TM-CM02 Biostatistics for Public Health

Lecture 1

Categorical data analysis

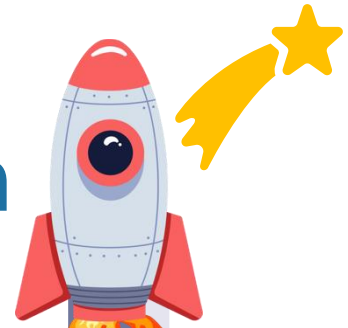
Kiesha Prem

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

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments



4

Logistic regression

Data prep, descriptive & inferential statistics, logistics regression

3

Linear regression

Data prep, descriptive & inferential statistics, linear regression

2

Hypothesis testing

Inferential statistics

1

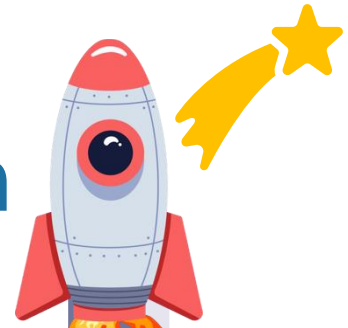
Categorical data analysis

Data prep, descriptive & inferential statistics

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments



4

Logistic regression

Data prep, descriptive & inferential statistics, logistics regression

3

Linear regression

Data prep, descriptive & inferential statistics, linear regression

2

Hypothesis testing

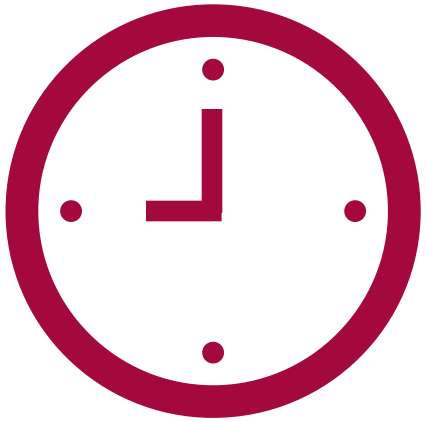
Inferential statistics

1

Categorical data analysis

Data prep, descriptive & inferential statistics

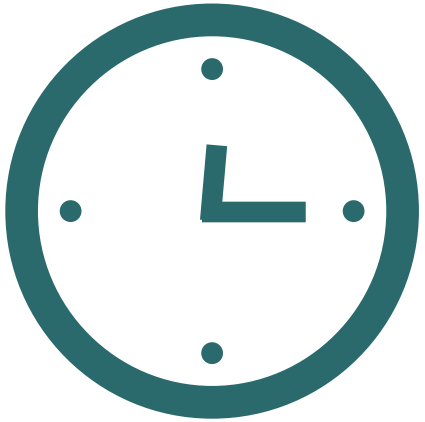
Today



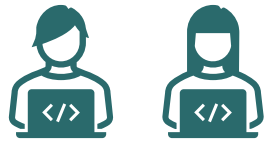
The class will start with lectures at 9 am
(There is a quiz next week!)

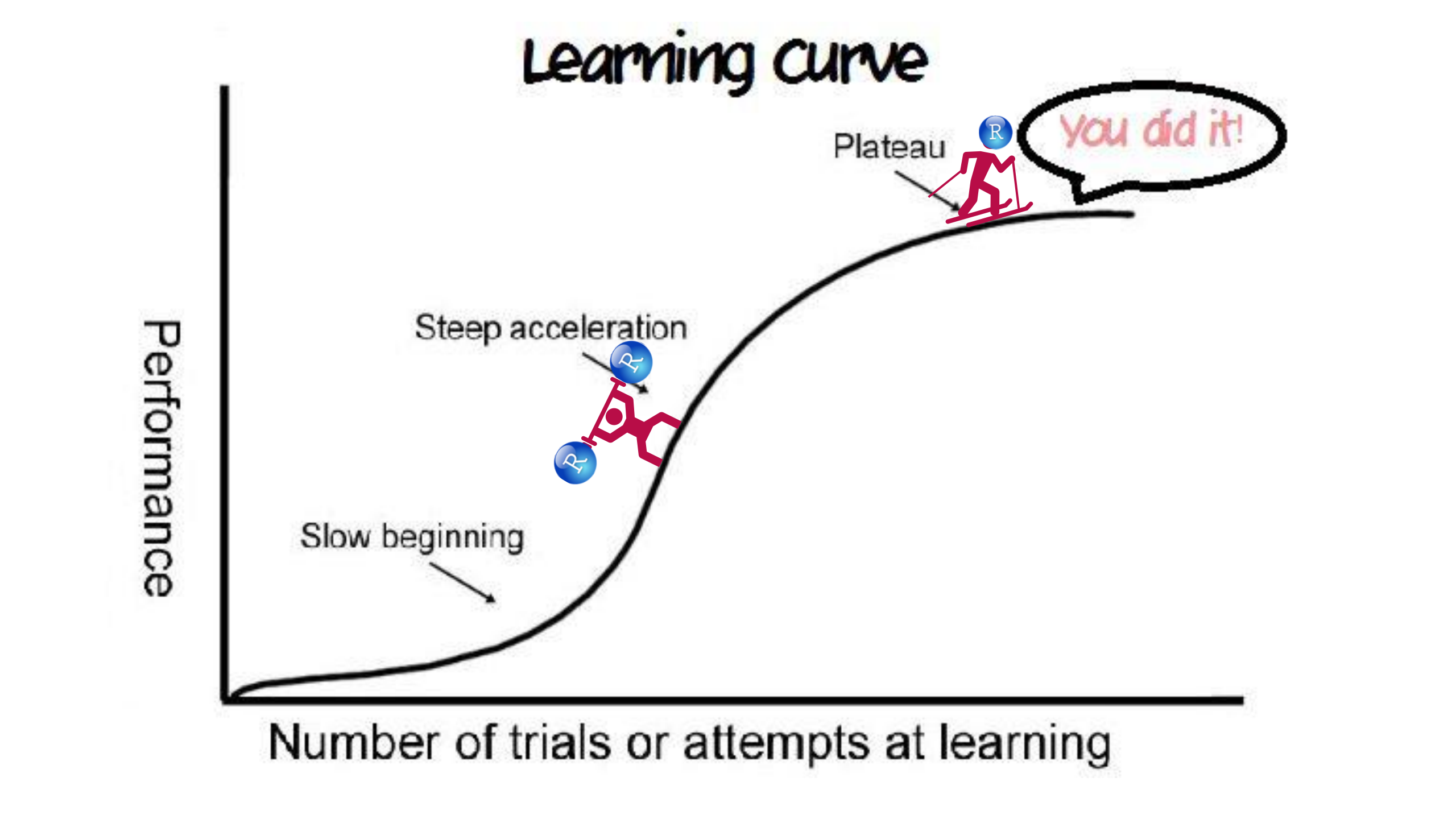
Lectures will end at around ~10 am

15-min Snack/bio break



Practicals will end at around ~1130 am




Learning curve
Plateau
you did it!
Steep acceleration
Slow beginning
Performance
Number of trials or attempts at learning

Categorical data analysis (part 1 and 2)

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.

What is categorical data?

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

What is categorical data?

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

In social sciences,

- to measure opinions and attitudes.


HAPPY
SAD
ANGRY
SCARED
WORRIED
SLEEPY


Please Rate
Please rate my service
Excellent
Good
Average
Poor
Very Poor
Our Service!

What is categorical data?

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

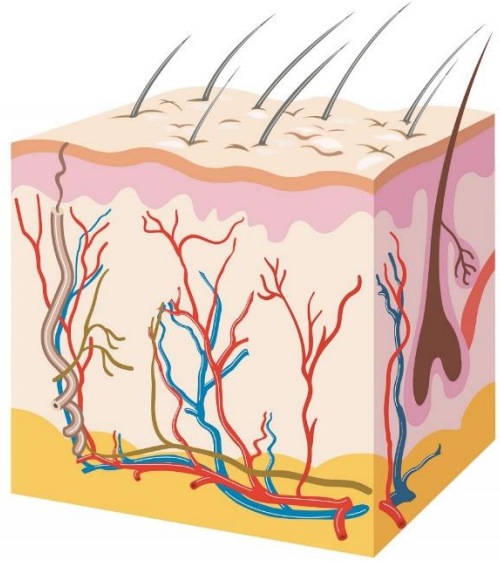
In social sciences,

- to measure opinions and attitudes.

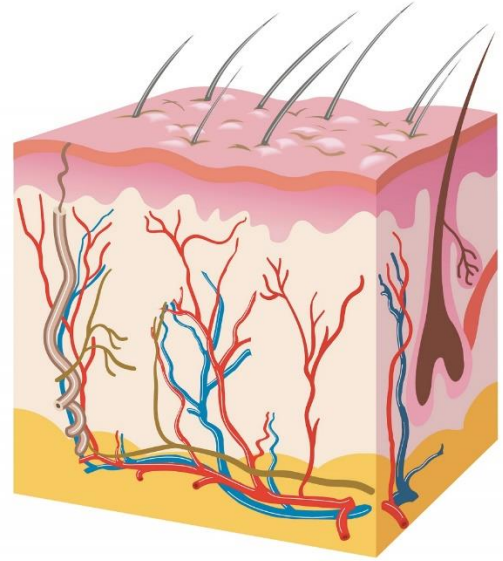
In health sciences,

- to classify the severity of a burn (first degree, second degree,…).
- to categorise the smoking status of individuals in the study (non-smoker, former smoker, current smoker).

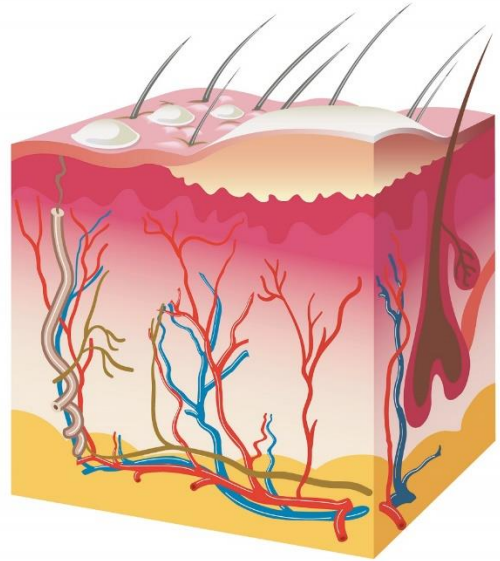
Skin Burns



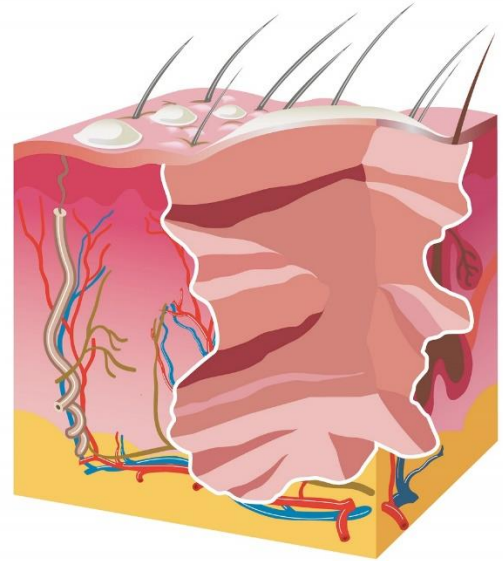
NORMAL SKIN



FIRST DEGREE BURN



SECOND DEGREE BURN



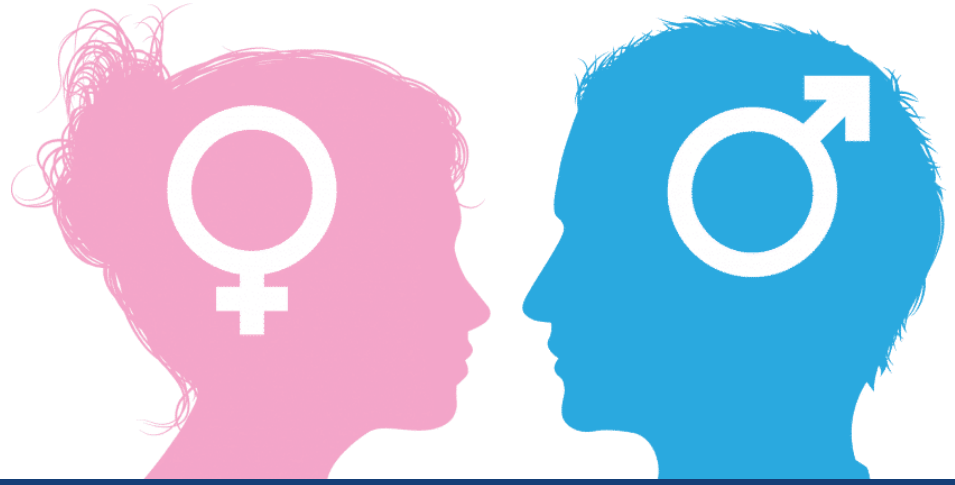
THIRD DEGREE BURN

Nominal Variables

Categories with no intrinsic order

- Sex: Male or Female
- Favourite type of music: Rock, Classical, Jazz,…

The statistical analyses on nominal variables should not depend on any ordering.



Ordinal Variables

Categorical variables with ordered scales.

- Response to a medical treatment: excellent, good, fair and poor
- An excellent response is clearly better than a fair one.

The statistical analyses designed for nominal variables can be used with nominal or ordinal variables.

The statistical analyses designed for ordinal variables cannot be used with nominal variables.

Nominal vs Ordinal

Which one of these are **categorical variables**?

	A	B	C	D	E	F	G	H
1	id	age	gender	bmi	ethnicity	smoke	cvd	ldl
2	1	72	Female	23.9	Indians	Never-Smoker	0	3.49
3	2	73	Female	26.2	Chinese	Never-Smoker	0	3.55
4	3	67	Female	19.9	Malays	Never-Smoker	0	3.15
5	4	65	Female	27.8	Indians	Never-Smoker	0	2.97
6	5	72	Male	22.0	Indians	Daily smoker	0	3.90
7	6	55	Female	20.9	Indians	Never-Smoker	0	2.29
8	7	72	Female	21.8	Malays	Daily smoker	1	3.92
9	8	66	Female	28.3	Malays	Never-Smoker	0	3.06
10	9	66	Male	27.5	Malays	Never-Smoker	0	3.06
11	10	62	Female	21.9	Chinese	Occasional smoker	0	3.14
12	11	67	Male	20.9	Malays	Never-Smoker	0	3.14
13	12	81	Female	11.6	Indians	Occasional smoker	0	4.51
14	13	71	Female	34.2	Malays	Never-Smoker	0	3.39
15	14	72	Male	22.5	Indians	Never-Smoker	0	3.46
16	15	63	Male	23.8	Malays	Never-Smoker	0	2.82

Which one of these are **categorical variables**?

	A	B	C	D	E	F	G	H
1	id	age	gender	bmi	ethnicity	smoke	cvd	ldl
2	1	72	Female	23.9	Indians	Never-Smoker	0	3.49
3	2	73	Female	26.2	Chinese	Never-Smoker	0	3.55
4	3	67	Female	19.9	Malays	Never-Smoker	0	3.15
5	4	65	Female	27.8	Indians	Never-Smoker	0	2.97
6	5	72	Male	22.0	Indians	Daily smoker	0	3.90
7	6	55	Female	20.9	Indians	Never-Smoker	0	2.29
8	7	72	Female	21.8	Malays	Daily smoker	1	3.92
9	8	66	Female	28.3	Malays	Never-Smoker	0	3.06
10	9	66	Male	27.5	Malays	Never-Smoker	0	3.06
11	10	62	Female	21.9	Chinese	Occasional smoker	0	3.14
12	11	67	Male	20.9	Malays	Never-Smoker	0	3.14
13	12	81	Female	11.6	Indians	Occasional smoker	0	4.51
14	13	71	Female	34.2	Malays	Never-Smoker	0	3.39
15	14	72	Male	22.5	Indians	Never-Smoker	0	3.46
16	15	63	Male	23.8	Malays	Never-Smoker	0	2.82

Which one of these are **categorical variables**?

Ordinal variable

	A	B	C	D	E	F	G	H
1	id	age	gender	bmi	ethnicity	smoke	cvd	ldl
2	1	72	Female	23.9	Indians	Never-Smoker	0	3.49
3	2	73	Female	26.2	Chinese	Never-Smoker	0	3.55
4	3	67	Female	19.9	Malays	Never-Smoker	0	3.15
5	4	65	Female	27.8	Indians	Never-Smoker	0	2.97
6	5	72	Male	22.0	Indians	Daily smoker	0	3.90
7	6	55	Female	20.9	Indians	Never-Smoker	0	2.29
8	7	72	Female	21.8	Malays	Daily smoker	1	3.92
9	8	66	Female	28.3	Malays	Never-Smoker	0	3.06
10	9	66	Male	27.5	Malays	Never-Smoker	0	3.06
11	10	62	Female	21.9	Chinese	Occasional smoker	0	3.14
12	11	67	Male	20.9	Malays	Never-Smoker	0	3.14
13	12	81	Female	11.6	Indians	Occasional smoker	0	4.51
14	13	71	Female	34.2	Malays	Never-Smoker	0	3.39
15	14	72	Male	22.5	Indians	Never-Smoker	0	3.46
16	15	63	Male	23.8	Malays	Never-Smoker	0	2.82

Nominal variable

How can we represent categorical data?

Using tables

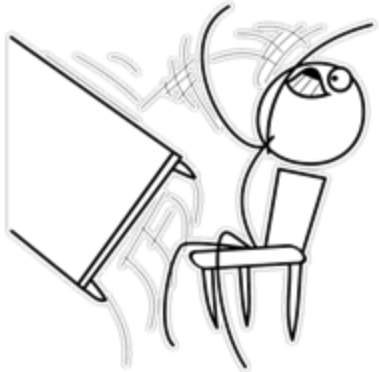


Table: Alcohol consumption of 2000 Singapore residents aged 18 to 69 years

Alcohol Consumption	Non-drinker	Occasional Drinker	Frequent Drinker	Regular Drinker
n	1078	718	152	52

Data for a single categorical variable can summarised by counting the number of observations in each category.

Using tables

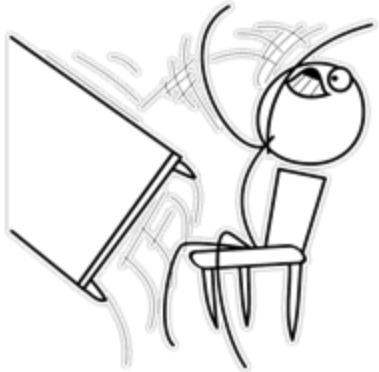


Table: Alcohol consumption of 2000 Singapore residents aged 18 to 69 years

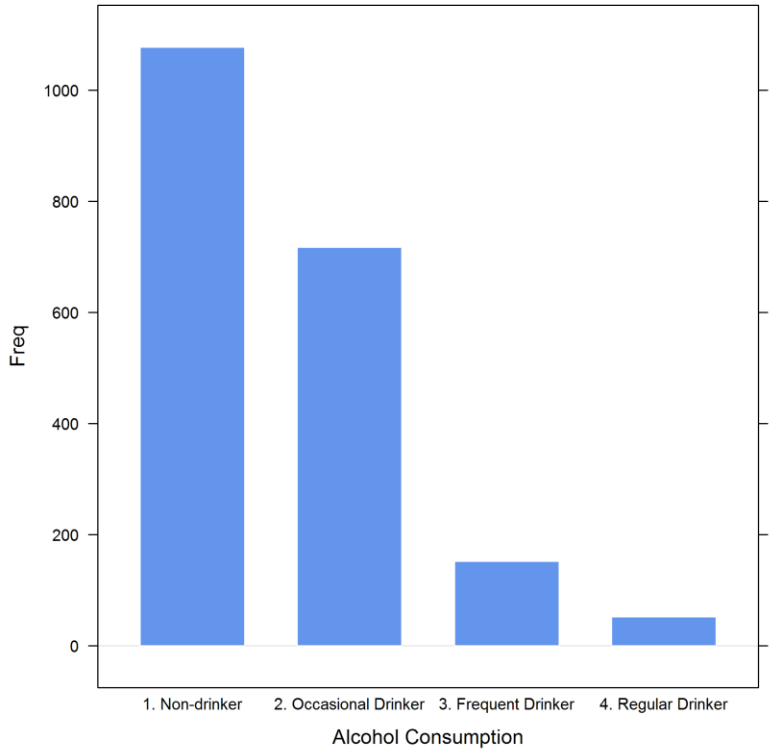
Alcohol Consumption	Non-drinker	Occasional Drinker	Frequent Drinker	Regular Drinker
n	1078	718	152	52

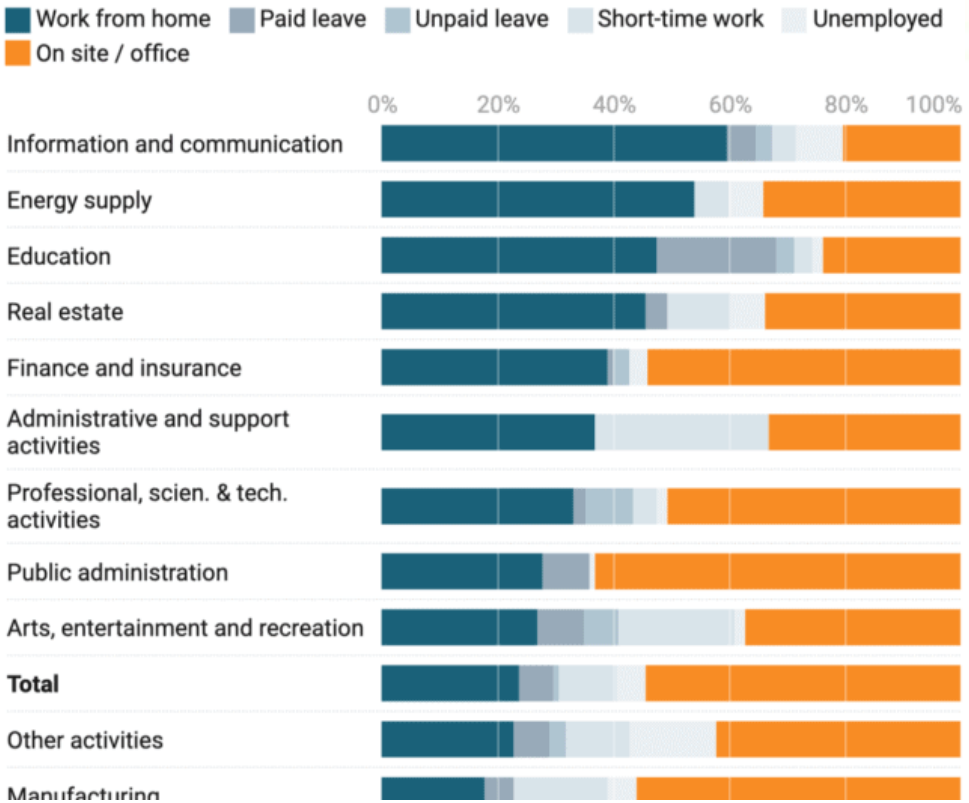
Each cell represents the number of participants in each category

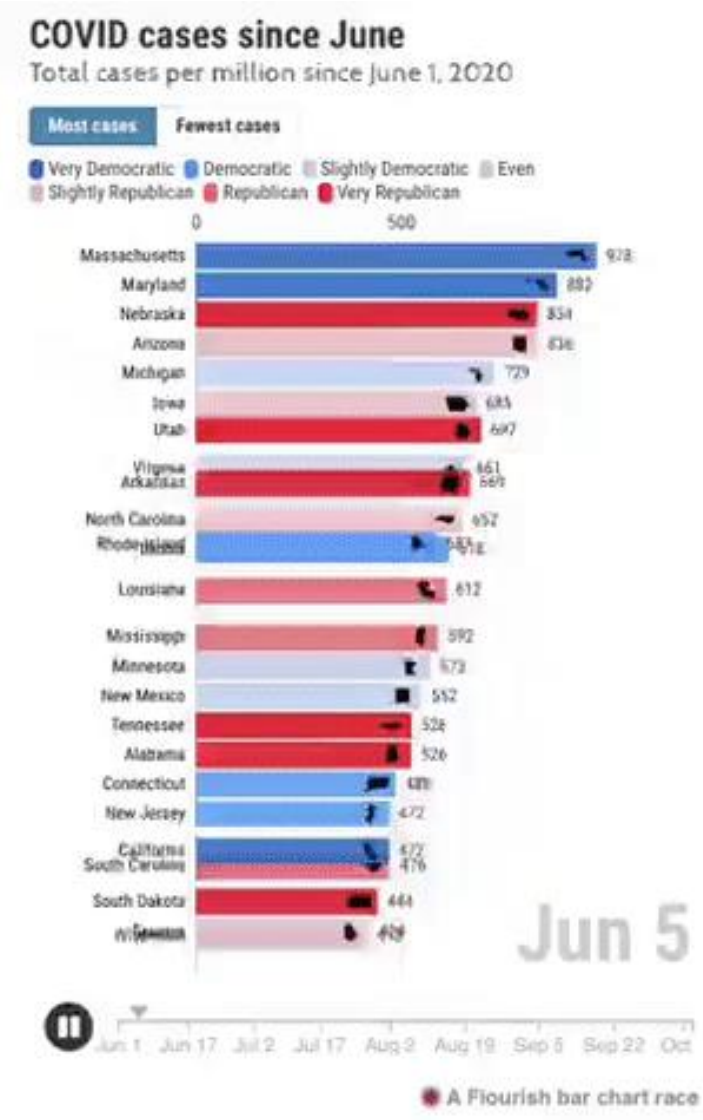
Data for a single categorical variable can summarised by counting the number of observations in each category.

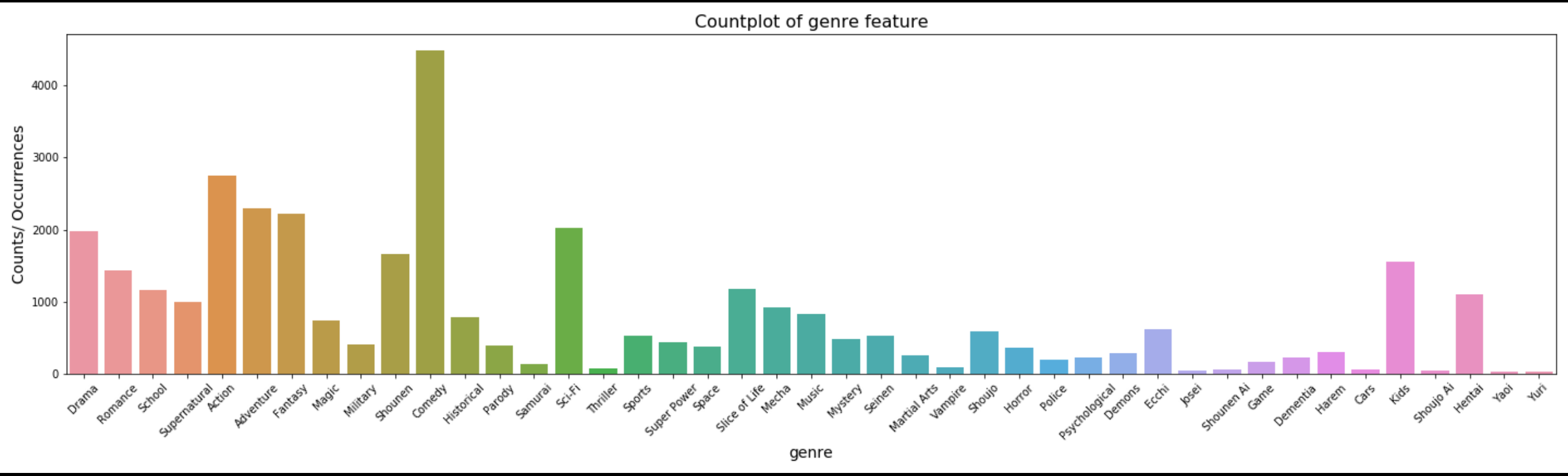
Using bar chart

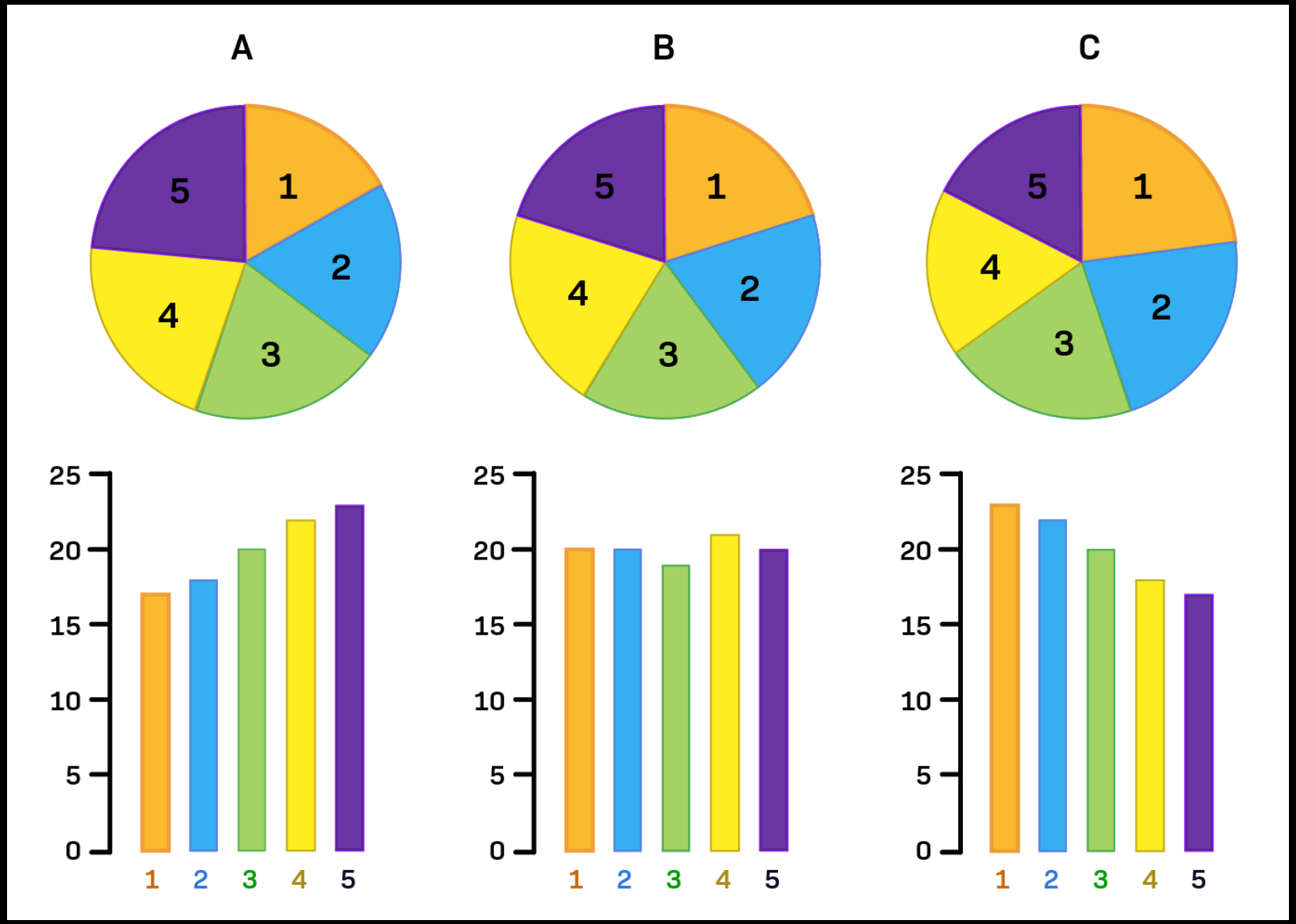
Visualise the data on a bar chart.








Countplot of genre feature
Counts/ Occurrences
0
1000
2000
3000
4000
genre
Drama
Romance
School
Supernatural
Action
Adventure
Fantasy
Magic
Military
Shounen
Comedy
Historical
Parody
Samurai
Sci-Fi
Thriller
Sports
Super Power
Space
Slice of Life
Mecha
Music
Mystery
Seinen
Martial Arts
Vampire
Shoujo
Horror
Police
Psychological
Demons
Ecchi
Josei
Shounen Ai
Game
Dementia
Harem
Cars
Kids
Shoujo Ai
Hentai
Yaoi
Yuri



Just eat pies…

Don't make them into charts

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%


Time for
R bu bu

Thank you