TM-CM02 Biostatistics for Public Health

Lecture 1

Categorical data analysis (part 2)

Kiesha Prem

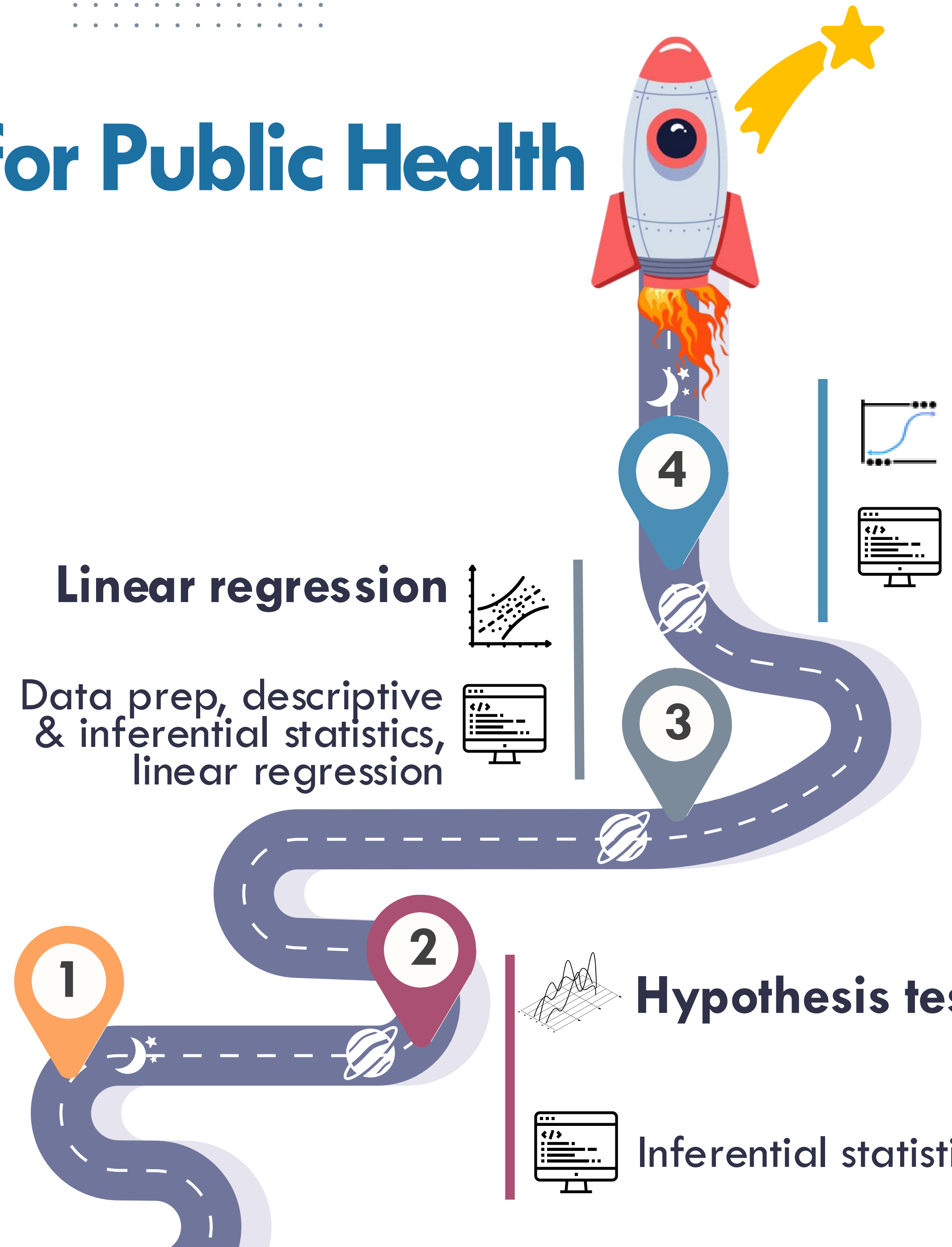
Saw Swee Hock School of Public Health, National University of Singapore

TM-CM02

Biostatistics for Public Health

quizzes

2 assignments



1. Categorical data analysis

Data prep, descriptive & inferential statistics

2. Hypothesis testing

Inferential statistics

3. Linear regression

Data prep, descriptive & inferential statistics, linear regression

4. Logistic regression

Data prep, descriptive & inferential statistics, logistics regression

NUS National University of Singapore | Saw Swee Hock School of Public Health

Categorical data analysis

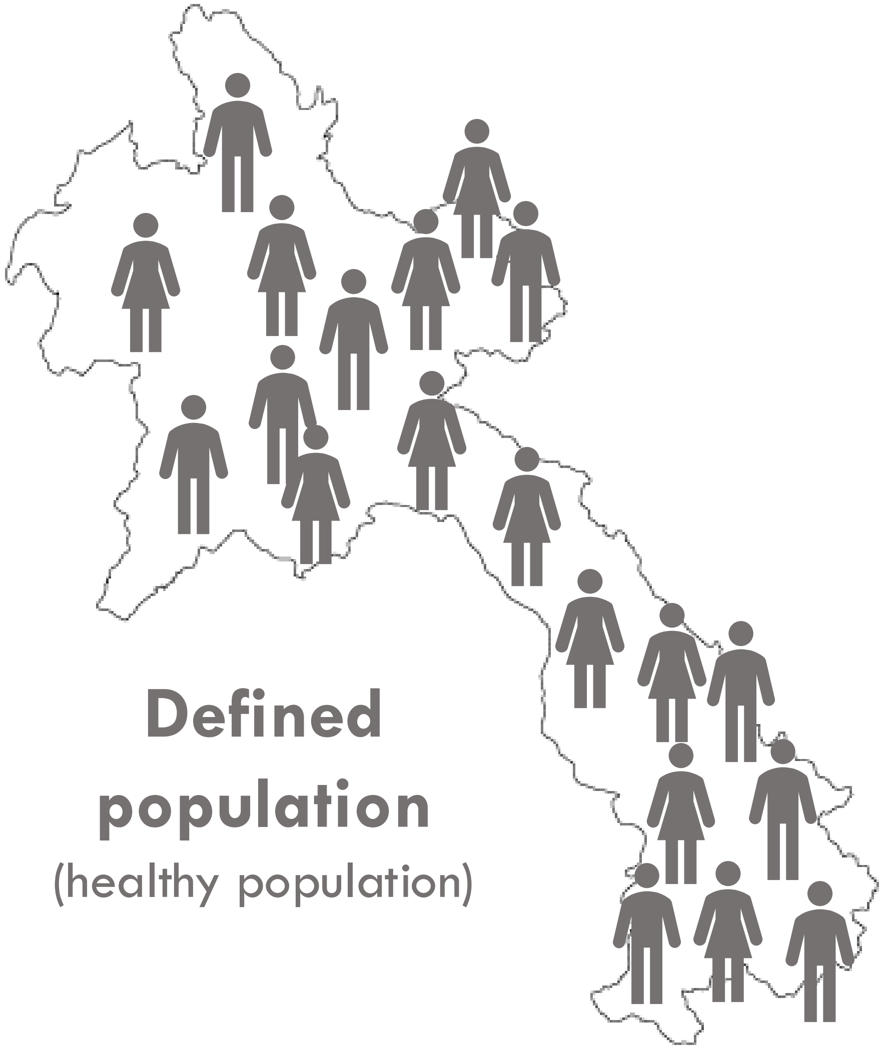
By the end of the session, you will:

- Understand the basics of categorical data and its relevance in public health.
- Perform descriptive analyses and visualise categorical data effectively.
- Apply statistical tests to assess relationships between categorical variables.
- Compute and interpret measures of association and confidence intervals.
- (Next week) Build and interpret logistic regression models in public health contexts.

First… let's review some basics

Risk, rate, and Odds

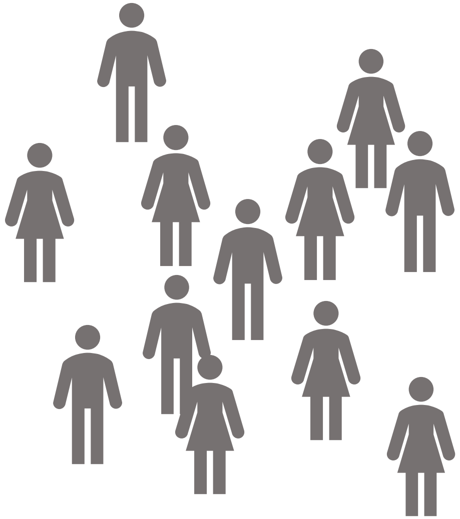
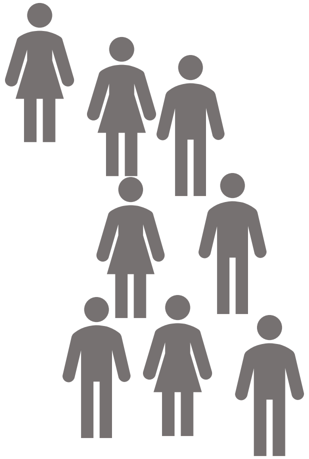
To calculate **Risk**



Defined population

(healthy population)

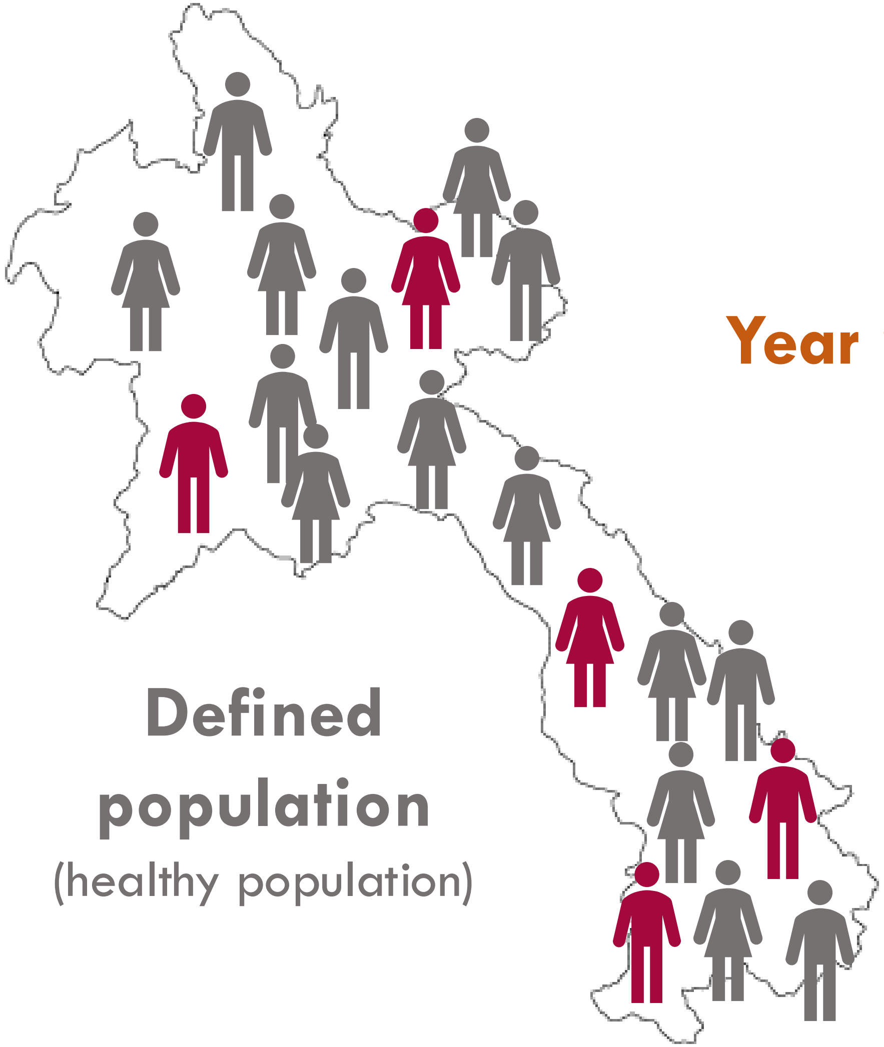
To calculate **Risk**

Year 2025

Defined population
(healthy population)

To calculate **Risk**

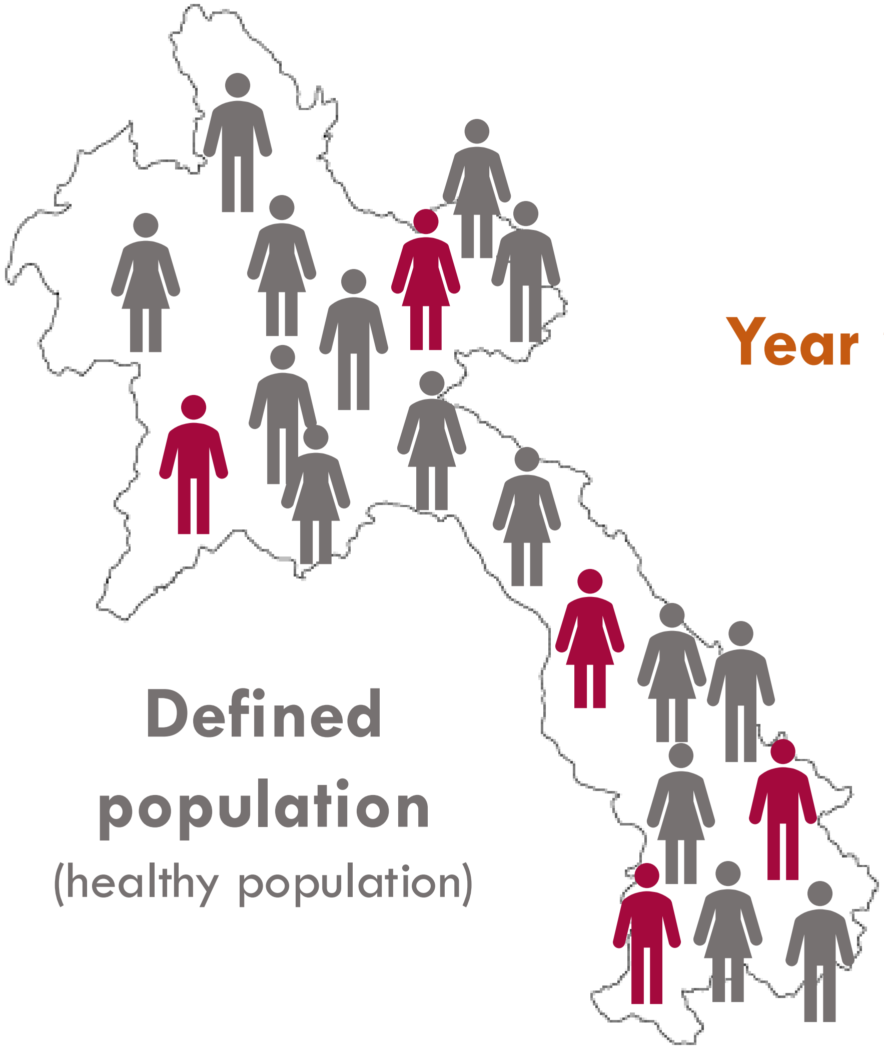


Defined population
(healthy population)

Year 2025

New cases of individuals living with Type 2 Diabetes

To calculate Risk



Defined population

(healthy population)

Year 2025

New cases of individuals living with Type 2 Diabetes

$$\text{Risk} = {}^{5}/_{20}$$

(Similar to the incidence)

To calculate **Rate**

To calculate **Rate**

Defined population
(healthy population)

Year 2025

New cases of individuals living with Type 2 Diabetes

To calculate Rate

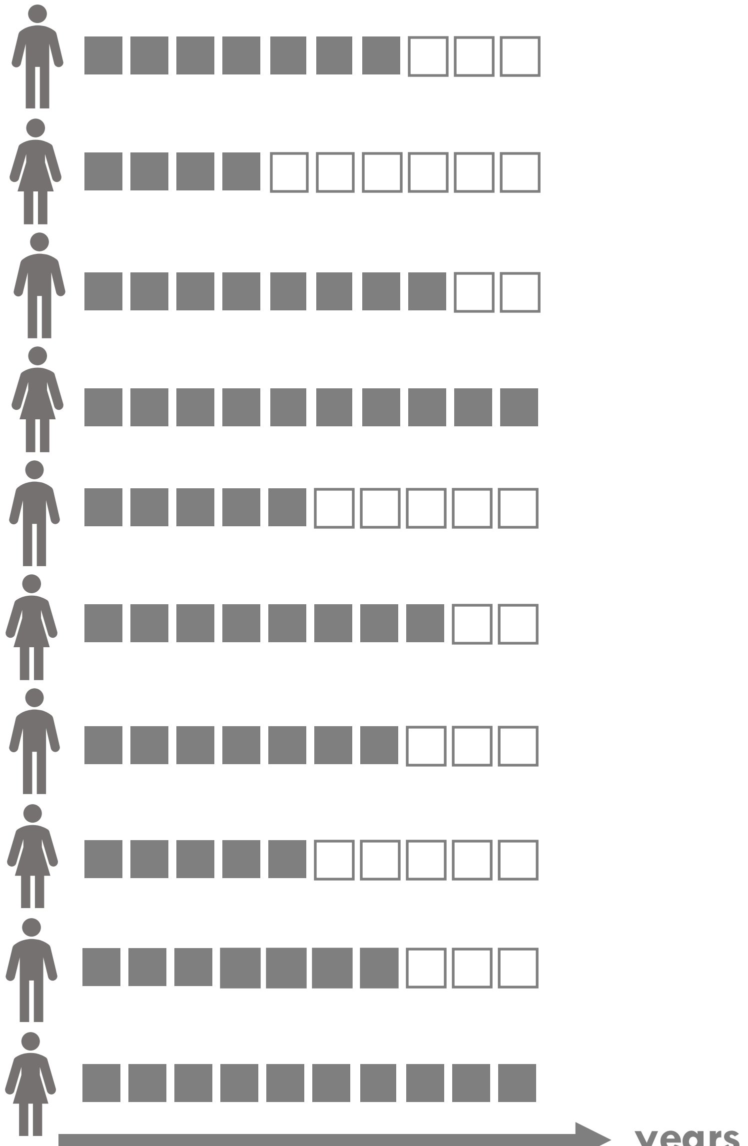
Sample

Defined population
(healthy population)

Year 2025

New cases of individuals living with Type 2 Diabetes

To calculate Rate



Defined population
(healthy population)

Year 2025

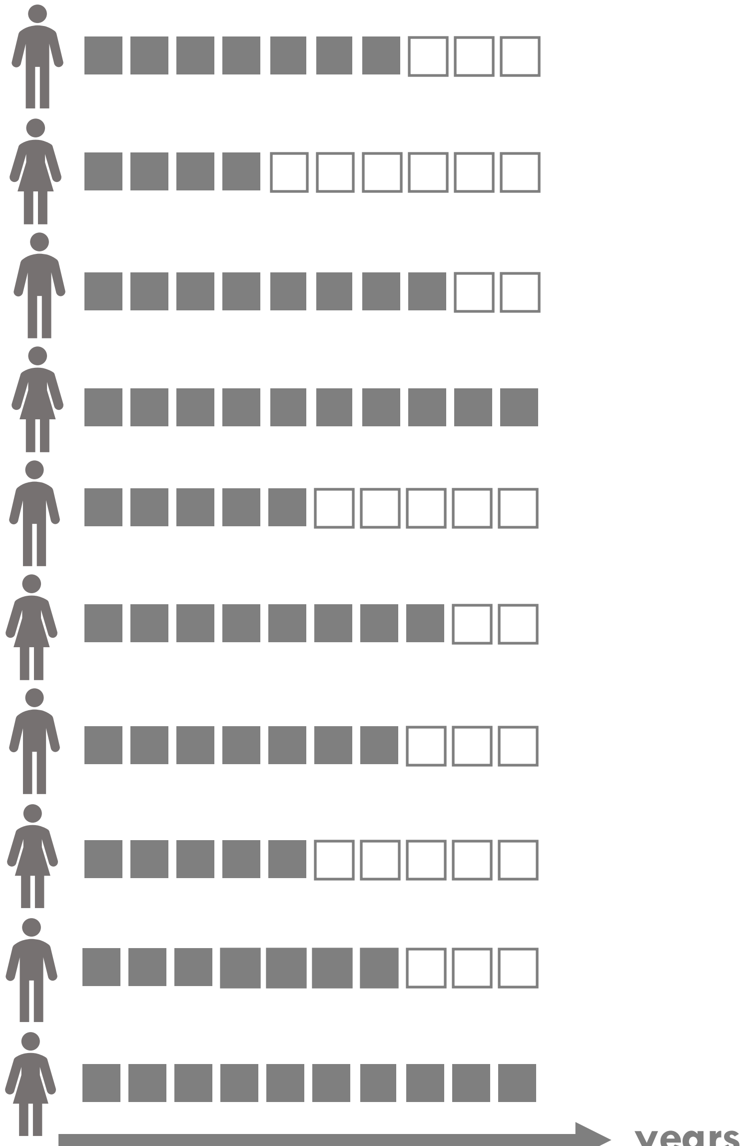
New cases of individuals living with Type 2 Diabetes

To calculate Rate

Defined population
(healthy population)

Year 2025

New cases of individuals living with Type 2 Diabetes



Not everyone in the study will be alive or around the whole duration
(hence, we need to account for the time contribution of each person)

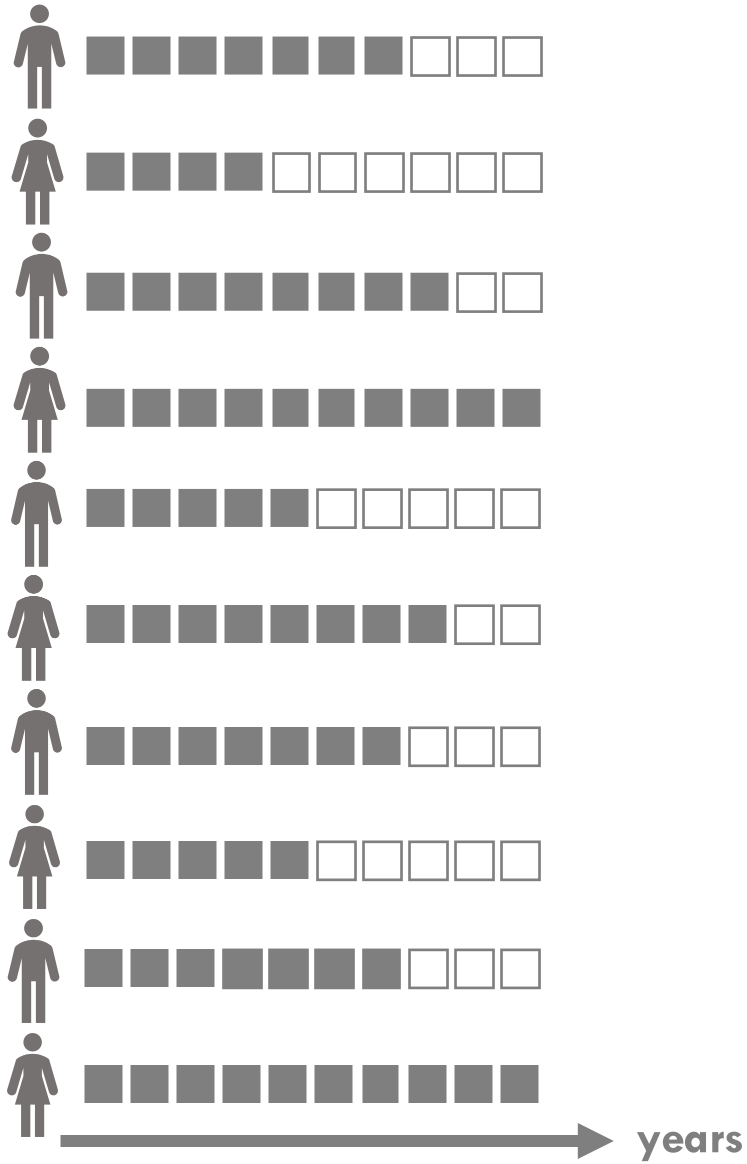
To calculate **Rate**

NUS National University of Singapore
Saw Swee Hock School of Public Health

Defined population
(healthy population)

Year 2025

New cases of individuals living with Type 2 Diabetes



Total person–years

= 71

Not everyone in the study will be alive or around the whole duration

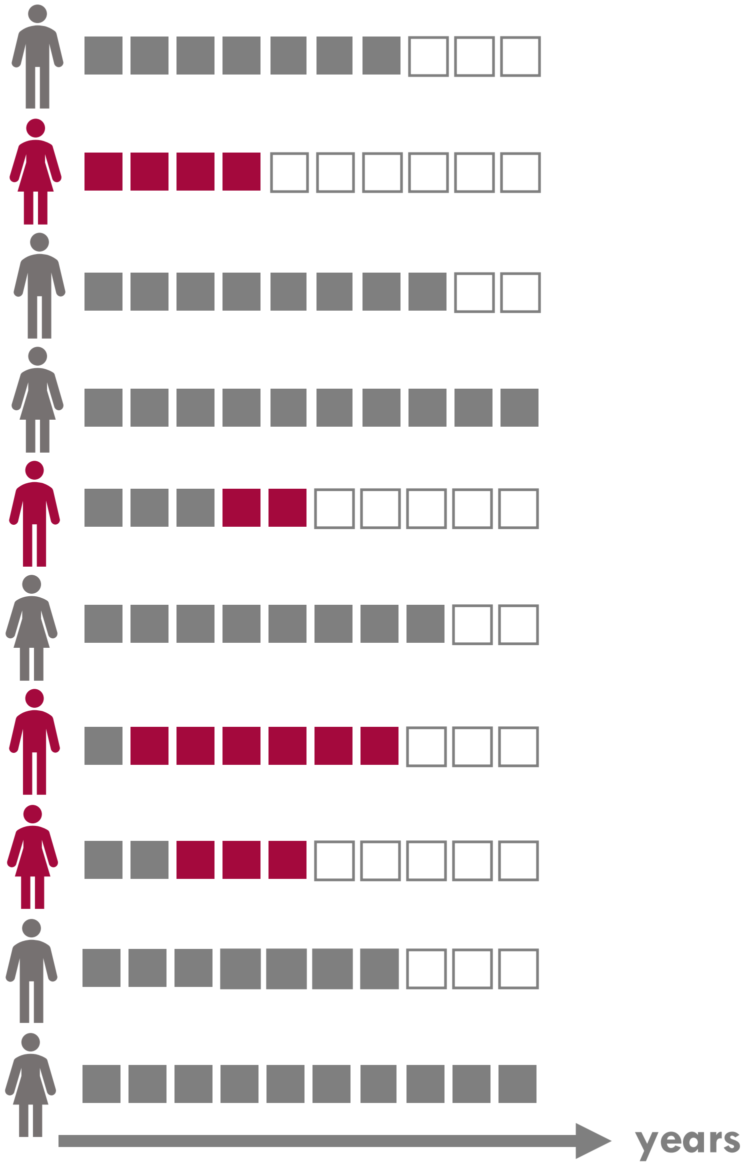
(hence, we need to account for the time contribution of each person)

To calculate Rate

Defined population
(healthy population)

Year 2025

New cases of individuals living with Type 2 Diabetes



Total person-years
$= 71$

Rate
$= \text{cases} / \text{person-years}$
$= 4 / 71 \text{ person-years}$

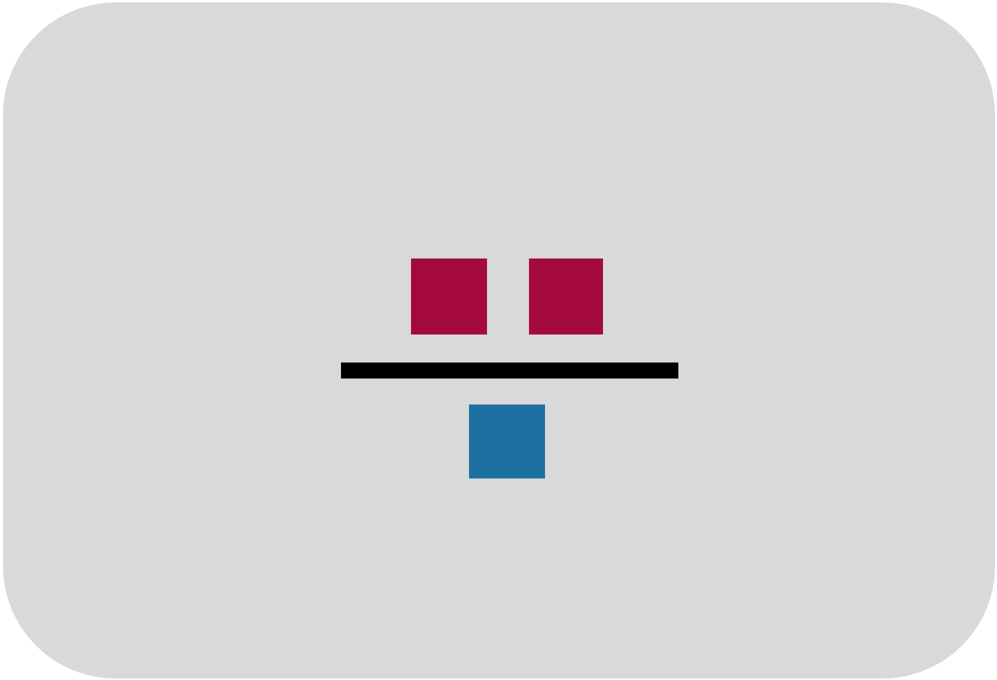
To calculate **Odds**

To calculate **Odds**

$$\frac{\text{Number of events}}{\text{Number of non events}} \quad \text{or} \quad \frac{\text{Pr(something happening)}}{\text{Pr(something not happening)}}$$

To calculate **Odds**

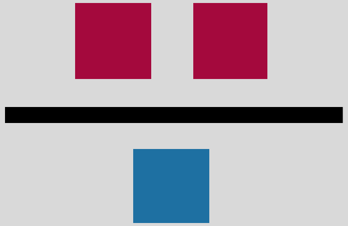
$$\frac{\text{Number of events}}{\text{Number of non events}} \quad \text{or} \quad \frac{\text{Pr(something happening)}}{\text{Pr(something not happening)}}$$



To calculate Odds

$$\frac{\text{Number of events}}{\text{Number of non events}} \quad \text{or} \quad \frac{\text{Pr(something happening)}}{\text{Pr(something not happening)}}$$

If the Odds of an event is >1



Odds of an event $= \frac{2}{1} = 2$

The event is more likely to happen than not

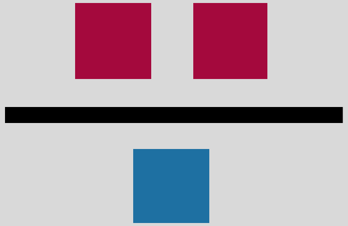
To calculate **Odds**

$$\frac{\text{Number of events}}{\text{Number of non events}}$$

or

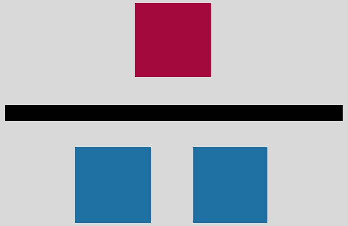
$$\frac{\text{Pr(something happening)}}{\text{Pr(something not happening)}}$$

If the **Odds** of an event is **>1**



Odds of an event $= \frac{2}{1} = 2$

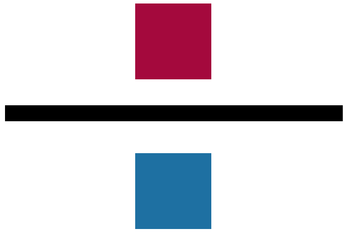
If the **Odds** of an event is **<1**



Odds of an event $= \frac{1}{2} = 0.5$

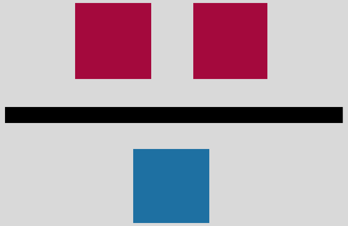
To calculate Odds

If the Odds of an event is =1



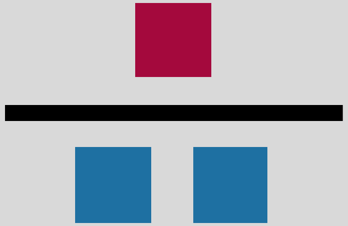
no difference

If the Odds of an event is >1



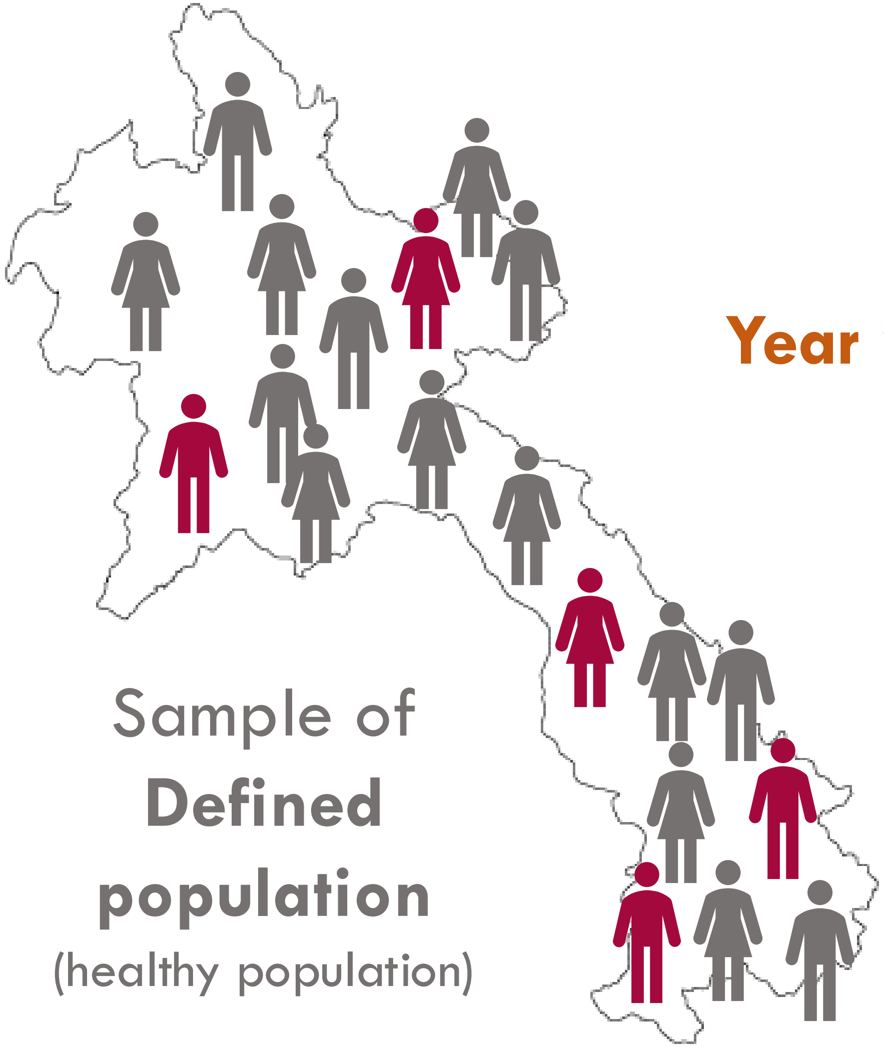
Odds of an event $= \frac{2}{1} = 2$

If the Odds of an event is <1



Odds of an event $= \frac{1}{2} = 0.5$

To calculate **Risk** and **Odds**



Sample of **Defined population** (healthy population)

Year 2025

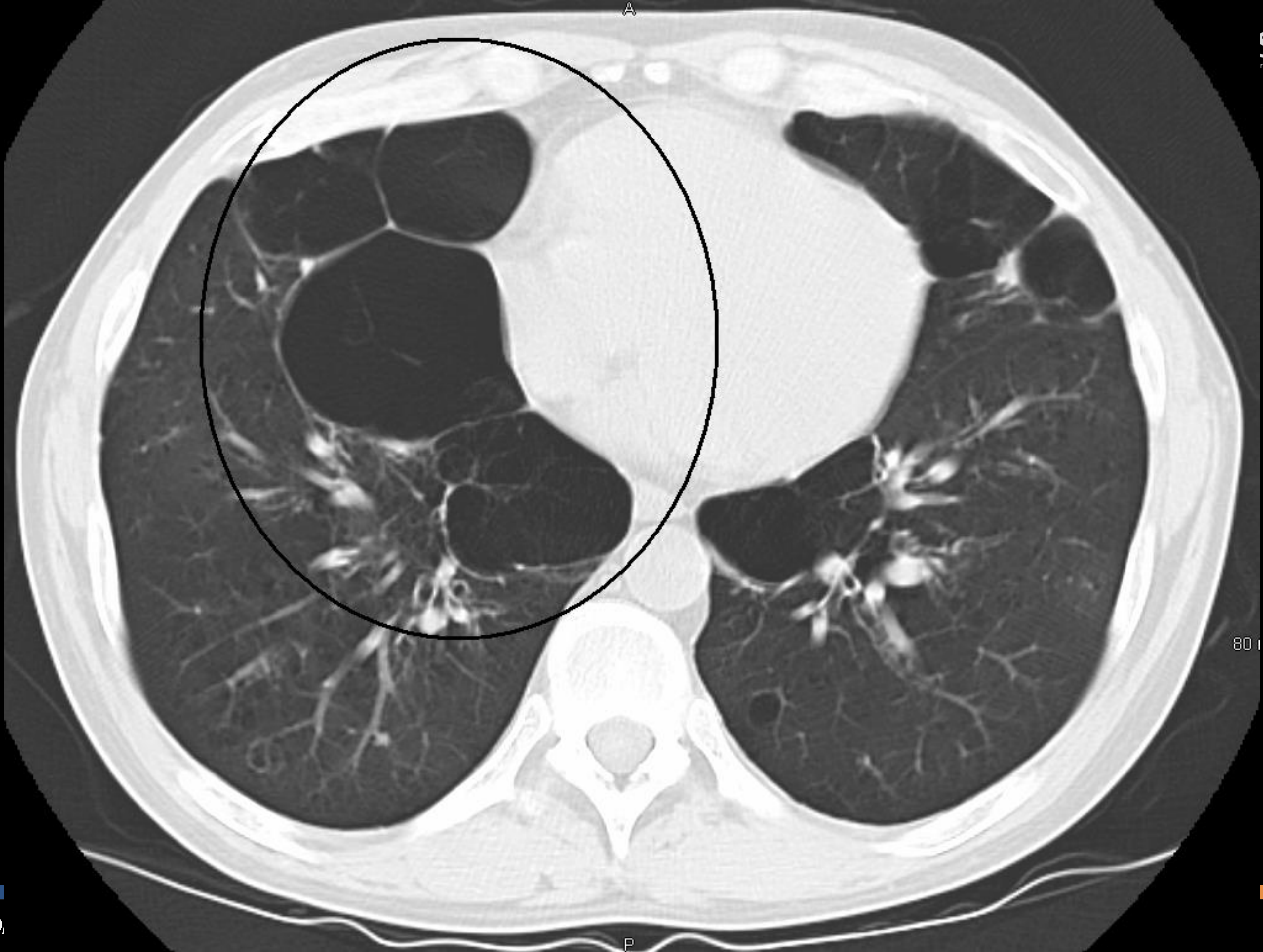
New cases of individuals living with Type 2 Diabetes

$$\textbf{Risk} = {}^{5}/_{20} = 0.25$$

$$\textbf{Odds} = {}^{5}/_{15} = 0.33$$

Is there any association between **lung cancer** and **smoking?**

Doll, R., & Hill, A. B. (1950). Smoking and Carcinoma of the Lung. *British Medical Journal*, *2*(4682), 739–748.


A
80
P

2×2 Contingency Tables

Use a contingency table to study the relationship between two categorical variables

- **Cross-tabulation**

Cells can display counts, percentages or proportions.

- For an $I \times J$ contingency table
  - X has *I* categories with ***I* rows** for each category of **X**
  - Y has J categories with ***J* columns** for each category of **Y**
  - *IJ* possible combinations of outcomes

Basic Contingency Table Example

Favorite Flavor	Boys		Girls	
Vanilla	8	32%	9	26%
Chocolate	10	40%	6	17%
Strawberry	5	20%	14	40%
Mint Chip	2	8%	6	17%
Total	25	100%	35	100%

Smoking and Carcinoma of the Lung

	Lung Cancer Status	
Smoking Status	Case	Control
Smoker	688	650
Non-smoker	21	59
Total	709	709

Sample cell proportions

The sample cell proportions relate to the cell counts by $p_{ij} = \frac{n_{ij}}{n}$.

	Lung Cancer Status	
Smoking Status	Case	Control
Smoker	0.485	0.458
Non-smoker	0.015	0.042

Measure of association

To assess the **strength of an association** between an exposure and the outcome of interest.

It indicates how more or less likely a group is to develop a particular disease compared to another group.

The two widely used measures:

1. Risk ratio (RR)
2. Odds ratio (OR)

Risk Ratio

Ratio of the probability of an event occurring in an exposed group to the probability of the event occurring in the non-exposed group.

$$\mathrm{RR} = \frac{\Pr(\text{event when exposed})}{\Pr(\text{event when not exposed})} = \frac{\Pr(Disease|\ Exposed)}{\Pr(Disease|\ Not\ Exposed)}$$

In epidemiology, risk ratio can be seen as the ratio of the risk of disease in the exposed group to the risk in the non-exposed group.

Risk Ratio

	Disease		
Exposure	Yes	No	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	N = a+b+c+d

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

Risk Ratio

Exposure	Disease: Yes	Disease: No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	N = a+b+c+d

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

Interpretation of Risk Ratio

If RR = 1

Risk in exposed group = Risk in non-exposed group

This indicates **no association**.

If RR > 1

Risk in exposed group > Risk in non-exposed group

This indicates a **positive association**.

If RR < 1

Risk in exposed group < Risk in non-exposed group

This indicates a **negative association**.

Odds

The odds of an event is the ratio of the probability that the event will occur to the probability that the event will not occur.

$$\text{odds of an event} = \frac{\Pr(\text{event will occur})}{\Pr(\text{event will not occur})}$$
$$= \frac{\mathrm{n(event)}}{\mathrm{n(non - event)}}$$

Non-negative values

Odds greater than 1 indicates a success is more likely than a failure

Odds

	Disease		
Exposure	Yes	No	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	N = a+b+c+d

- The odds that an exposed person develops a disease is $\frac{a}{b}$.
- The odds that a non-exposed person develops the disease is $\frac{c}{d}$.

Odds Ratio

Odds ratio (OR) is the ratio of the odds of disease in the exposed group to odds of disease in the non-exposed group.

$$OR = \frac{\text{odds that an exposed person develops a disease}}{\text{odds that a non} - \text{exposed person develops a disease}}$$

$$= \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{\boldsymbol{a \times d}}{\boldsymbol{b \times c}}$$

Odds Ratio

	Disease		
Exposure	**Yes**	**No**	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	N = a+b+c+d

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

Odds Ratio

	Disease		
Exposure	Yes	No	**Total**
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	N = a+b+c+d

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

Interpretation of Odds Ratio

If **OR = 1**

The odds of having the outcome are equal for those exposed and those who are not exposed.

This indicates **no association**.

If **OR > 1**

The odds of having the outcome are higher for those exposed and those who are not exposed.

This indicates a **positive association**.

If **OR < 1**

The odds of having the outcome are lower for those exposed and those who are not exposed.

This indicates a **negative association**.

Is there any association between **lung cancer** and **smoking?**

Smoking and Carcinoma of the Lung

	Lung Cancer Status	
Smoking Status	Case	Control
Smoker	688	650
Non-smoker	21	59
Total	709	709

Significant tests

Significance tests assess the **evidence against the null hypothesis** by the calculation of a test statistic and obtaining a corresponding **p-value**.

A relatively **low p-value** provides evidence against the null hypothesis, whereas a relatively **high p-value** suggests there is little or no evidence against the null hypothesis.

The **null hypothesis** $(\boldsymbol{H_0})$ usually states that there is no difference between two means or proportions or that a ratio measure is equal to one. Alternatively, we may test that a mean or proportion is equal to a non-zero value.

Significant tests

Steps for conducting significance tests:

1. **State the null hypothesis** (H_0)
2. **State the alternate hypothesis** (H_α)
3. **Calculate test statistic** (parameter of interest divided by standard error)
4. **Look up and interpret p-value:**
   - Remember that statistical significance is not equivalent to medical or biological significance!
   - Interpret a p-value in terms of the level of evidence (α) against the null hypothesis.

Chi-square tests of independence

To identify whether there is a significant association between the two categorical variables.

H_0: The two categorical variables are independent

H_1: The two categorical variables are associated.

The chi-square χ^2 test statistics is

$$\chi^2 = \sum_{i=1}^{K} \frac{(|O_i - E_i|)^2}{E_i}.$$

If H_0 is true, χ^2 test statistics follows a χ_1^2 distribution.

Chi-square tests of independence

- It merely indicates if there is significant association between the two categorical variables.
- They are **not** able to quantify the strength and direction of the association.
- You will need the risk ratios or odds ratios to describe the strength of association.

Chi-square tests of independence

Assumptions: at least 80% of the cells have an expected count of 5 or more.

Another way to look at categorical data is to use a **logistic regression.**

NUS National University of Singapore | Saw Swee Hock School of Public Health

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments

1 Categorical data analysis

Data prep, descriptive & inferential statistics

2 Hypothesis testing

Inferential statistics

3 Linear regression

Data prep, descriptive & inferential statistics, linear regression

4 Logistic regression

Data prep, descriptive & inferential statistics, logistics regression

Thank you