log(OR))]}$$

COVID vaccine

BNT162b2	COVID		Total
	Positive	Negative	
Yes	8	21712	21720
No	162	21566	21728
Total	170	43278	43448

$$\text{Odds of getting COVID for a vaccinated group} = \frac{8}{21712} = 0.000368$$

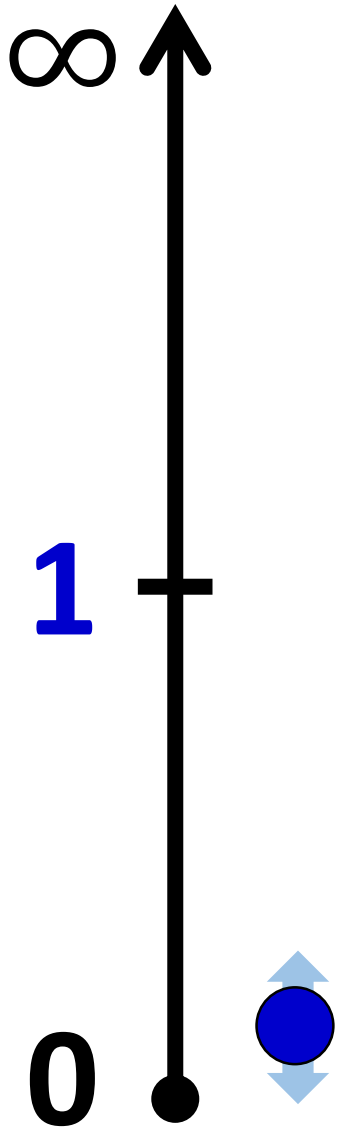
$$\text{Odds of getting COVID for a non-vaccinated group} = \frac{162}{21566} = 0.000751$$

COVID vaccine

BNT162b2	COVID		Total
	Positive	Negative	
Yes	8	21712	21720
No	162	21566	21728
Total	170	43278	43448

$$\text{Odds Ratio} = \frac{0.000368}{0.000751} = 0.049$$

COVID vaccine



$$\text{Odds Ratio} = \frac{0.000368}{0.000751} = 0.049$$

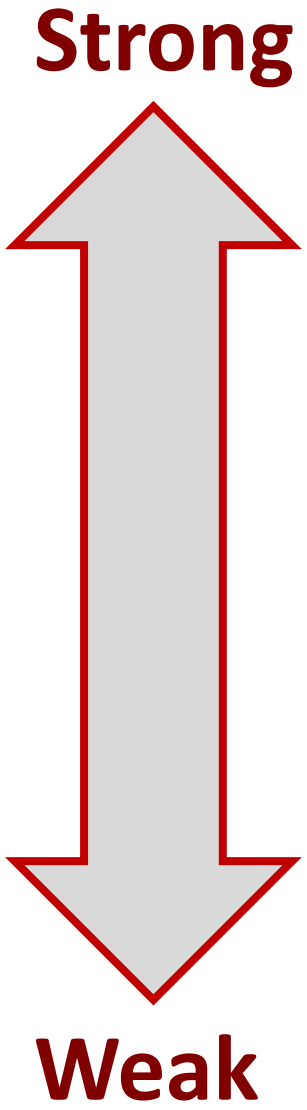
95% Confidence interval = 0.024 to 0.099

Negative association

Odds of getting COVID in a non-vaccinated group is about 20 times higher than it is for vaccinated group.

$$OR = 0.049\ (0.024 - 0.099)$$

Hypothesis testing



Barnard's test

Fisher's Exact test

Pearson's Chi square test

Hypothesis testing

Pearson's Chi square test

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Each cell >5

Hypothesis testing

Fisher's Exact test

Exposure	Outcome		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Fix marginal Freq.

Hypothesis testing

Barnard's test

No condition, but not popular

Hypothesis testing

Step 1

H_0: There is **NO** association between exposure and outcome

H_1 : There is association between exposure and outcome

Hypothesis testing

Step 2

Performed hypothesis testing

Hypothesis testing

Step 3

Check p-value!

If p-value < 0.05, reject H_0 → There is an association

If p-value > 0.05, do not reject H_0 → There is no association

Hypothesis testing in R

Barnard's test

```
Library(Exact)
exact.test(x)
```

Fisher's Exact test

```
fisher.test(x)
```

Pearson's Chi square test

```
chisq.test(x)
```

COVID vaccine

BNT162b2	COVID		Total
	Positive	Negative	
Yes	8	21712	21720
No	162	21566	21728
Total	170	43278	43448

H_0: There is **NO** association between vaccination and getting COVID

H_1: There is association between vaccination and getting COVID

OR = 0.04905065 (0.02411292, 0.09977913)

P-value < 2.2e-16

Summary

Type of study	Type of association	Hypothesis testing
Cohort study	Relative risk (RR)	Pearson's Chi square test Fisher's Exact test Barnard's test
Case-control study	Odds Ratio (OR)	Pearson's Chi square test Fisher's Exact test Barnard's test
Cross-sectional study	Relative risk (RR)	Pearson's Chi square test Fisher's Exact test Barnard's test