Introduction to statistics

(Day 1)

Housekeeping

- **Who:**
  - Paola Dalmasso
    paola.dalmasso@unito.it
  - Alessia Visconti
    alessia.visconti@unito.it

- **Exam:**
  - Multiple-choice questions
    (*via* Moodle)

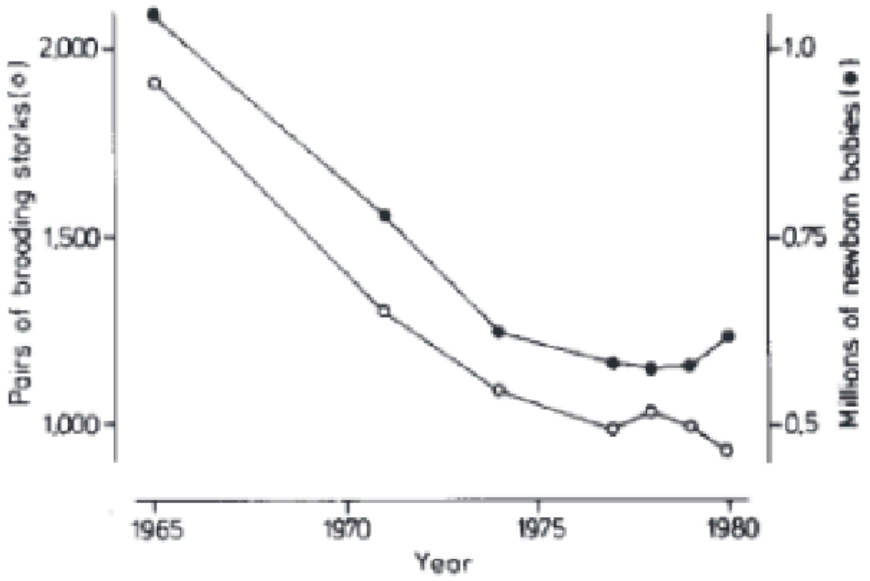
Introduction

Why are we here?

Will you buy this mouthwash?

Storks and babies

SIR—There is concern in West Germany over the falling birth rate. The accompanying graph[1,2] might suggest a solution that every child knows makes sense.



Sies, H., *A new parameter for sex education*. Nature, 1988, doi:10.1038/332495a0

Milk and tumours

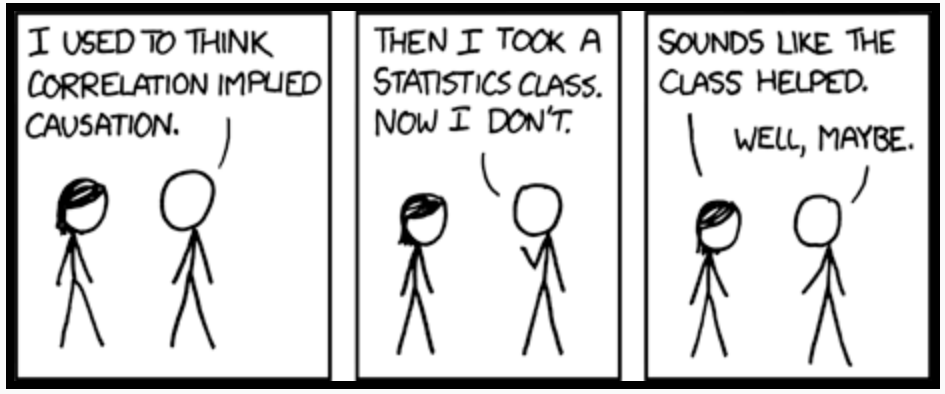
- English women, who consume large quantities of milk, develop some types of tumours 18 times more frequently than Japanese women, who rarely drink it

- Tumours usually appear late in life.
- British women live, on average, 12 years more than Japanese women

Huff, D., *How to lie with statistics*. W. W. Norton & Company, 1954

Why are we here?

- Because "numbers" (or rather, the way they are presented) are sometimes deceiving

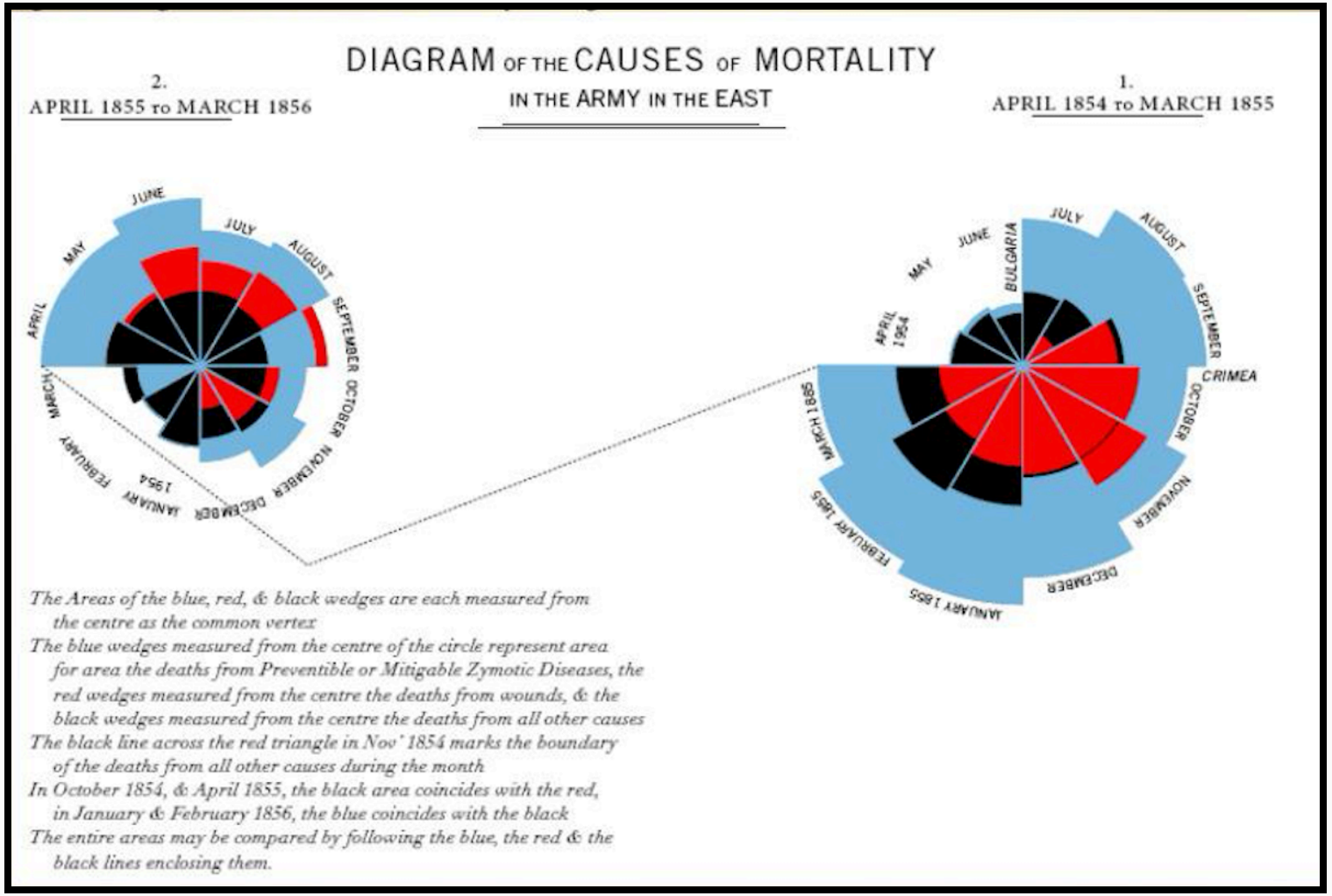


Why are we here?

- Why do soldiers die?

Who is this person?

Why do soldiers die?



Why are we here?

- Because "numbers" (or rather, the way they are presented) are sometimes deceiving
- Because “numbers” (and the way they are presented) help us describe, understand, and change the world

Why are we here?

- To learn how to read, understand, and critically analyse scientific papers
- To be able to carry out research involving the acquisition, processing, and analysis of data

Scientific research

Problem

- Why do soldiers die?

Plan

- How do we answer?

Data

Scutari Hospital. J.A. Benwell. about 1856

Analysis

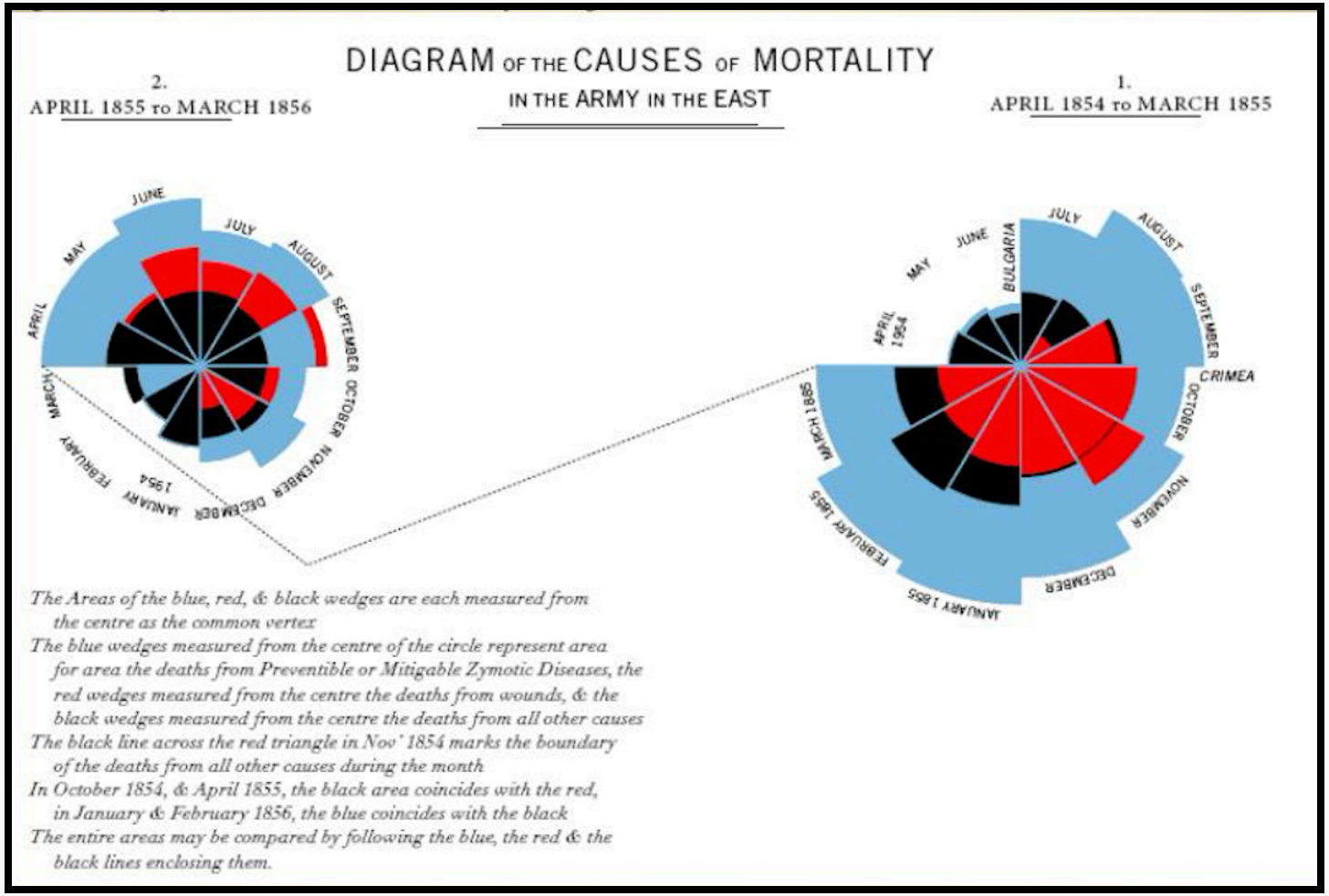
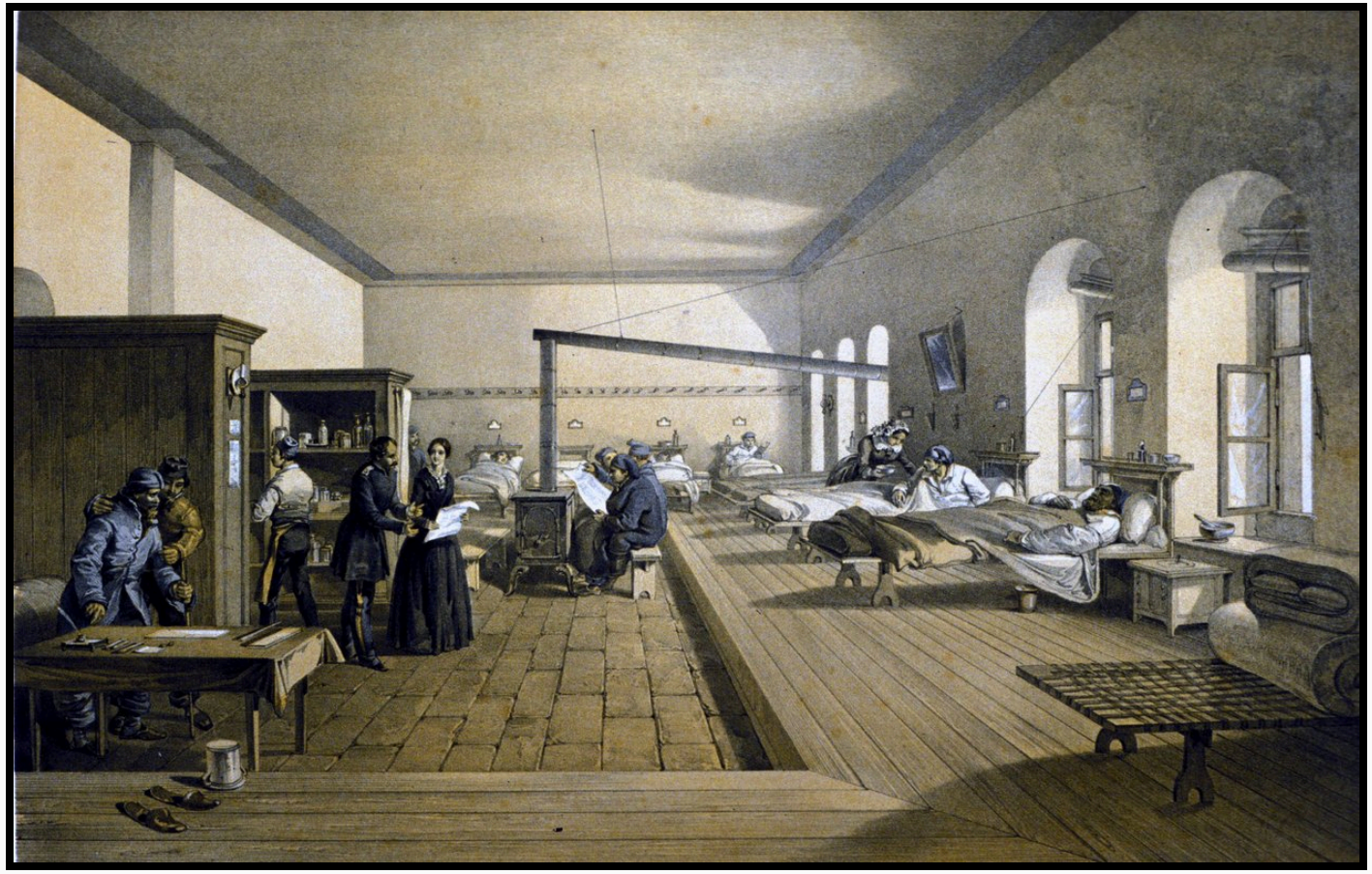


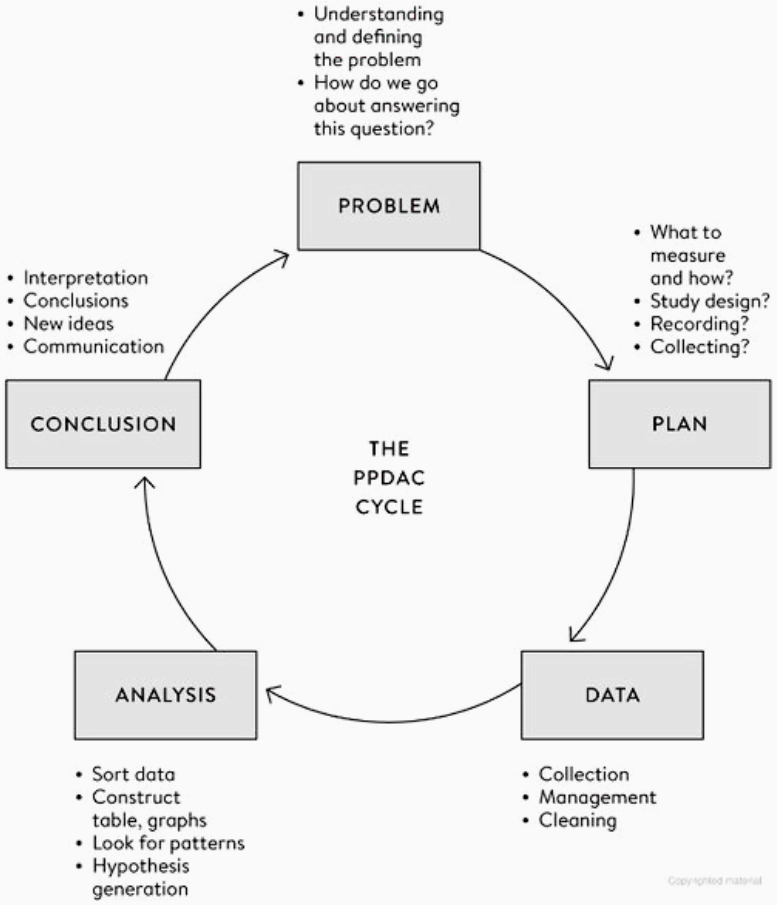
DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST
2.
APRIL 1855 to MARCH 1856
1.
APRIL 1854 to MARCH 1855
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRUARY
MARCH
APRIL
MAY
1854
BULGARIA
CRIMEA
JANUARY 1855
FEBRUARY 1855
MARCH 1855
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes
The black line across the red triangle in Nov' 1854 marks the boundary of the deaths from all other causes during the month
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Conclusions



Florence Nightingale (1820 - 1910) at Scutari Hospital in Turkey around 1855, unknown artist

The data problem-solving cycle



Spiegelhalter, D., *The Art of Statistics: Learning From Data*, Pelican, 2019

Inductive reasoning

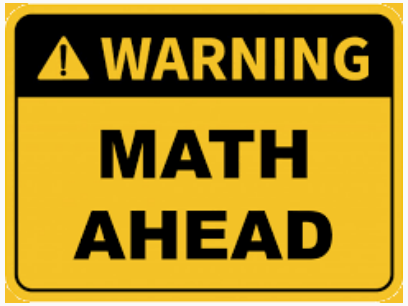
Derive a general rule from observations

What is Statistics?

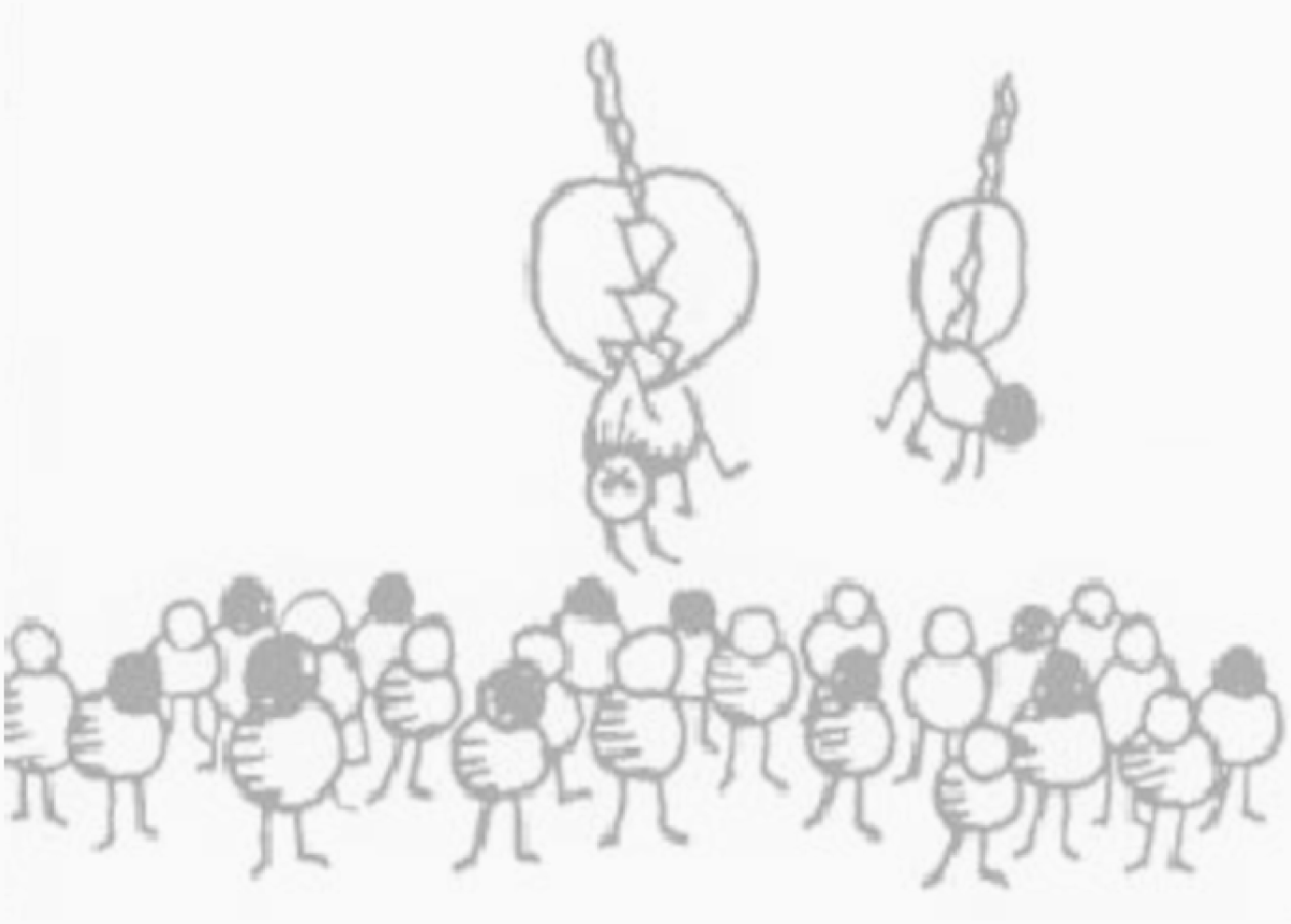
- The collection, organisation, summarisation, and analysis of data
  → *Descriptive* statistics
- The drawing of inferences about a body of data when only a part of the data is observed
  → *Inferential* statistics

What will we learn?

- How to collect data
- How to summarise data
- How to make decision with data

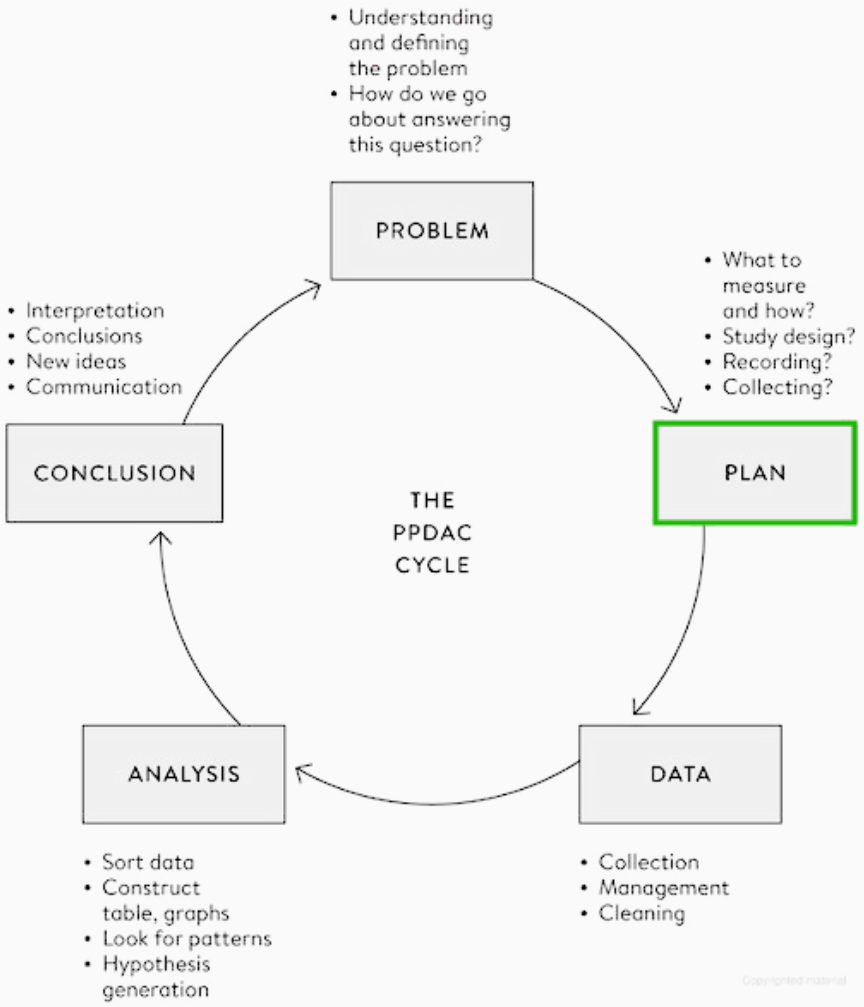


Sampling



Learning objectives

- Understand the difference between population and sample
- Understand the difference between sampling strategies
- Understand sampling error and bias



Spiegelhalter, D., *The Art of Statistics: Learning From Data*, Pelican, 2019

Population *vs* sample

Istat | Istituto Nazionale di Statistica

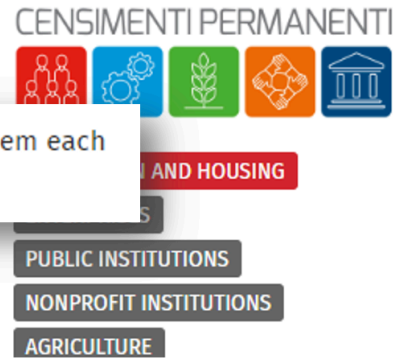
POPULATIONS AND SAMPLES

PERMANENT CENSUS OF POPULATION AND HOUSING

The permanent census of the population and housing begins in October 2018. For the first time ISTAT conducts not a ten-yearly but an annual survey of the main characteristics of the country's resident population and its social and economic conditions at national, regional and local levels.

The new permanent census of population and housing do not involve all Italian families, but a sample of them each year: about 1,400,000 families resident in 2,800 Italian municipalities.

Moreover, only a percentage of the municipalities (about 1,100 of them) will take part by census operations every year; the remainder will be called to participate once every four years. In this way, all municipalities will be surveyed at least once by 2021.



Population *vs* sample (in the clinic)

Delirium as a Predictor of Mortality in Mechanically Ventilated Patients in the Intensive Care Unit

E. Wesley Ely, MD, MPH
Ayumi Shintani, PhD, MPH
Brenda Truman, RN, MSN
Theodore Speroff, PhD
Sharon M. Gordon, PsyD
Frank E. Harrell, Jr, PhD
Sharon K. Inouye, MD, MPH
Gordon R. Bernard, MD
Robert S. Dittus, MD, MPH

Context In the intensive care unit (ICU), delirium is a common yet underdiagnosed form of organ dysfunction, and its contribution to patient outcomes is unclear.

Objective To determine if delirium is an independent predictor of clinical outcomes, including 6-month mortality and length of stay among ICU patients receiving mechanical ventilation.

Design, Setting, and Participants Prospective cohort study enrolling 275 consecutive mechanically ventilated patients admitted to adult medical and coronary ICUs of a US university-based medical center between February 2000 and May 2001. Patients were followed up for development of delirium over 2158 ICU days using the Confusion Assessment Method for the ICU and the Richmond Agitation-Sedation Scale.

Exercise #1

Objective To determine whether intravenous dexamethasone increases the number of ventilator-free days among patients with COVID-19–associated ARDS.

Design, Setting, and Participants Multicenter, randomized, open-label, clinical trial conducted in 41 intensive care units (ICUs) in Brazil. Patients with COVID-19 and moderate to severe ARDS, according to the Berlin definition, were enrolled from April 17 to June 23, 2020. Final follow-up was completed on July 21, 2020. The trial was stopped early following publication of a related study before reaching the planned sample size of 350 patients.

? Which is the population of this study?

a) Patients with Acute respiratory distress syndrome (ARDS)
b) Patients with COVID-19–associated ARDS
c) Brazilian patients with COVID-19–associated ARDS
d) Patients without COVID-19–associated ARDS

01:00

Exercise #1 -- Solution

Objective To determine whether intravenous dexamethasone increases the number of ventilator-free days among patients with COVID-19–associated ARDS.

Design, Setting, and Participants Multicenter, randomized, open-label, clinical trial conducted in 41 intensive care units (ICUs) in Brazil. Patients with COVID-19 and moderate to severe ARDS, according to the Berlin definition, were enrolled from April 17 to June 23, 2020. Final follow-up was completed on July 21, 2020. The trial was stopped early following publication of a related study before reaching the planned sample size of 350 patients.

? Which is the population of this study?

a) Patients with Acute respiratory distress syndrome (ARDS)
b) Patients with COVID-19–associated ARDS ✅
c) Brazilian patients with COVID-19–associated ARDS
d) Patients without COVID-19–associated ARDS

Exercise #2

Objective To determine whether intravenous dexamethasone increases the number of ventilator-free days among patients with COVID-19–associated ARDS.

Design, Setting, and Participants Multicenter, randomized, open-label, clinical trial conducted in 41 intensive care units (ICUs) in Brazil. Patients with COVID-19 and moderate to severe ARDS, according to the Berlin definition, were enrolled from April 17 to June 23, 2020. Final follow-up was completed on July 21, 2020. The trial was stopped early following publication of a related study before reaching the planned sample size of 350 patients.

? Which is the sample used in this study?

a) Patients with Acute respiratory distress syndrome (ARDS)
b) Patients with COVID-19–associated ARDS
c) Brazilian patients with COVID-19–associated ARDS
d) Patients without COVID-19–associated ARDS

01:00

Exercise #2 -- Solution

Objective To determine whether intravenous dexamethasone increases the number of ventilator-free days among patients with COVID-19–associated ARDS.

Design, Setting, and Participants Multicenter, randomized, open-label, clinical trial conducted in 41 intensive care units (ICUs) in Brazil. Patients with COVID-19 and moderate to severe ARDS, according to the Berlin definition, were enrolled from April 17 to June 23, 2020. Final follow-up was completed on July 21, 2020. The trial was stopped early following publication of a related study before reaching the planned sample size of 350 patients.

? Which is the sample used in this study?

a) Patients with Acute respiratory distress syndrome (ARDS)
b) Patients with COVID-19–associated ARDS
c) Brazilian patients with COVID-19–associated ARDS ✅
d) Patients without COVID-19–associated ARDS

Opportunity *vs* random sample

🎯 An **opportunity** sample is the sample drawn from the part of the population that is close to hand (and which may not represent the whole population)

📌 All the patients presenting to a given clinic in a given period of time are enrolled

Opportunity *vs* random sample

🎯 A **random** sample is the sample in which the probability of getting any particular sample may be calculated (and which should represent the whole population)

📌 A randomly selected set of patients with the disease is enrolled

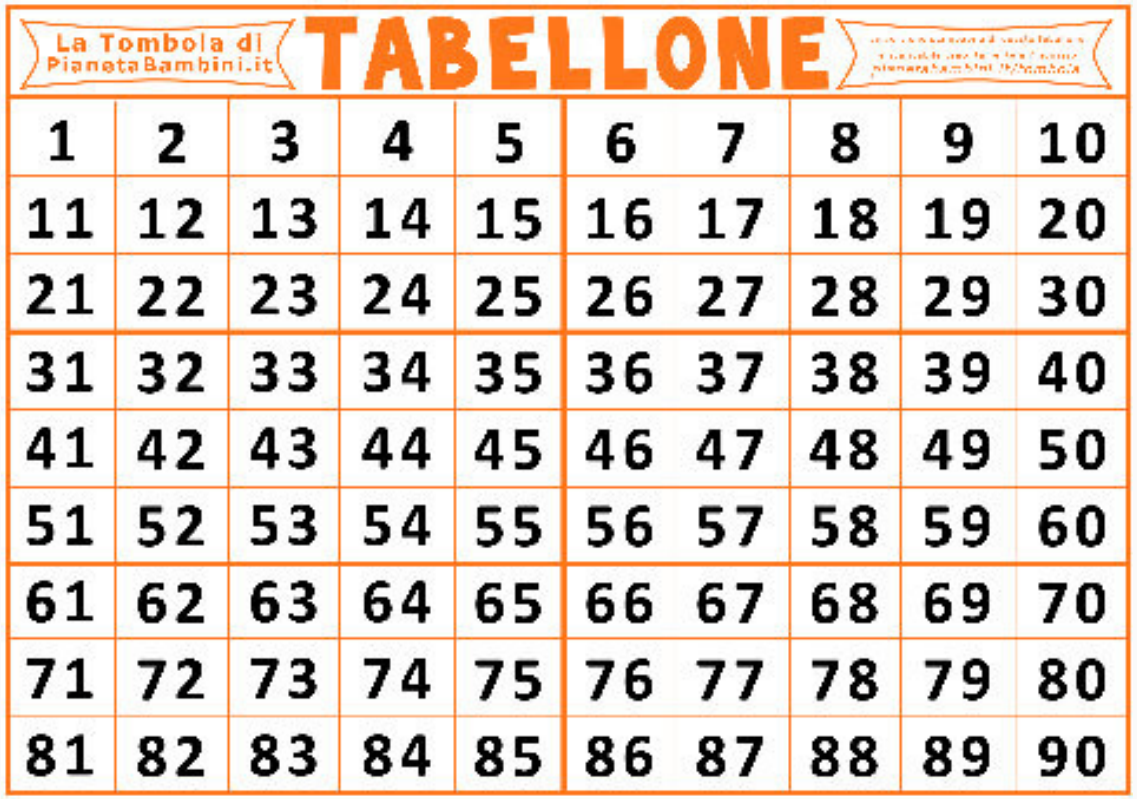
Strategy 1: Simple random sampling

🎯 A sample of size n drawn from a population of size N ensuring that every possible sample of size n is equally likely

Strategy 1: Simple random sampling

📌 $N = 90$

$n = 10$

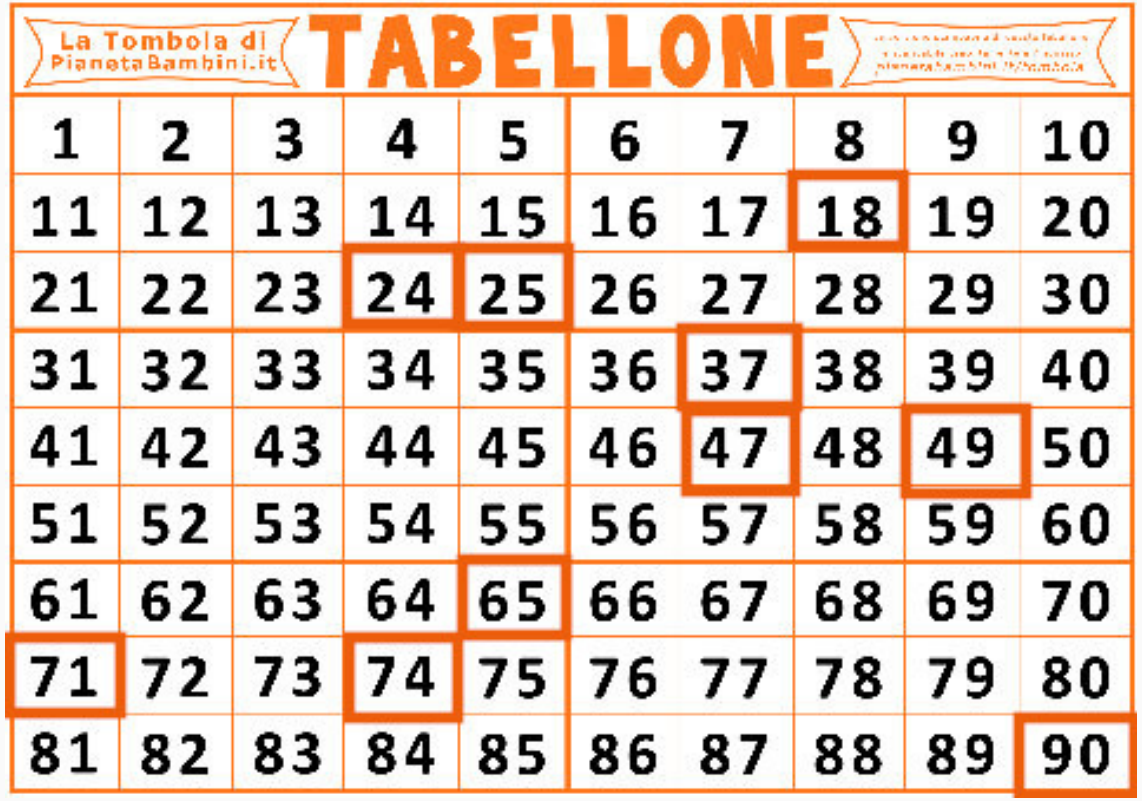


1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90

Strategy 1: Simple random sampling

📌 $N = 90$
$n = 10$

$49, 65, 25, 74, 18$
$90, 47, 24, 71, 37$



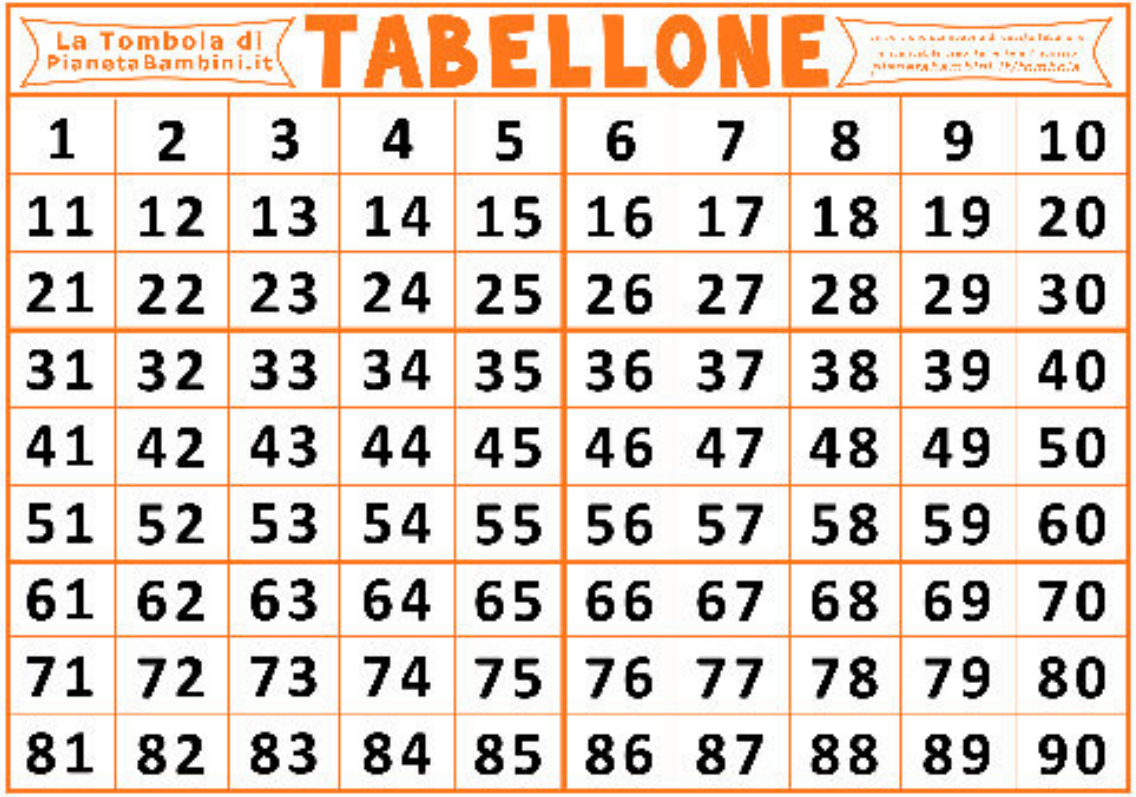
Strategy 2: Systematic Sampling

📌 $N = 90$

$n = 10$

$x = 42$

$step = N/n = 90/10 = 9$



1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90

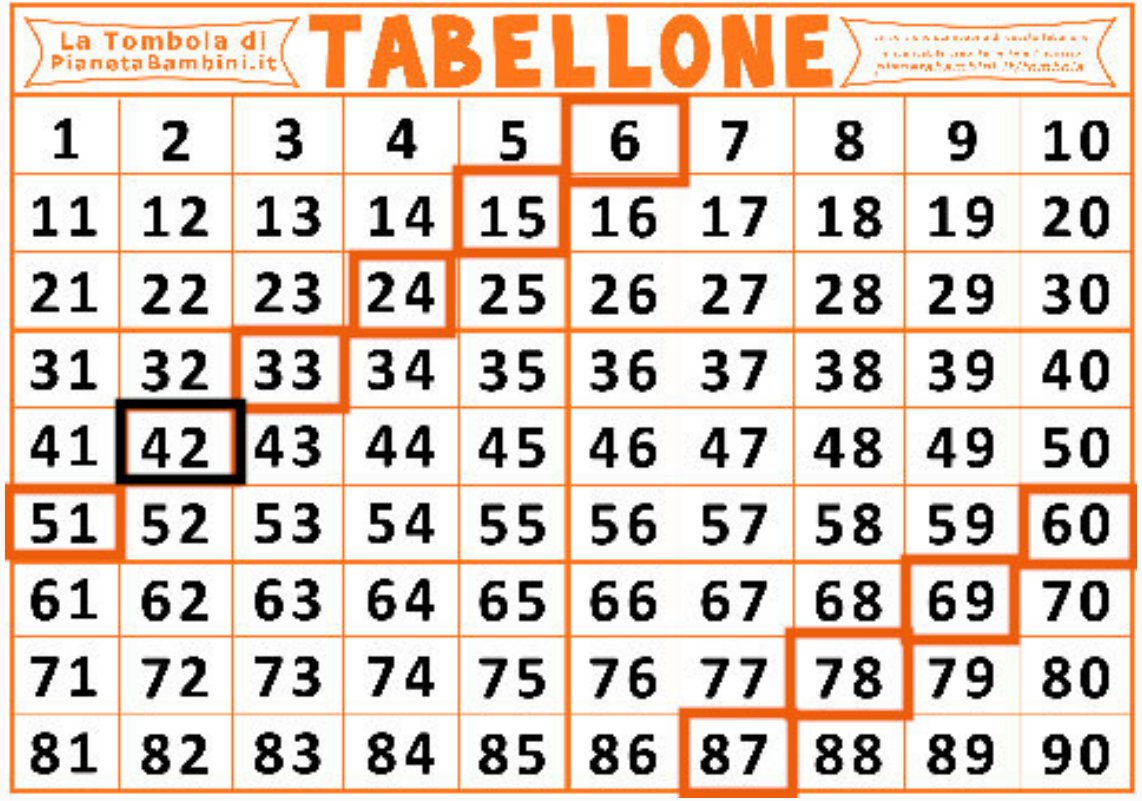
Strategy 2: Systematic Sampling

📌 $N = 90$

$n = 10$

$x = 42$

$step = N/n = 90/10 = 9$



Strategy 3: Stratified Random Sampling

🎯 The population is divided into homogenous group (strata) and a simple random sample is drawn from each stratum

Variation #1: stratified systematic sample
Variation #2: stratified sampling proportional to size

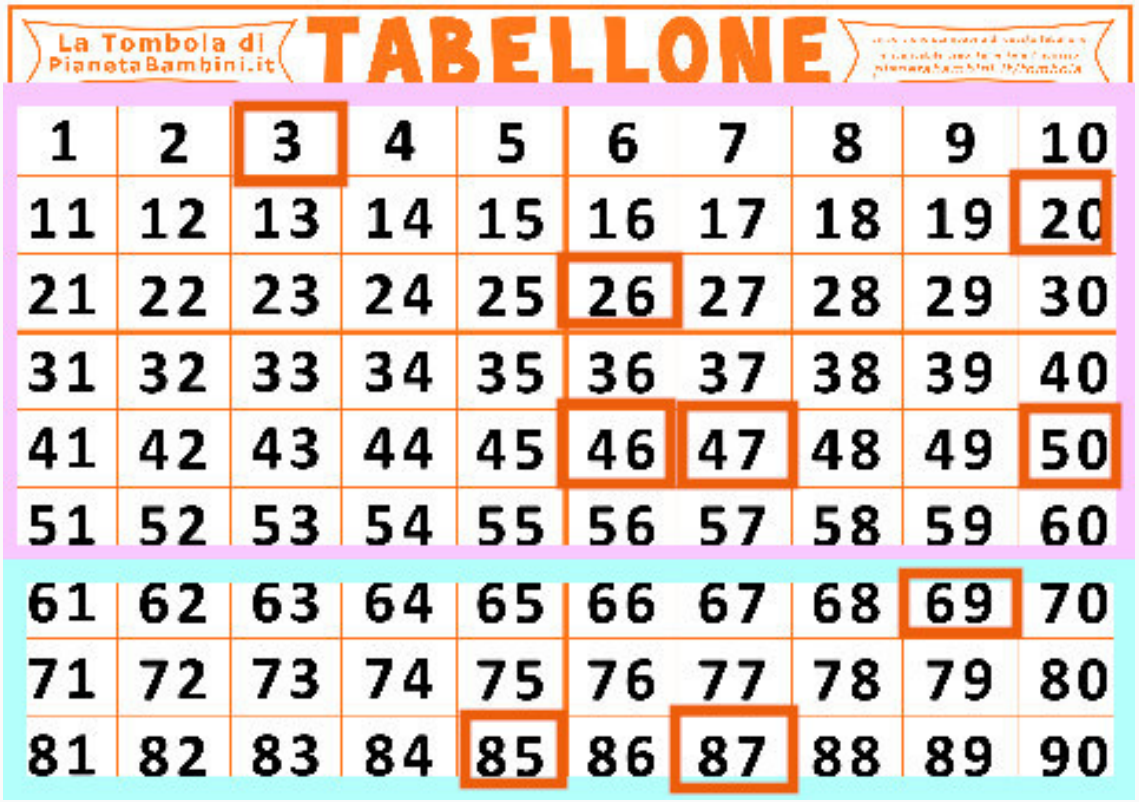
Strategy 3: Stratified Random Sampling

📌

$$N = 90$$
$$N_{female} = 60$$
$$N_{male} = 30$$

$$n = 9$$
$$n_{female} = 6$$
$$n_{male} = 3$$

$$Females : 46, 20, 26, 50, 47, 3$$
$$Males : 69, 85, 87$$



Strategy 4: Cluster sampling

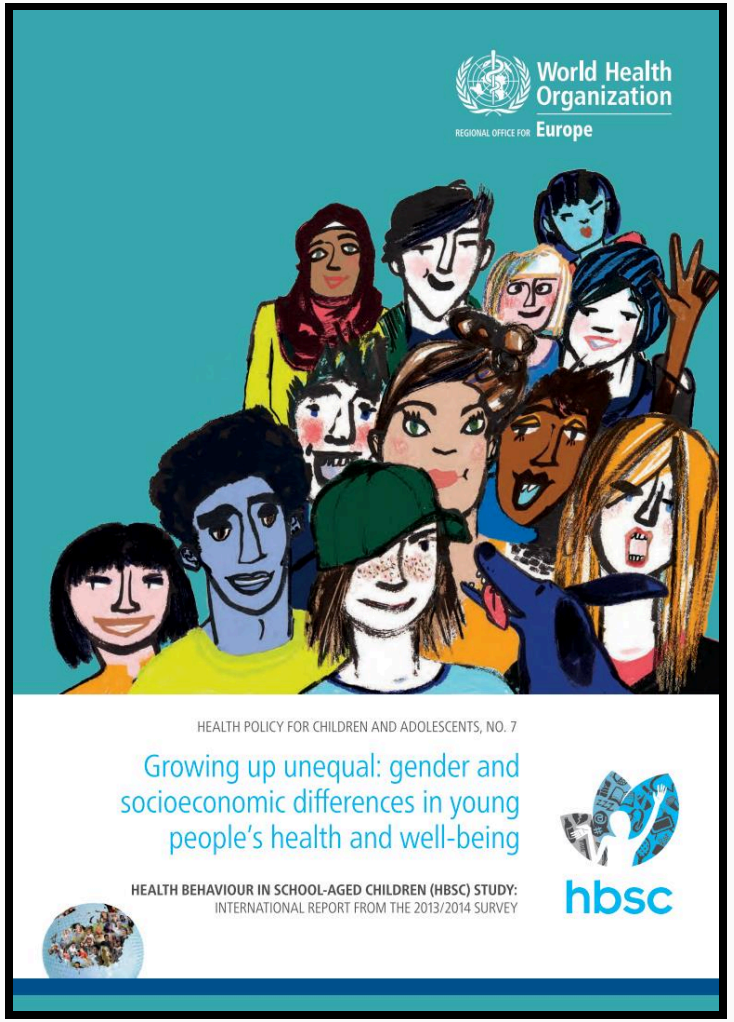
🎯 The population is divided into clusters, and a simple random sample is drawn

Variation: one stage (observing everything) *vs*
two stage (sampling within clusters)

Strategy 4: Cluster sampling

Sampling in the wild

https://hbsc.org



Exercise #3

? A representative of a cheese factory is asking questions on cheese consumption to every 5th customer entering the supermarket

Which kind of sampling strategy are they using?

a) simple random sampling
b) systematic sampling
c) stratified sampling
d) none of the above

01:00

Exercise #3 -- Solution

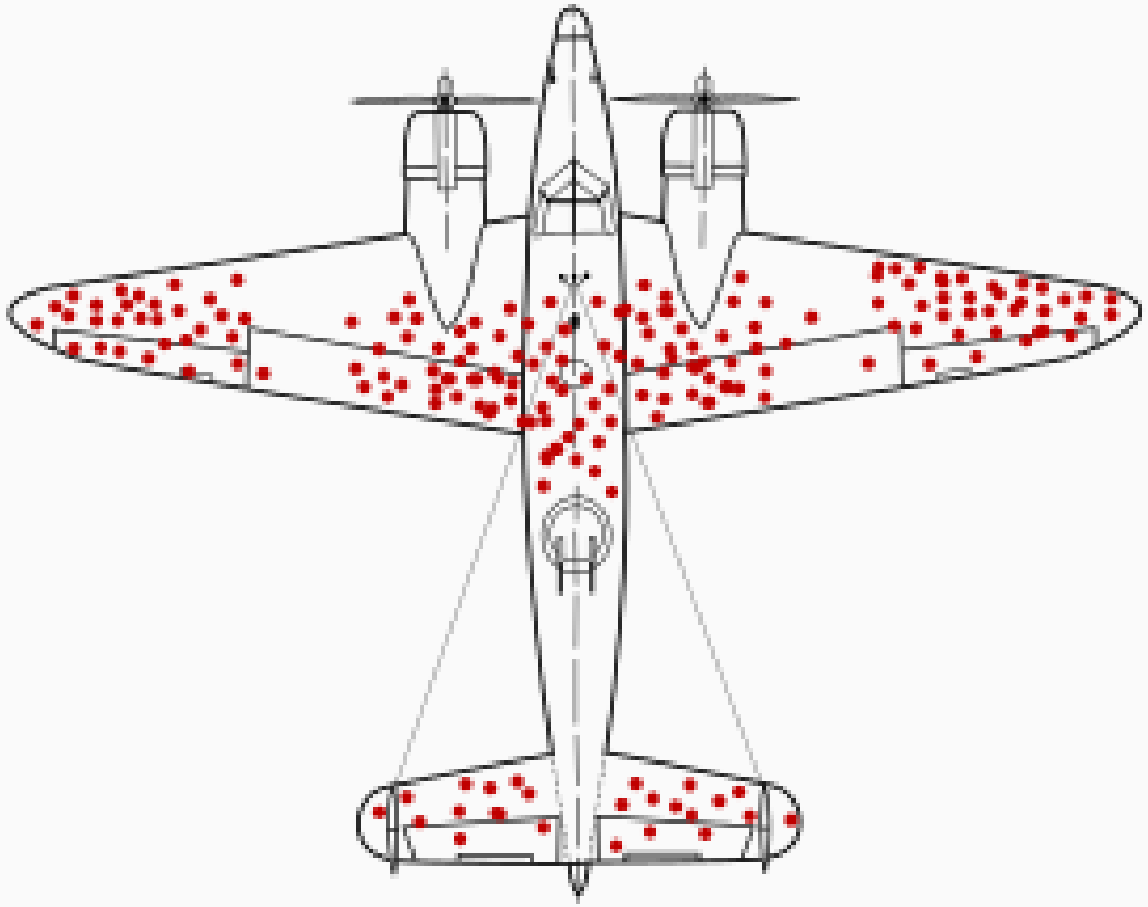
❓ A representative of a cheese factory is asking questions on cheese consumption to every 5th customer entering the supermarket

Which kind of sampling strategy are they using?

a) simple random sampling
b) systematic sampling
c) stratified sampling
d) none of the above ✅

Selection bias

- Survivor bias



Selection bias

- Survivor bias

    "*Buildings used to be more beautiful/longer lasting*"

Selection bias

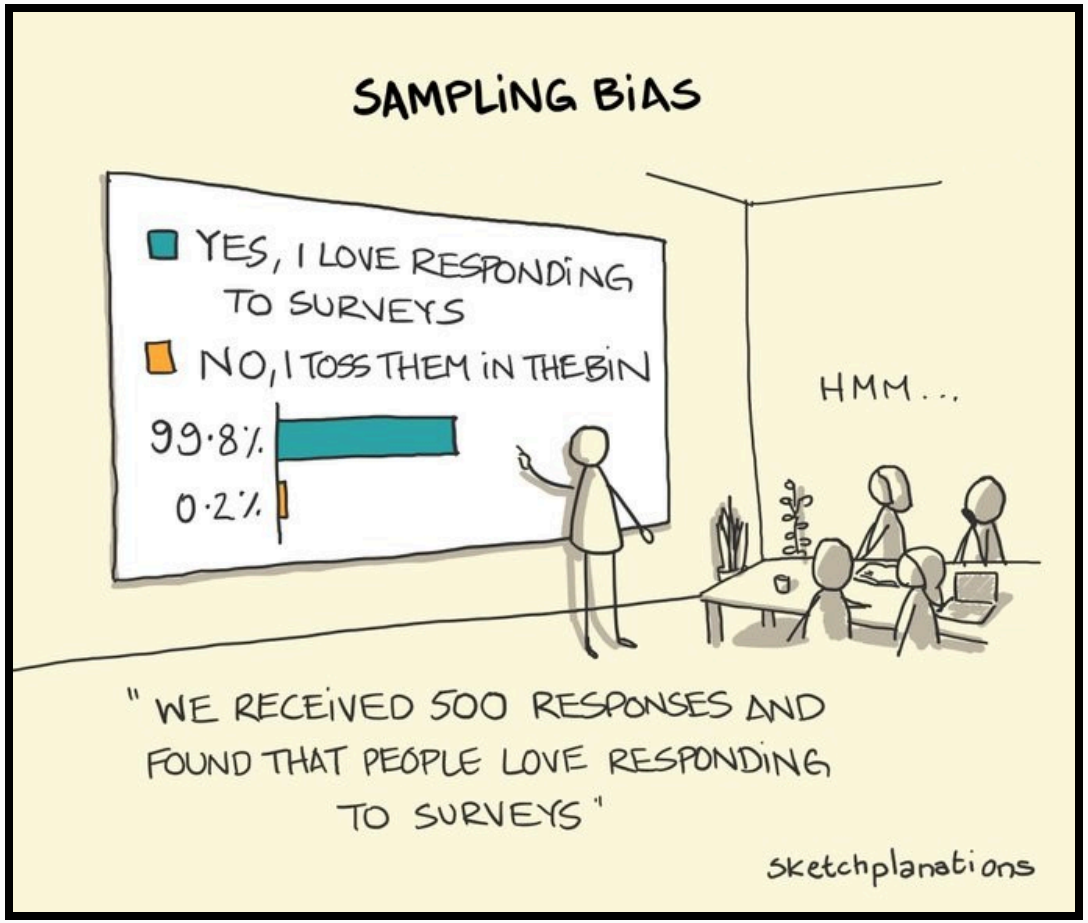
- Survivor bias

  "*Buildings used to be more beautiful/longer lasting*"

  "*I don't understand why, nowadays, one should to do X, I never did it, and I'm still here to tell it*"

Selection bias

- Survivor bias
- Volunteer bias



Selection bias

- Survivor bias
- Volunteer bias

> A teacher wonders if extra lessons improve exam performance. She prepares them, inviting interested students to sign up.

Selection bias

- Survivor bias
- Volunteer bias
- Lost to follow up bias

Selection bias

- Survivor bias
- Volunteer bias
- Lost to follow up bias

> A pharma company is testing a new drug on a court of 100 cancer patients recruited in a center of excellence, 30 of whom did not show up at the follow up. What do we conclude about this new drug knowing that these 30 people...

Selection bias

- Survivor bias
- Volunteer bias
- Lost to follow up bias

A pharma company is testing a new drug on a court of 100 cancer patients recruited in a center of excellence, 30 of whom did not show up at the follow up. What do we conclude about this new drug knowing that these 30 people...

- died?

Selection bias

- Survivor bias
- Volunteer bias
- Lost to follow up bias

A pharma company is testing a new drug on a court of 100 cancer patients recruited in a center of excellence, 30 of whom did not show up at the follow up. What do we conclude about this new drug knowing that these 30 people...

- died?
- stopped the drug?

Selection bias

- Survivor bias
- Volunteer bias
- Lost to follow up bias

A pharma company is testing a new drug on a court of 100 cancer patients recruited in a center of excellence, 30 of whom did not show up at the follow up. What do we conclude about this new drug knowing that these 30 people....

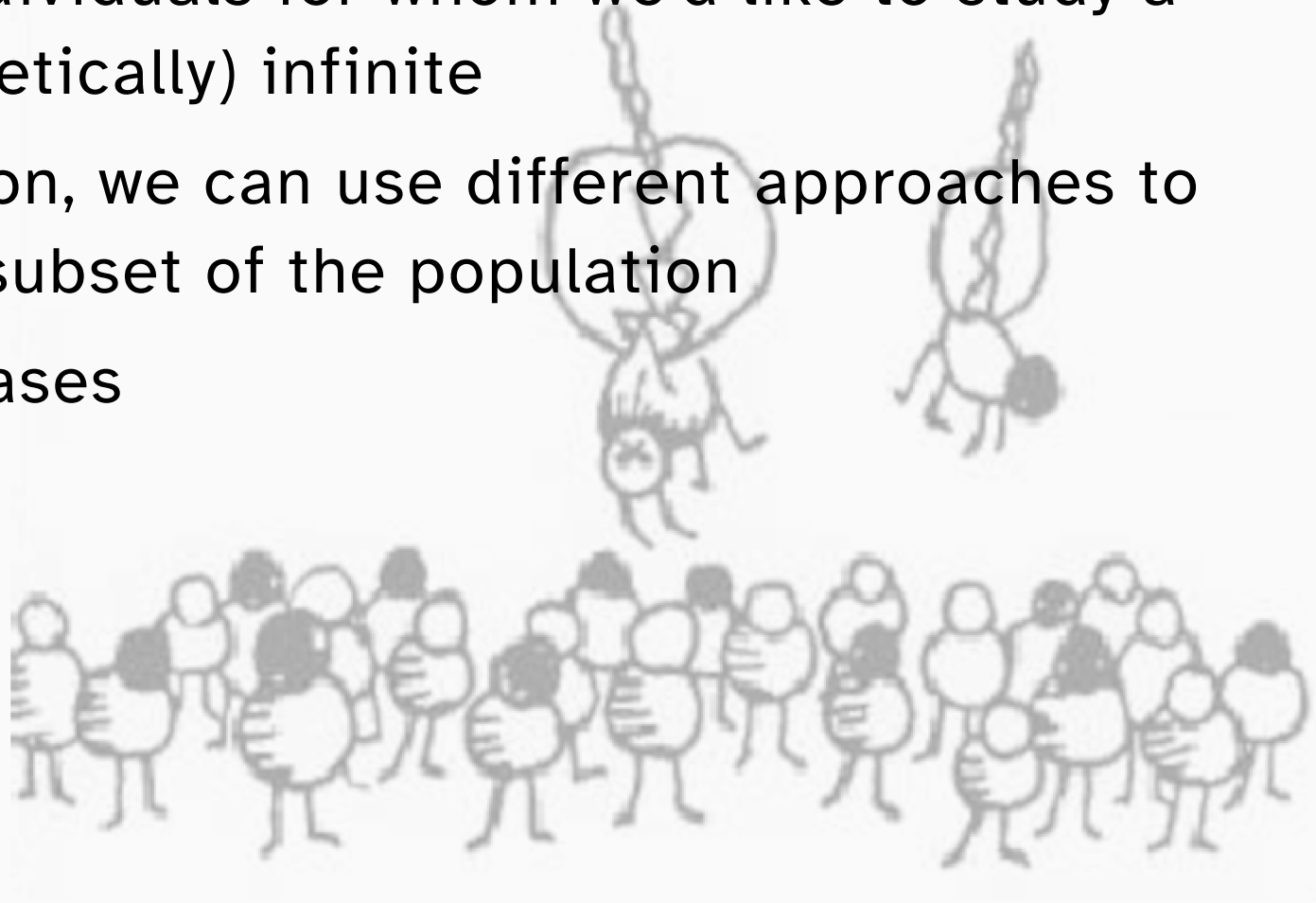
- died?
- stopped the drug?
- went back home?

Selection bias

- Survivor bias
- Volunteer bias
- Lost to follow up bias
- . . .

Summary

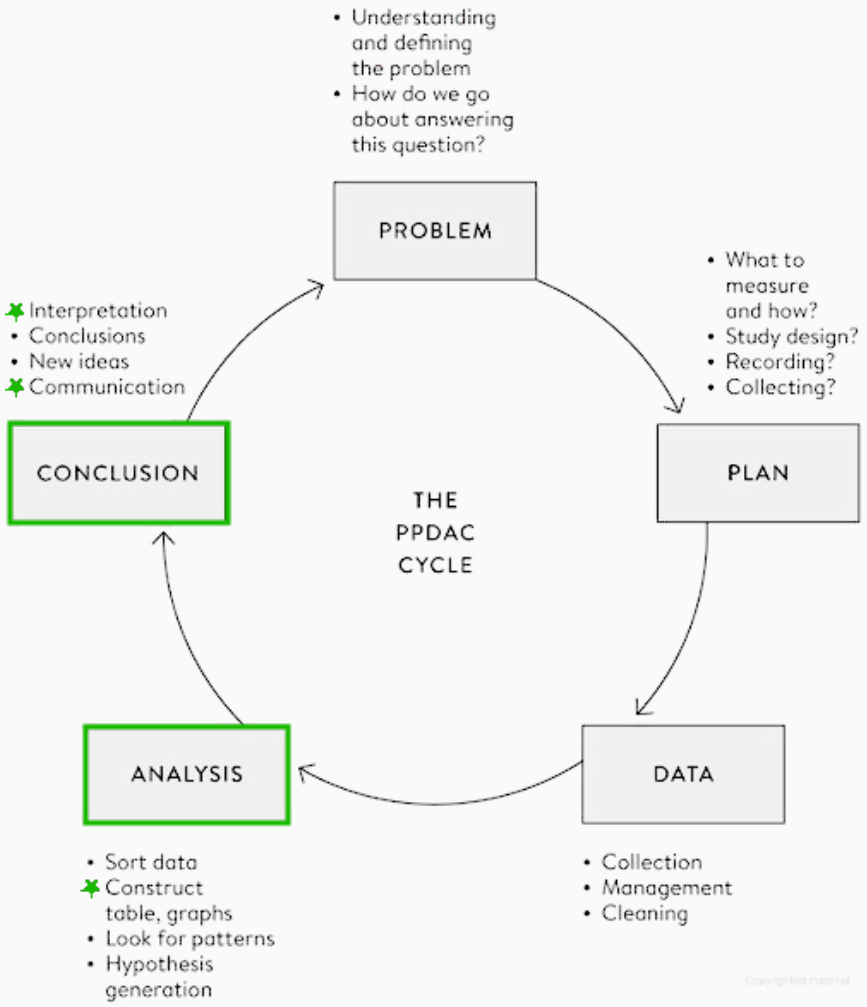
- A population includes all individuals for whom we'd like to study a phenomenon, and it's (theoretically) infinite
- When can't study a population, we can use different approaches to sample a (representative?) subset of the population
- Samples may suffer from biases



Summarise data

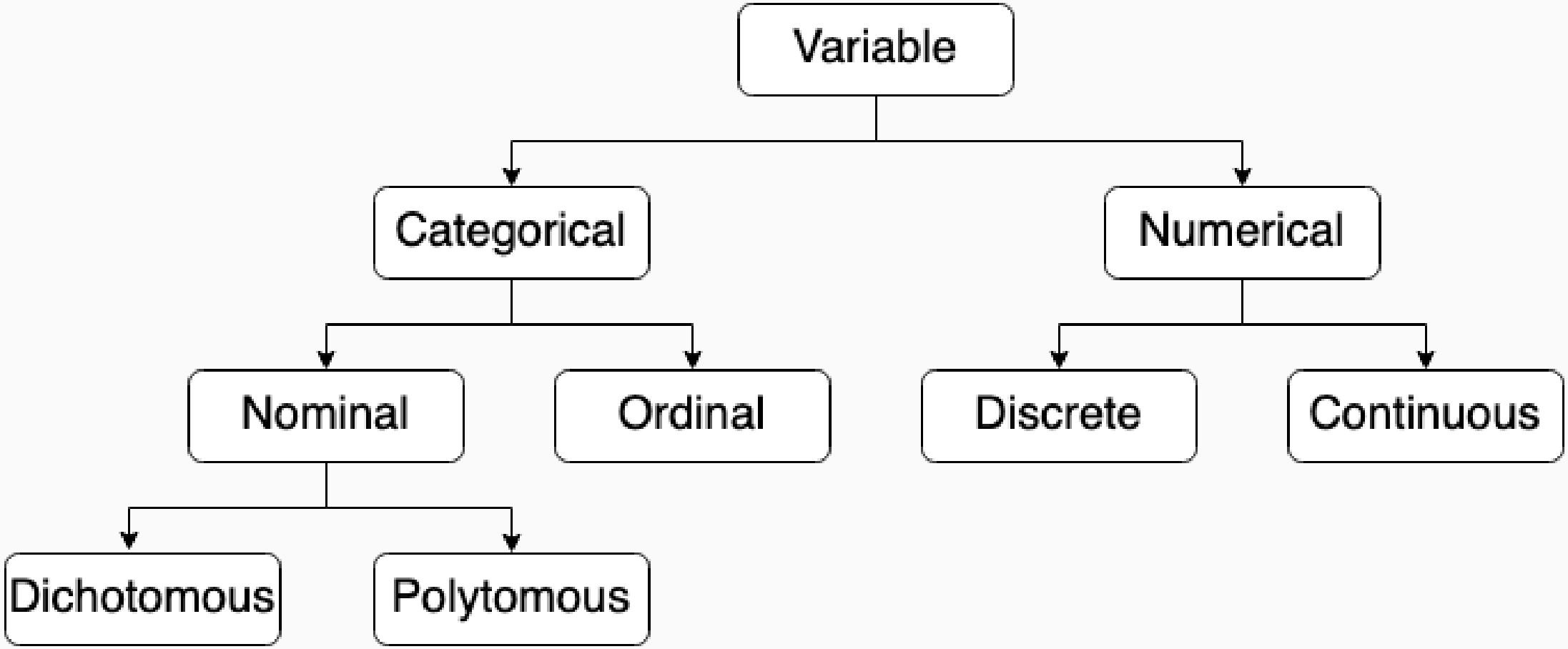
Learning objectives

- Understand the differences between data types
- Be able to summarise each data type
- Understand the difference between parameters and statistics
- Understand why visualise your data is important

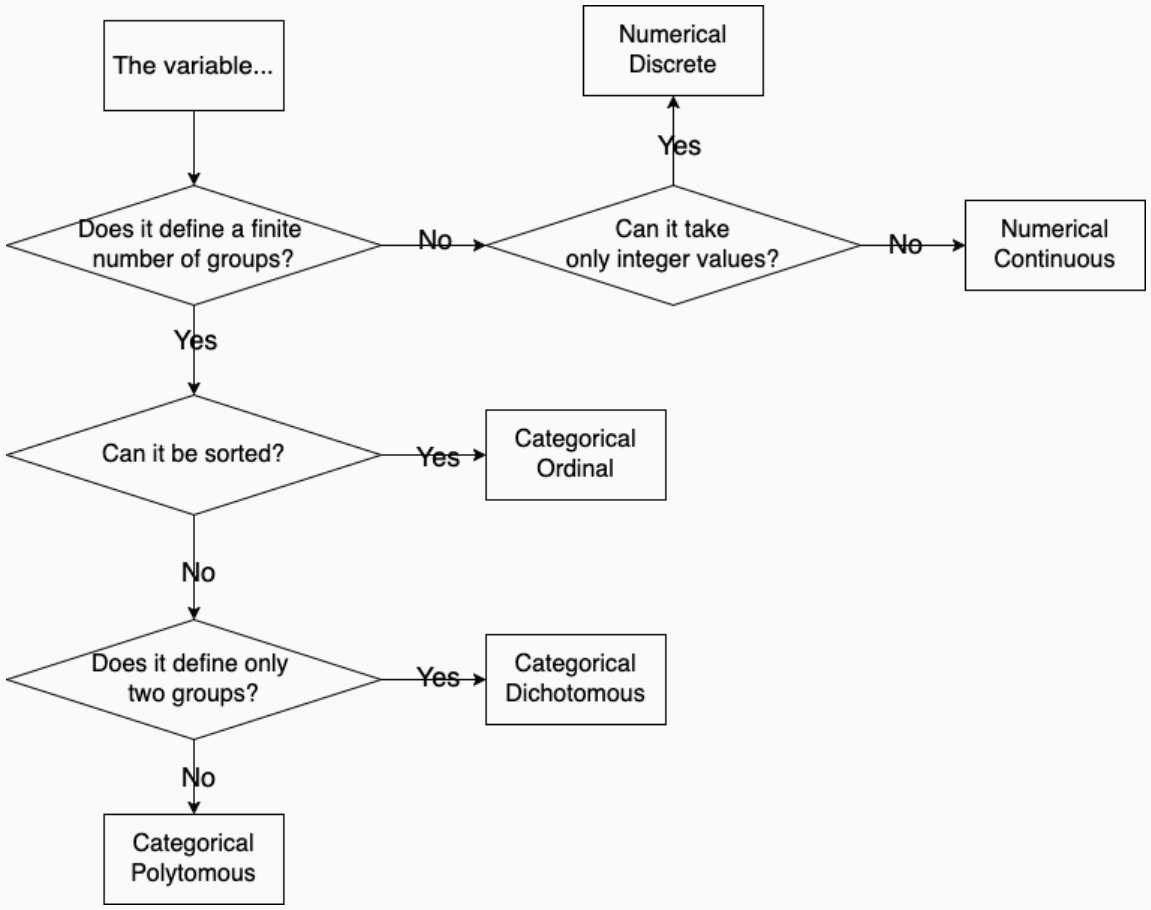


Spiegelhalter, D., *The Art of Statistics: Learning From Data*, Pelican, 2019

Type of data

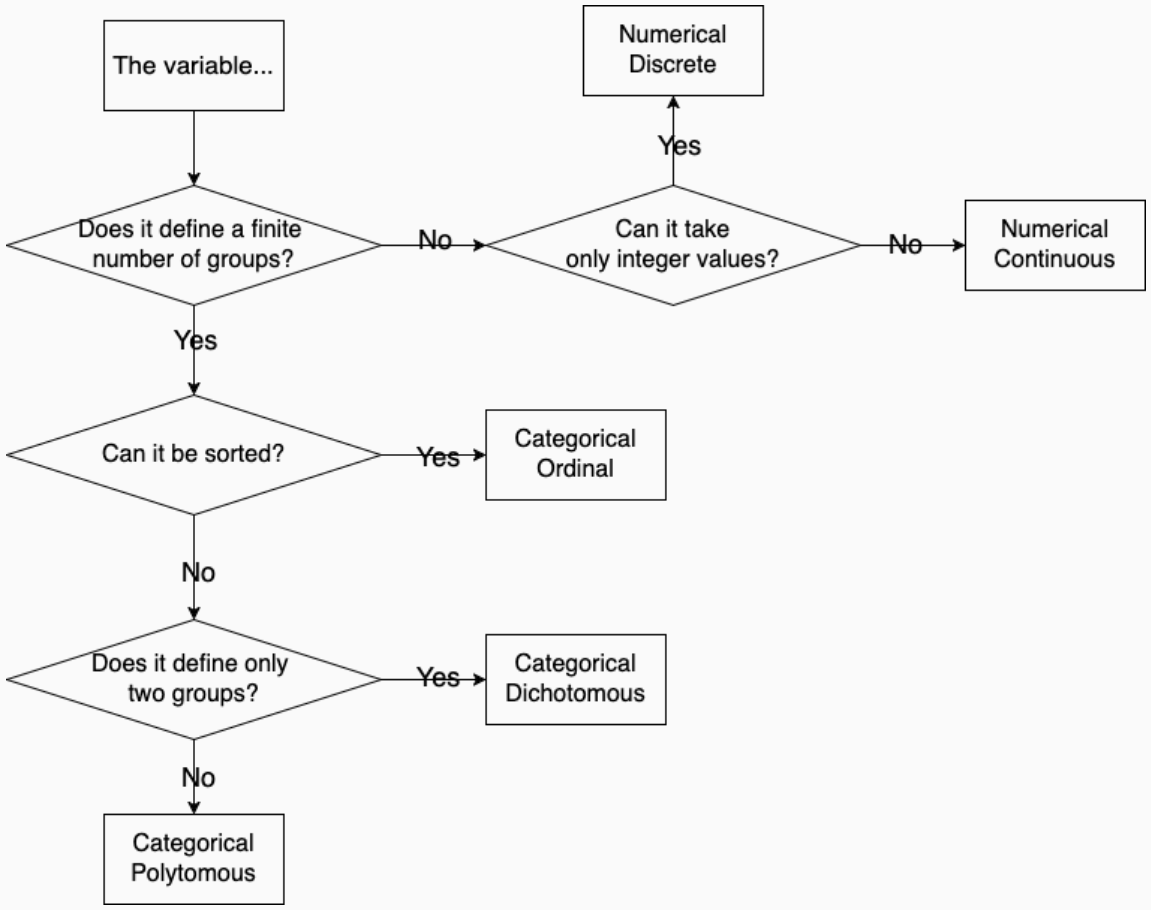

Variable
Categorical
Numerical
Nominal
Ordinal
Discrete
Continuous
Dichotomous
Polytomous

What type of data is this?



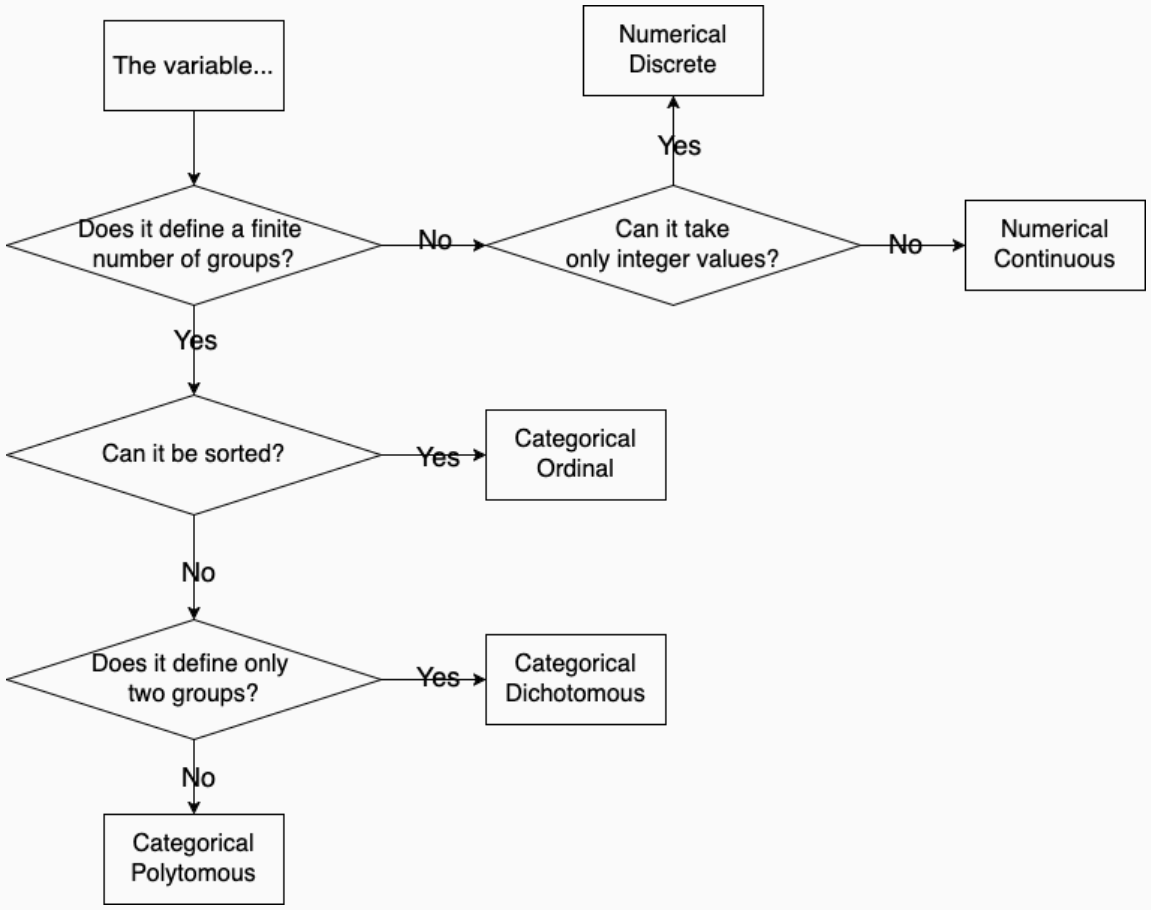
What type of data is this?

? The number of death in a hospital



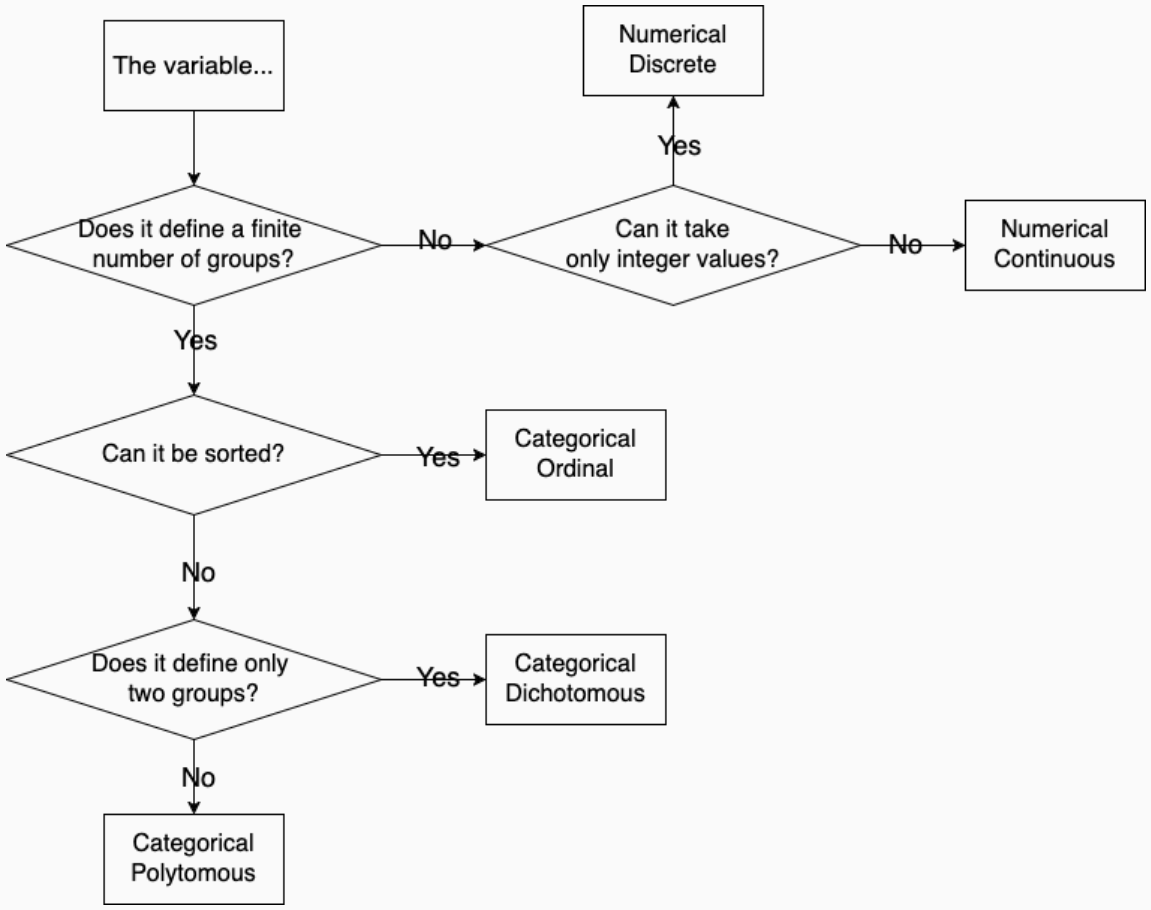
What type of data is this?

? The size of a T-shirt



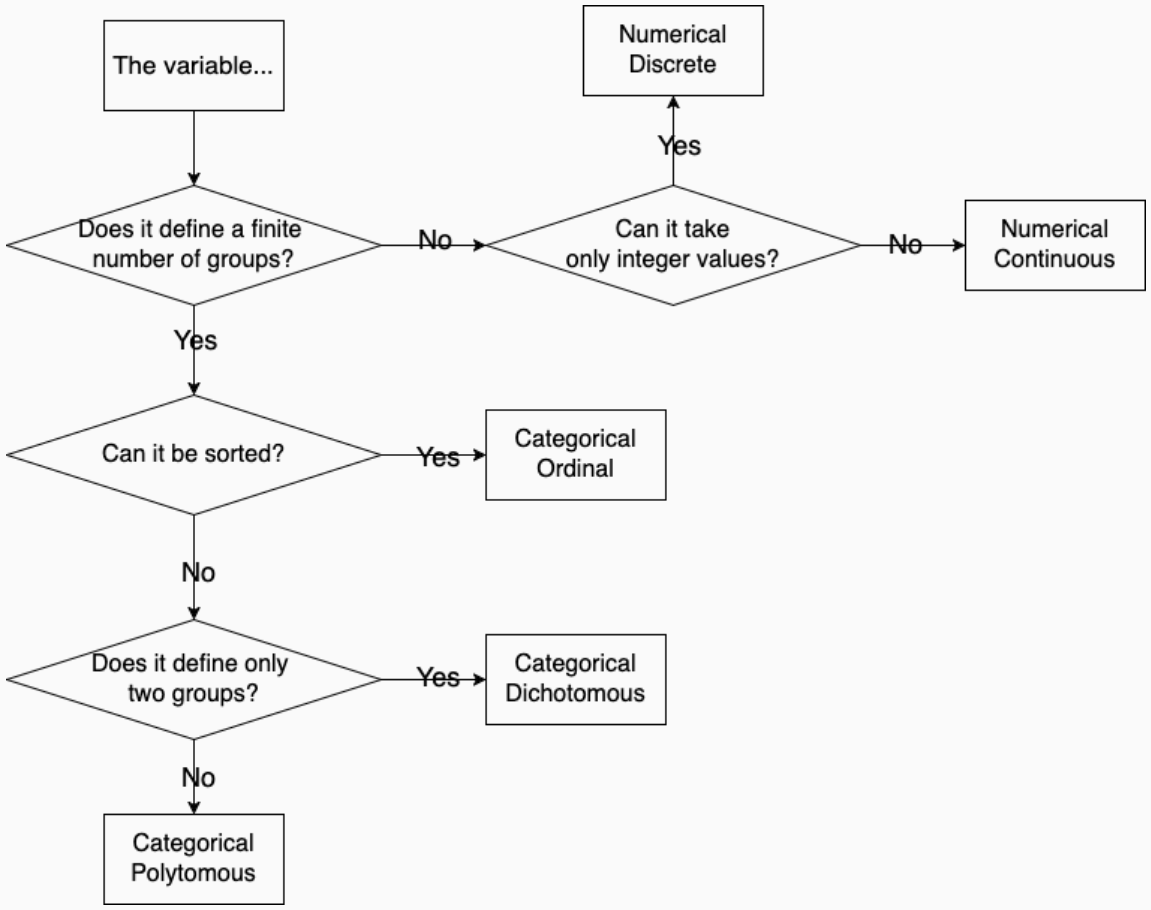
What type of data is this?

? One's nationality



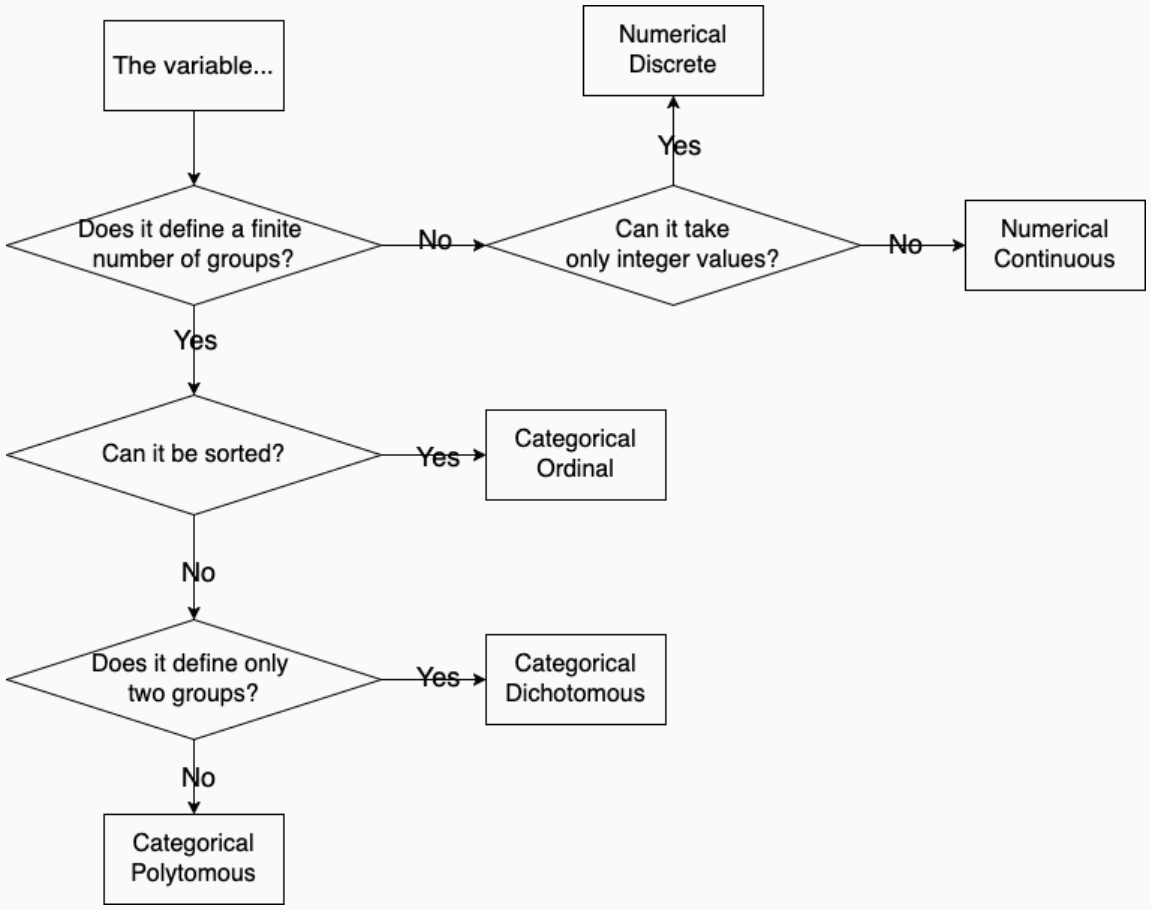
What type of data is this?

? Fasting blood sugar levels



What type of data is this?

? Passing the driving licence test



Exercise #4

? Which types of data are included in this table?

02:00

Visconti A., *et al.*, Total serum *N*-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Table 1 Patient characteristics.

	All cohorts
N (pre-treatment)	88
N (follow-up)	66
Sex	
Male	57 (64.8%)
Female	31 (35.2%)
Age (years)	60.5 ± 15.0
BMI (kg/m^2)	28.0 ± 5.4
***BRAF* mutant**	40 (45.5%)
LDH (≤ULN)	58 (65.9%)
Metastatic stage	
Stage III unresectable	2 (2.3%)
M1a	14 (15.9%)
M1b	17 (19.3%)
M1c	32 (36.4%)
M1d	23 (26.1%)
ECOG performance status	
0	47 (53.4%)
1	31 (35.2%)
2	8 (9.1%)
3	2 (2.3%)
ICI therapy	
Ipilimumab	1 (1.1%)
Pembrolizumab	20 (22.7%)
Nivolumab	30 (34.1%)
Ipilimumab + Nivolumab	37 (42.0%)

Exercise #4 -- Solution

? Which types of data are included in this table?

Table 1 Patient characteristics.

	All cohorts
N (pre-treatment)	88
N (follow-up)	66
Sex Categorical Dichotomous	
Male	57 (64.8%)
Female	31 (35.2%)
Age (years) Numerical Continuous	60.5 ± 15.0
BMI (kg/m²) Numerical Continuous	28.0 ± 5.4
***BRAF* mutant** Categorical Dichotomous	40 (45.5%)
LDH (≤ULN) Categorical Dichotomous	58 (65.9%)
Metastatic stage Categorical ordinal	
Stage III unresectable	2 (2.3%)
M1a	14 (15.9%)
M1b	17 (19.3%)
M1c	32 (36.4%)
M1d	23 (26.1%)
ECOG performance status Categorical ordinal	
0	47 (53.4%)
1	31 (35.2%)
2	8 (9.1%)
3	2 (2.3%)
ICI therapy Categorical polytomous	
Ipilimumab	1 (1.1%)
Pembrolizumab	20 (22.7%)
Nivolumab	30 (34.1%)
Ipilimumab + Nivolumab	37 (42.0%)

Visconti A., *et al.*, Total serum *N*-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Why is it important?

Table 1 Patient characteristics. Categorical variables are presented as number (percentage). Continuous variables are presented as mean ± standard deviation.

Visconti A., *et al.*, Total serum *N*-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Table 1 Patient characteristics.

	All cohorts
N (pre-treatment)	88
N (follow-up)	66
Sex	
Male	57 (64.8%)
Female	31 (35.2%)
Age (years)	60.5 ± 15.0
BMI (kg/m^2)	28.0 ± 5.4
***BRAF* mutant**	40 (45.5%)
LDH (≤ULN)	58 (65.9%)
Metastatic stage	
Stage III unresectable	2 (2.3%)
M1a	14 (15.9%)
M1b	17 (19.3%)
M1c	32 (36.4%)
M1d	23 (26.1%)
ECOG performance status	
0	47 (53.4%)
1	31 (35.2%)
2	8 (9.1%)
3	2 (2.3%)
ICI therapy	
Ipilimumab	1 (1.1%)
Pembrolizumab	20 (22.7%)
Nivolumab	30 (34.1%)
Ipilimumab + Nivolumab	37 (42.0%)

Categorical variables

Frequency table

- absolute frequency (the number)
- relative frequency (the percentage)

Table 1 Patient characteristics.

	All cohorts
N (pre-treatment)	88
Sex	
Male	57 (64.8%)
Female	31 (35.2%)
Metastatic stage	
Stage III unresectable	2 (2.3%)
M1a	14 (15.9%)
M1b	17 (19.3%)
M1c	32 (36.4%)
M1d	23 (26.1%)
ECOG performance status	
0	47 (53.4%)
1	31 (35.2%)
2	8 (9.1%)
3	2 (2.3%)
ICI therapy	
Ipilimumab	1 (1.1%)
Pembrolizumab	20 (22.7%)
Nivolumab	30 (34.1%)
Ipilimumab + Nivolumab	37 (42.0%)

Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Exercise #5

? Which sex is more frequent?

Table 1 Patient characteristics.

	All cohorts
N (pre-treatment)	88
Sex	
Male	57 (64.8%)
Female	31 (35.2%)

Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

00:30

Exercise #5 -- Solution

? Which sex is more frequent?

Table 1 Patient characteristics.

	All cohorts
N (pre-treatment)	88
Sex	
Male	57 (64.8%)
Female	31 (35.2%)

Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Contingency table

Study Objectives: Persistent insomnia, although very common in general practice, often proves problematic to manage. This study investigates the clinical effectiveness and the feasibility of applying cognitive behavior therapy (CBT) methods for insomnia in primary care.
Design: Pragmatic randomized controlled trial of CBT versus treatment as usual.
Setting: General medical practice.
Participants: Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

Espie. CA, *et al.*, *Randomized Clinical Effectiveness Trial of Nurse-Administered Small-Group Cognitive Behavior Therapy for Persistent Insomnia in General Practice*, Sleep, 2007, doi:10.1186/s12885-023-10511-3

Exercise #6

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males			
Females			
Total			

05:00

Exercise #6 -- Solution

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males			
Females	72	65	
Total	107	94	201

Exercise #6 -- Solution

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males	35	29	
Females	72	65	
Total	107	94	201

Exercise #6 -- Solution

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males	35	29	64
Females	72	65	137
Total	107	94	201

We've just filled a contingency table (using the absolute frequencies)

Exercise #6 -- Solution

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males	35 (35/107)	29 (29/94)	64 (64/201)
Females	72 (72/107)	65 (65/94)	137 (64/201)
Total	107	94	201

Option 1: let's divide "by columns", that is, calculate the percentage of males and females in each arm of the experiment

Exercise #6 -- Solution

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males	35 (32.7%)	29 (30.9%)	64 (31.8%)
Females	72 (67.3%)	65 (69.1%)	137 (68.2%)
Total	107	94	201

Option 1: let's divide "by columns", that is, calculate the percentage of males and females in each arm of the experiment

Exercise #6 -- Solution

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males	35 (35/64)	29 (29/64)	64
Females	72 (72/137)	65 (65/137)	137
Total	107 (107/201)	94 (94/201)	201

Option 2: let's divide "by rows", that is, calculate the percentage of subjects in each arm of the experiments that is either male or female

Exercise #6 -- Solution

? Complete the table with the correct absolute and relative frequencies using the information contained in the abstract

> **Participants:** Two hundred one adults (mean age, 54 years) randomly assigned to receive CBT (n = 107; 72 women) or treatment as usual (n = 94; 65 women).

	CBT	Standard	Total
Males	35 (54.7%)	29 (45.3%)	64
Females	72 (52.6%)	65 (47.4%)	137
Total	107 (53.2%)	94 (46.8%)	201

Option 2: let's divide "by rows", that is, calculate the percentage of subjects in each arm of the experiments that is either male or female

Exercise #7

? Is the proportion of men and women similar?

a) True b) False c) I need more elements to decide

? Comparing the two arms, is the proportion of men and women similar?

a) True b) False c) I need more elements to decide

	CBT	Standard	Total
Males	35 (32.7%)	29 (30.9%)	64 (31.8%)
Females	72 (67.3%)	65 (69.1%)	137 (68.2%)
Total	107	94	201

01:00

Exercise #7 -- Solution

❓ Is the proportion of men and women similar?

a) True b) False ✅ c) I need more elements to decide

❓ Comparing the two arms, is the proportion of men and women similar?

a) True b) False c) I need more elements to decide

	CBT	Standard	Total
Males	35 (32.7%)	29 (30.9%)	64 (31.8%)
Females	72 (67.3%)	65 (69.1%)	137 (68.2%1)
Total	107	94	201

Exercise #7 -- Solution

❓ Is the proportion of men and women similar?

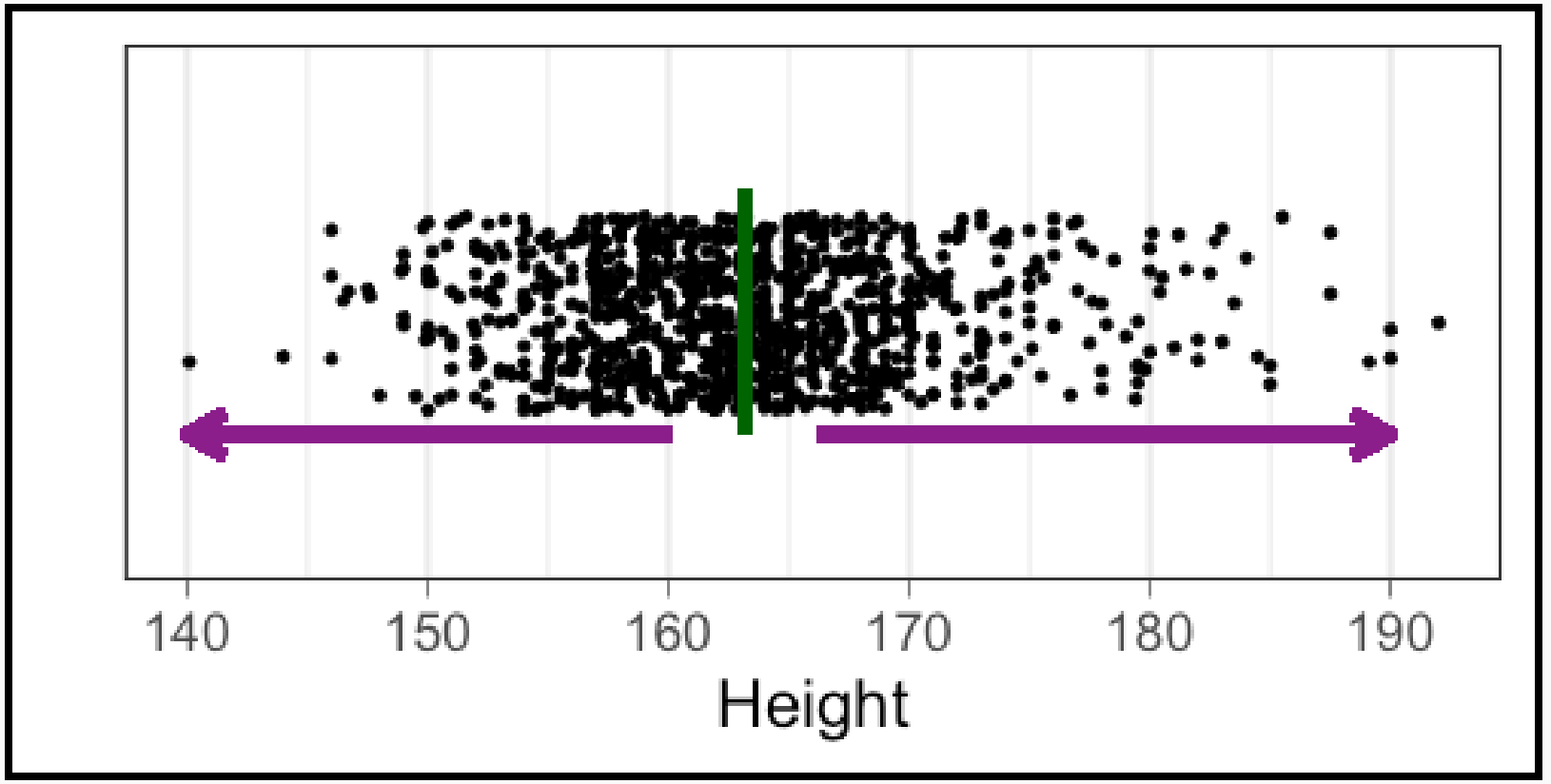
a) True b) False ✅ c) I need more elements to decide

❓ Comparing the two arms, is the proportion of men and women similar?

a) True ✅ b) False c) I need more elements to decide

	CBT	Standard	Total
Males	35 (32.7%)	29 (30.9%)	64 (31.8%)
Females	72 (67.3%)	65 (69.1%)	137 (68.2%1)
Totale	107	94	201

Measures of central tendency and dispersion



Measure of central tendency: mode

🎯 The most frequent item

📌 $x = \{1, 1, 1, 3, 4, 4, 7, 8, 8, 9, 9\}$

$\text{mode}(x) = 1$

Exercise #8

? Calculate the mode of the following data sets

$$y = \{1, 1, 1, 3, 4, 4, 4, 7, 8, 8, 9, 9\}$$
$$\text{mode}(y) = \ ?$$

$$z = \{1, 3, 4, 7, 8, 9, 11, 17, 21, 42\}$$
$$\text{mode}(z) = \ ?$$

01:00

Exercise #8 -- Solution

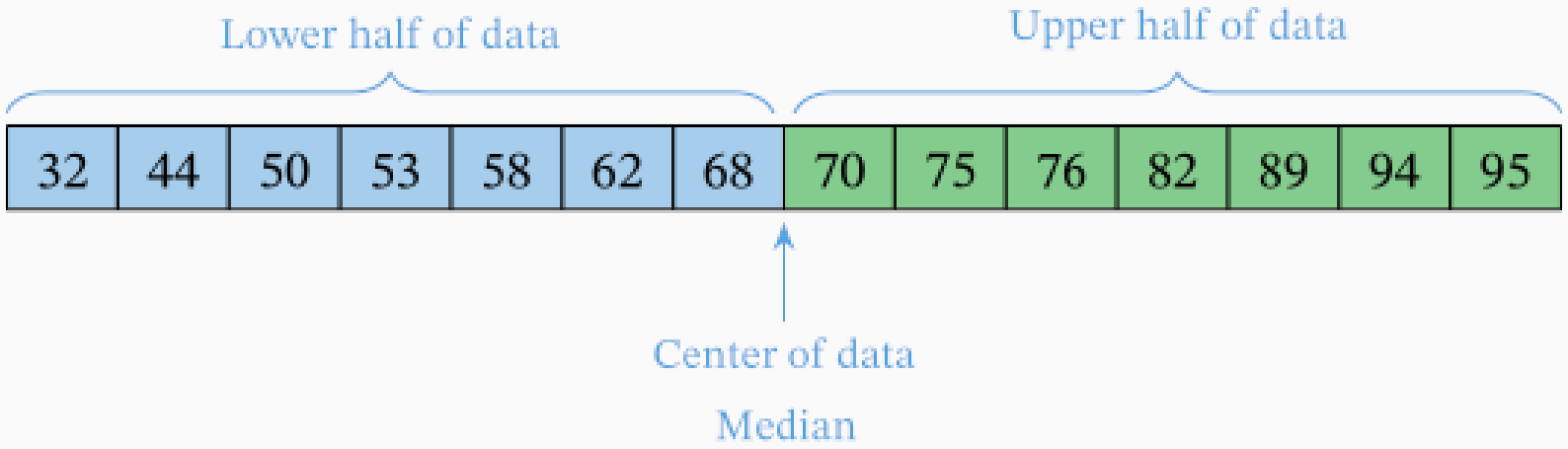
? Calculate the mode of the following data sets

$$y = \{1, 1, 1, 3, 4, 4, 4, 7, 8, 8, 9, 9\}$$
$$\text{mode}(y) = 1 \ \& \ 4$$

$$z = \{1, 3, 4, 7, 8, 9, 11, 17, 21, 42\}$$
$$\text{mode}(z) = \text{It doesn't exist}$$

Measure of central tendency: median

🎯 The "middle" value



⚠️ Data should be sorted!

Measure of central tendency: median

🎯 The "middle" value

📌 $n = 7, x = \{1, 3, 3, 6, 7, 8, 9\}$

$$\text{median}(x) = x_{(n+1)/2} = x_{(7+1)/2} = x_4 = 6$$

📌 $n = 8, x = \{1, 2, 3, 4, 5, 6, 8, 9\}$

$$\text{median}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2} = \frac{x_{(8/2)} + x_{((8/2)+1)}}{2} = \frac{x_4 + x_5}{2} = \frac{4+5}{2} = 4.5$$

⚠️ Data should be sorted!

Exercise #9

? Calculate the median of the following data sets

$$y = \{6, 34, 40, 55, 75\}$$
$$\text{median}(y) = \ ?$$

$$z = \{6, 34, 40, 55, 175\}$$
$$\text{median}(z) = \ ?$$

01:00

Exercise #9 -- Solution

❓ Calculate the median of the following data sets

$$y = \{6, 34, 40, 55, 75\}$$
$$\text{median}(y) = y_3 = 40$$

$$z = \{6, 34, 40, 55, 175\}$$
$$\text{median}(z) = \ ?$$

⚠️ Data should be sorted!

Exercise #9 -- Solution

❓ Calculate the median of the following data sets

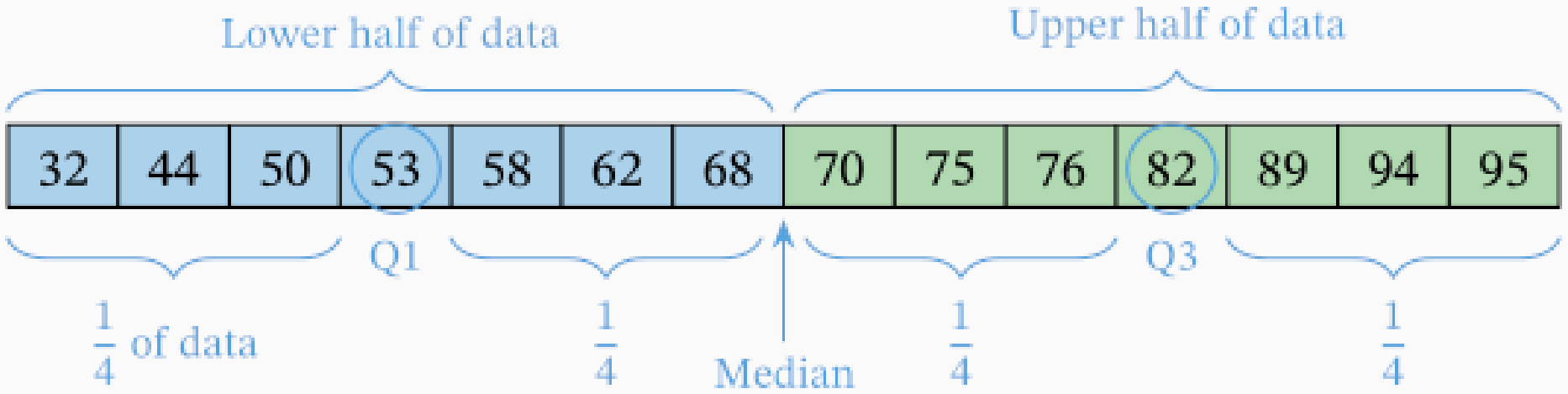
$$y = \{6, 34, 40, 55, 75\}$$
$$\text{median}(y) = y_3 = 40$$

$$z = \{6, 34, 40, 55, 175\}$$
$$\text{median}(z) = z_3 = 40$$

⚠️ Data should be sorted!

🎯 Robust to outliers

Quartiles



⚠️ Data should be sorted!

Measure of central tendency: mean

🎯 Arithmetic mean

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

📌 $x = \{4, 36, 45, 50, 75\}$

$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{4+36+45+50+75}{5} = 42$

Exercise #10

? Calculate the mean of the following data sets

$$y = \{6, 34, 40, 55, 75\}$$
$$\bar{y} = \ ?$$

$$z = \{6, 34, 40, 55, 175\}$$
$$\bar{z} = \ ?$$

02:00

Exercise #10 -- Solution

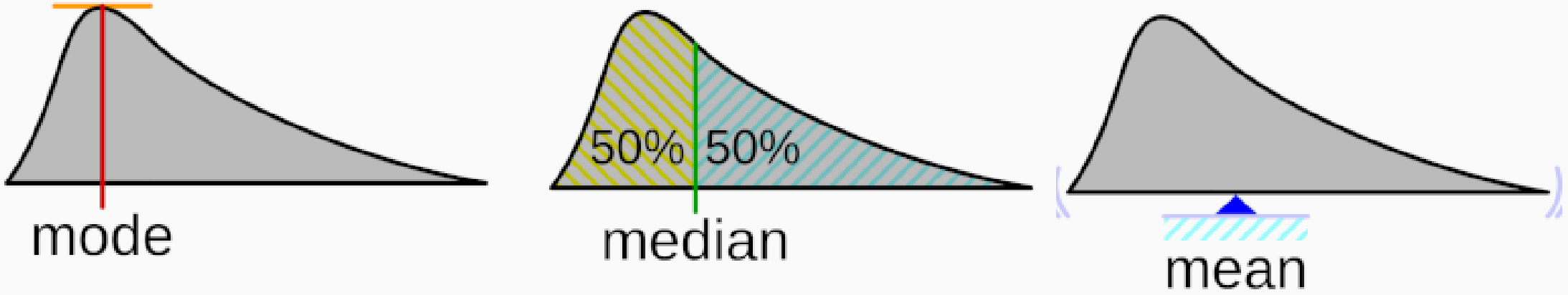
❓ Calculate the mean of the following data sets

$$y = \{6, 34, 40, 55, 75\}$$
$$\bar{y} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) = \frac{6+34+40+55+75}{5} = 42$$

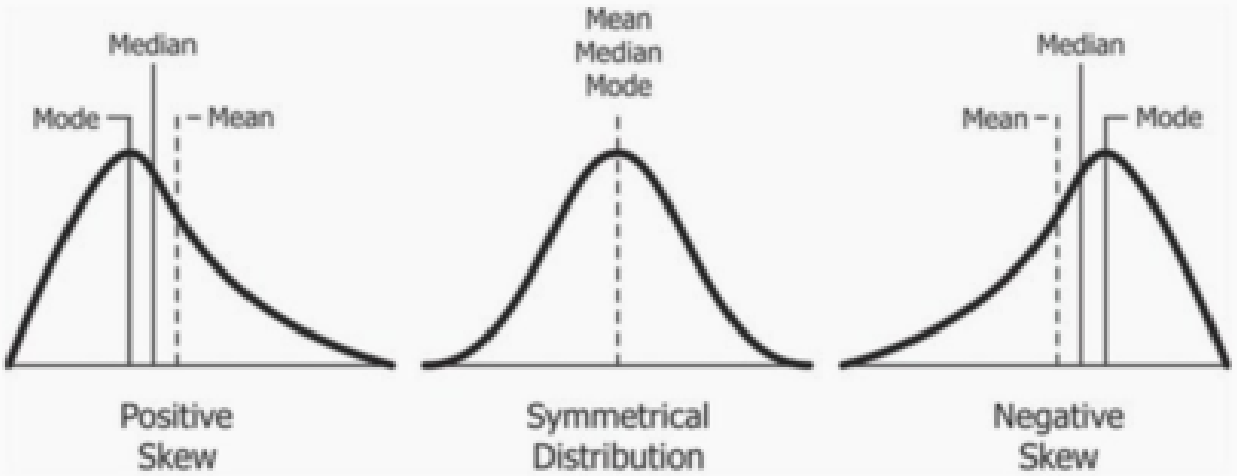
$$z = \{6, 34, 40, 55, 175\}$$
$$\bar{z} = \frac{1}{n}\left(\sum_{i=1}^{n} z_i\right) = \frac{4+36+45+50+175}{5} = 62$$

🎯 Sensitive to outliers

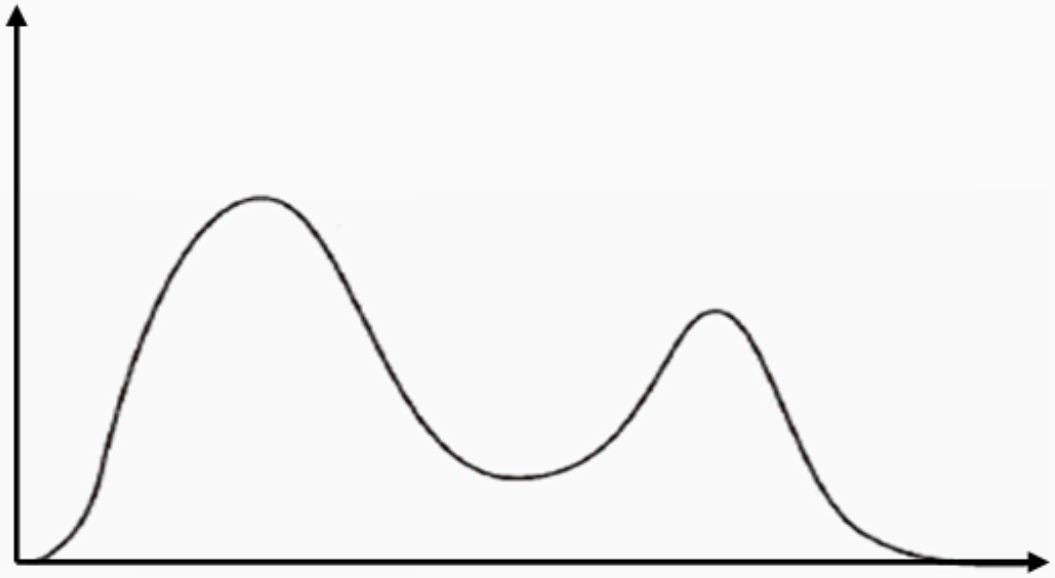
Mode *vs* median *vs* mean



The shape of a distribution



The shape of a distribution



Exercise 11

? In the results section, the authors reported the following

The mean length of stay was 22.4 days (median: 14 days).

The shape of the distribution is...

a) symmetric
b) positive skewed
c) negative skewed
d) I need more information to answer

01:00

Exercise 11 -- Solution

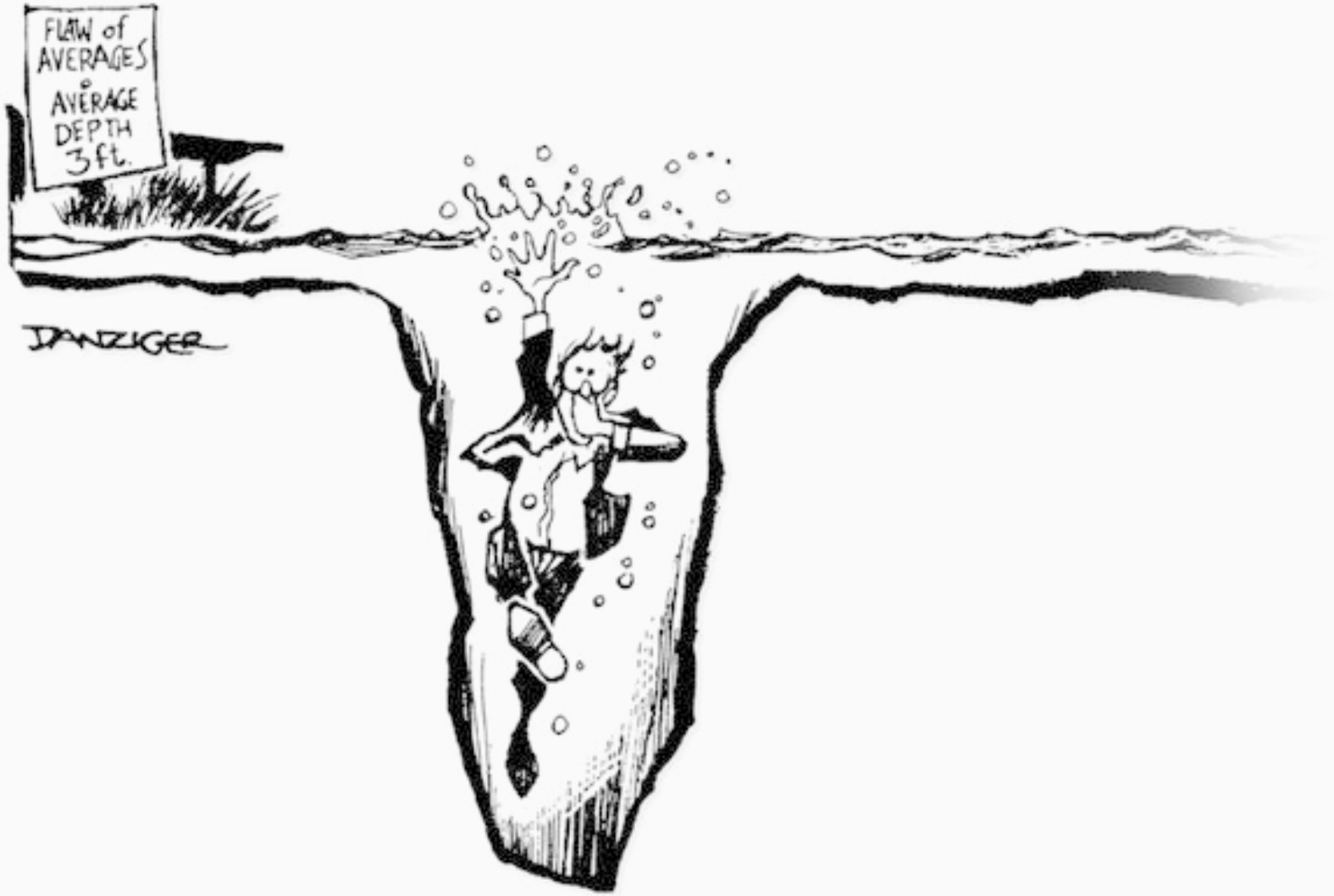
❓ In the results section, the authors reported the following

The mean length of stay was 22.4 days (median: 14 days).

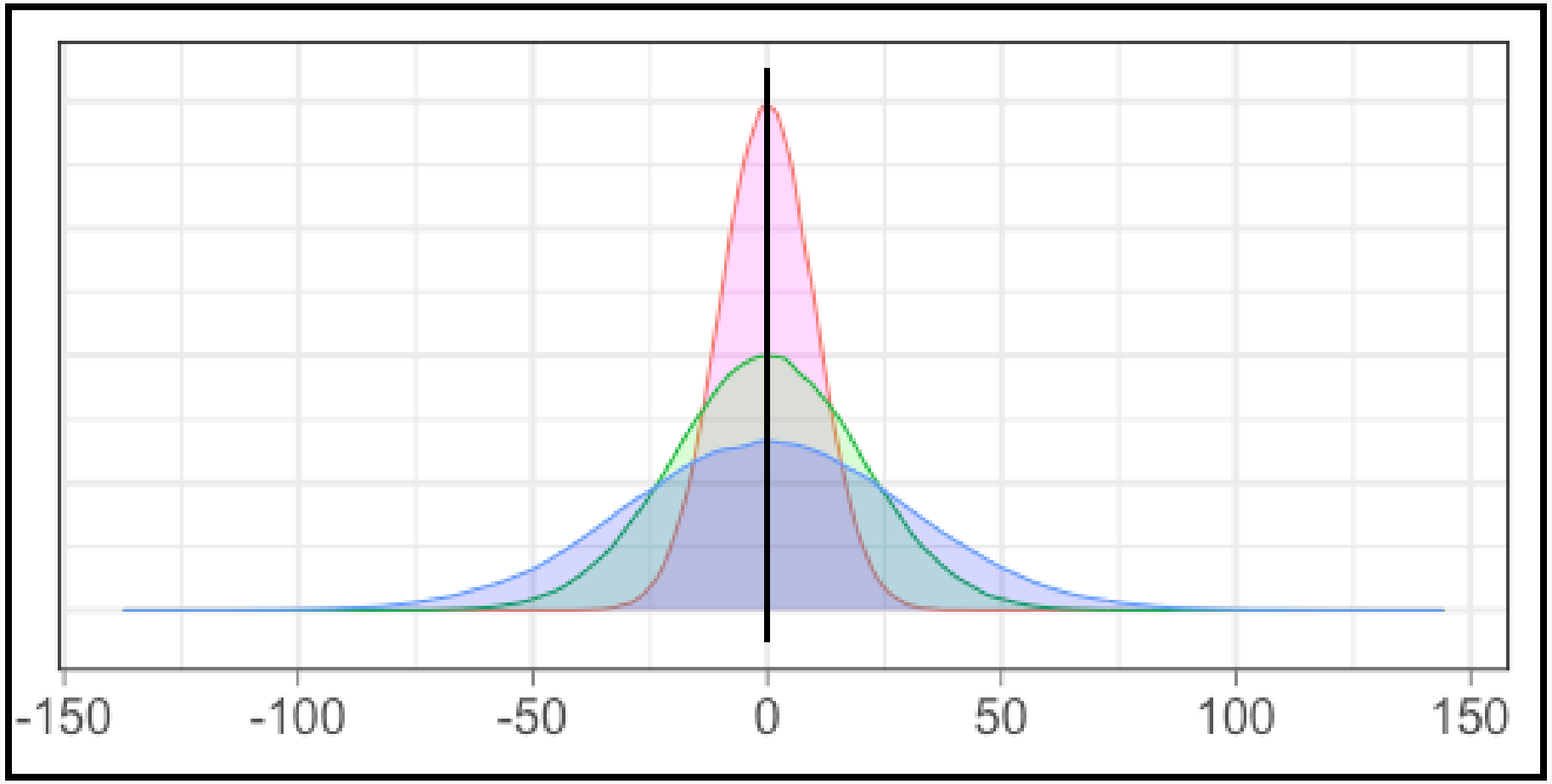
The shape of the distribution is...

a) symmetric
b) positive skewed ✅
c) negative skewed
d) I need more information to answer

Measures of dispersion



Measures of dispersion

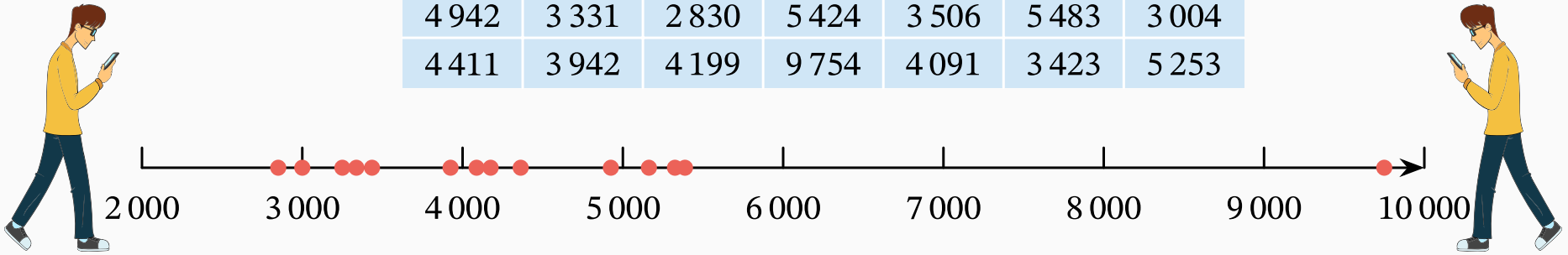


Measure of dispersion: range

🎯 $\text{range}(x) = \max(x) - \min(x)$

Number of Messages Exchanged in 1 Month

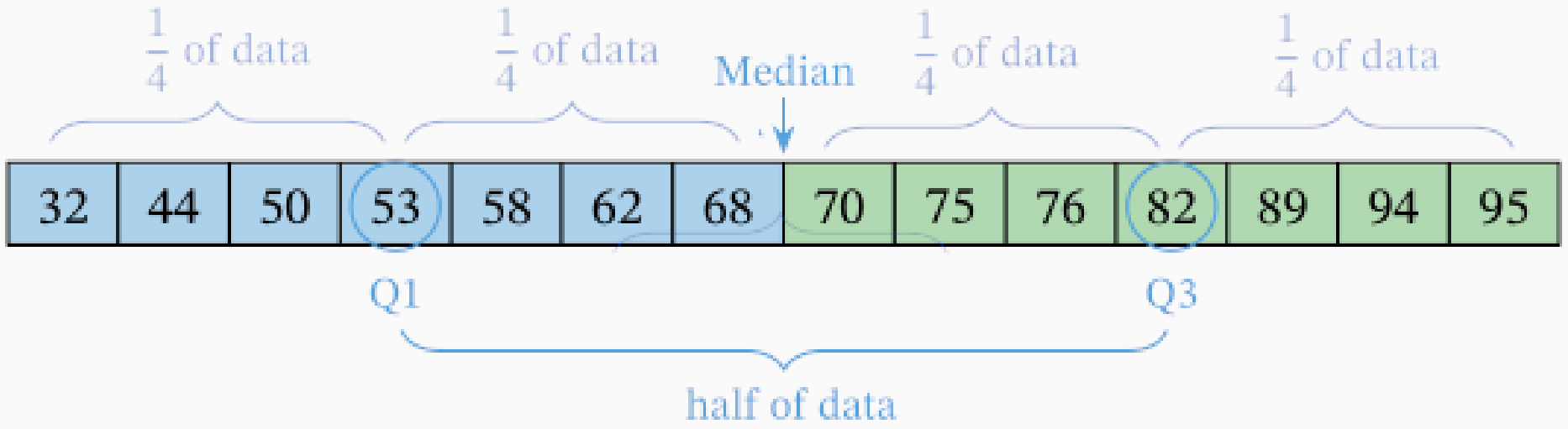
4 942	3 331	2 830	5 424	3 506	5 483	3 004
4 411	3 942	4 199	9 754	4 091	3 423	5 253



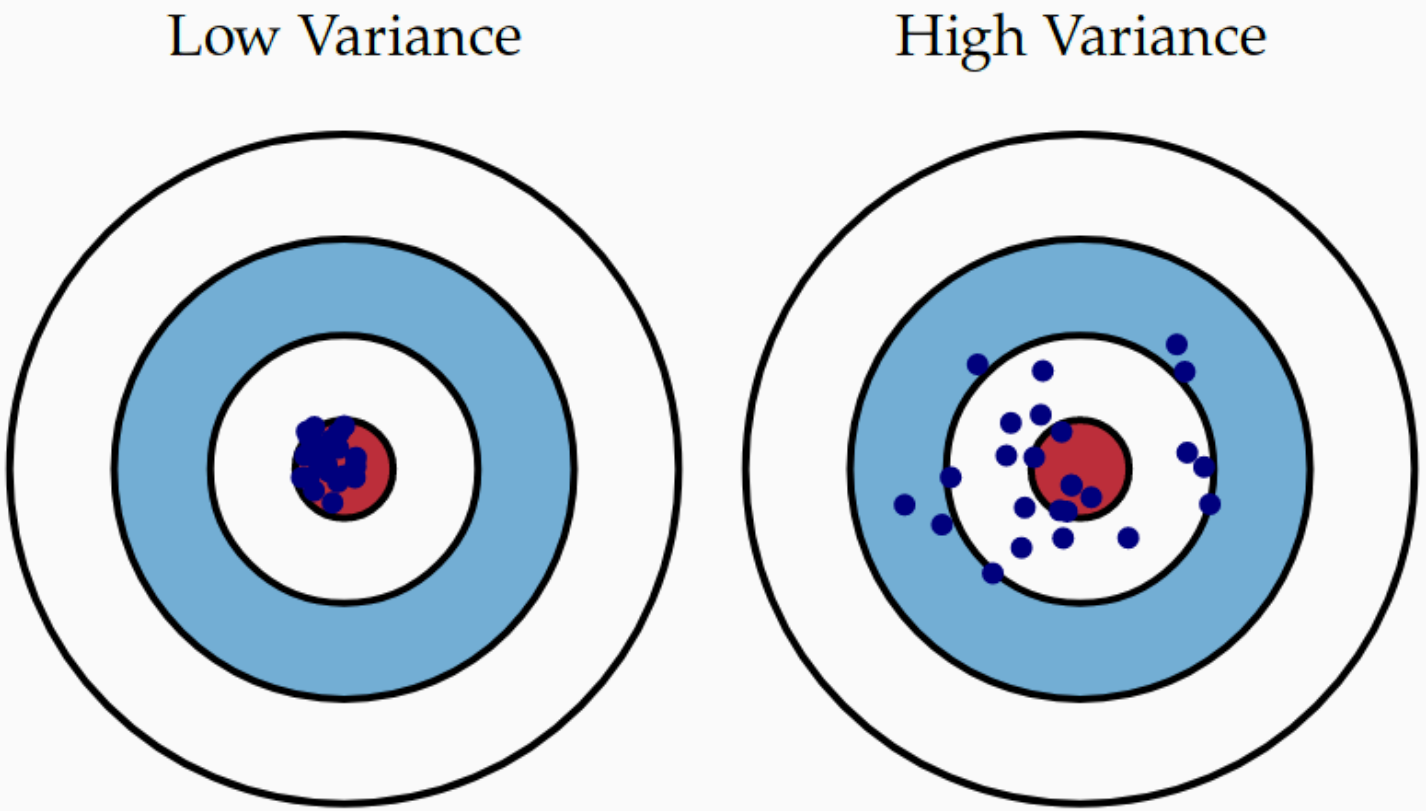
$$\text{range}(x) = \max(x) - \min(x) = 9,754 - 2,830 = 6,924$$

Measure of dispersion: interquartile range

🎯 $\mathrm{IQR}(x) = \mathrm{Q3}(x) - \mathrm{Q1}(x)$



Measure of dispersion: variance



Measure of dispersion: variance

🎯 $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

where $\bar{x} = \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)$

📌 $x = \{1, 2, 3\}$ $\bar{x} = \frac{1+2+3}{3} = 2$

$$s = \frac{1}{3-1} \times [(1-2)^2 + (2-2)^2 + (3-2)^2] =$$
$$= \frac{1}{2} \times [1^2 + 0^2 + 1^2] = \frac{1}{2} \times 2 = 1$$

Measure of dispersion: standard deviation

🎯 $$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

📌 $$x = \{1, 2, 3\} \qquad \bar{x} = \frac{1+2+3}{3} = 2$$

$$s = \sqrt{\frac{1}{3-1} \times [(1-2)^2 + (2-2)^2 + (3-2)^2]} =$$

$$= \sqrt{\frac{1}{2} \times [1^2 + 0^2 + 1^2]} = \sqrt{\frac{1}{2} \times 2} = \sqrt{1} = 1$$

Centrality, dispersion, and data types

Data type	Centrality Measure	Dispersion Measure
Nominal	Mode	-
Ordinal	Mode, Median	Range, IQR
Numeric	Mode, Median, Mean	Range, IQR, standard deviation

Exercise #12

? In the results section, the authors reported the following

Coronary-artery calcium scores averaged 68.9±244.2 (range 0 to 1526) in patients and 8.8±41.8 (range 0 to 243.4) in controls.

Which measures do we use to describe this variable?

a) mean and standard deviation
b) median and interquartile range
c) median and standard deviation
d) I need more information to answer

01:00

Exercise #12 -- Solution

❓ In the results section, the authors reported the following

Coronary-artery calcium scores averaged 68.9±244.2 (range 0 to 1526) in patients and 8.8±41.8 (range 0 to 243.4) in controls.

Which measures do we use to describe this variable?

a) mean and standard deviation
b) median and interquartile range ✅
c) median and standard deviation
d) I need more information to answer

Outliers in the wild

TABLE 3. Length of In-Patient Stay, by Surgical Procedure

Procedure	No. of procedures	Length of stay, d: Mean ± SD	Length of stay, d: Median (IQR)
Breast surgery	1,338	3.3 ± 4.4	3 (0-5)
Coronary artery bypass graft	570	9.6 ± 15.2	8 (7-9)
Cesarean section	4,831	4.9 ± 6.4	4 (3-5)
Repair of fractured neck of femur	2,303	13.8 ± 12.2	10 (7-17)
Hip replacement	6,432	8.7 ± 5.9	7 (6-9)
Abdominal hysterectomy	1,484	5.4 ± 4.0	5 (4-6)
Knee replacement	4,483	8.2 ± 5.0	7 (6-9)
Major vascular surgery	269	22.4 ± 23.1	14 (8-30)
Overall	21,710	7.8 ± 8.0	6 (4- 9)

The mean length of stay was 7.8 days but was greatly influenced by 2 patients with lengths of stay of almost 1 year. The median length of stay was 6 days, with 90% of patients discharged within 14 days after the procedure. Table 3 displays measures of central tendency (mean and median values) and dispersion (SDs and interquartile ranges) for the length of stay for each type of surgical procedure.

Reilly, J. *et al.*. *Procedure-Specific Surgical Site Infection Rates and Postdischarge Surveillance in Scotland*, Infection Control and Hospital Epidemiology, 2006, doi:10.1086/509839

Parameters *vs* statistics

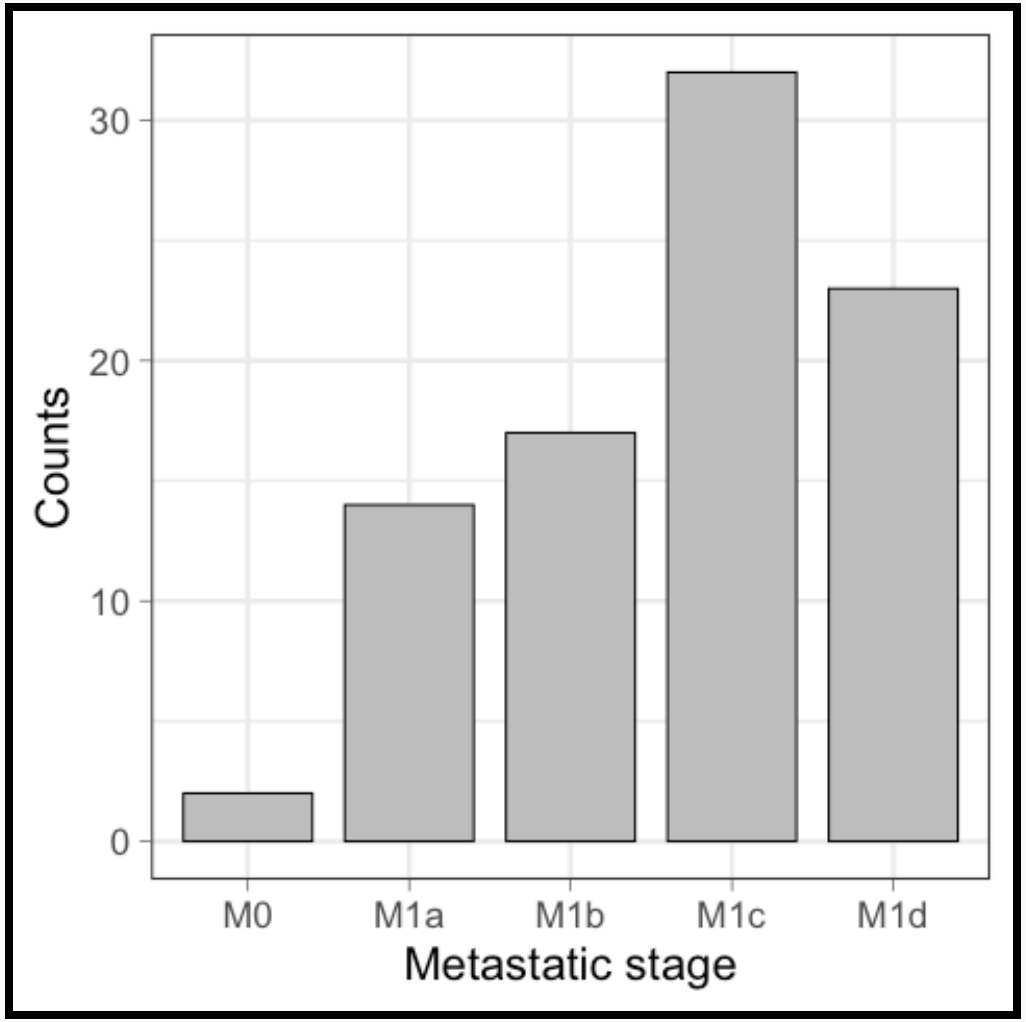
- Parameters: calculated on the population
- Statistics: calculated on the sample

	Parameter	Statistic
Size	N	n
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	π	p

Notes on data visualization

Bar chart

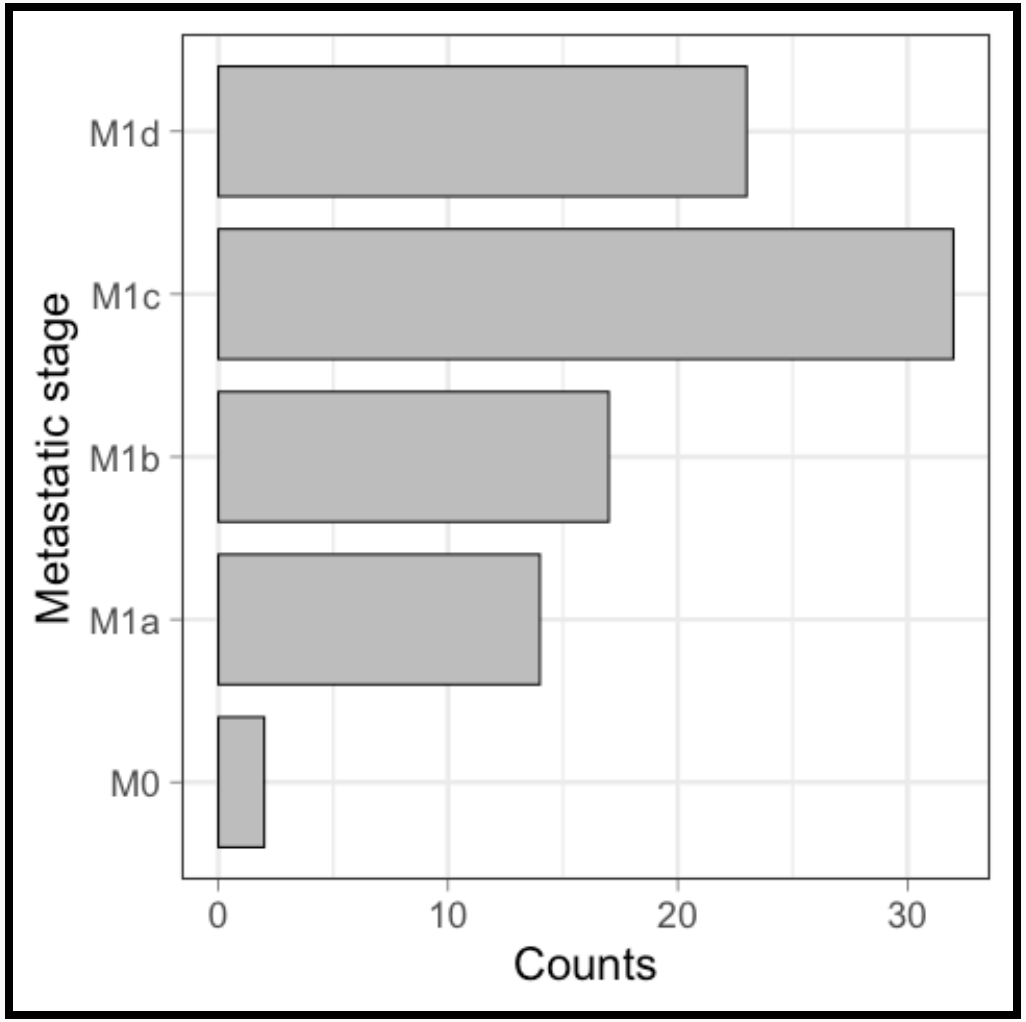
- categorical data
  - absolute frequency
  - relative frequency



Visconti A., *et al.*, Total serum *N*-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Horizontal bar chart

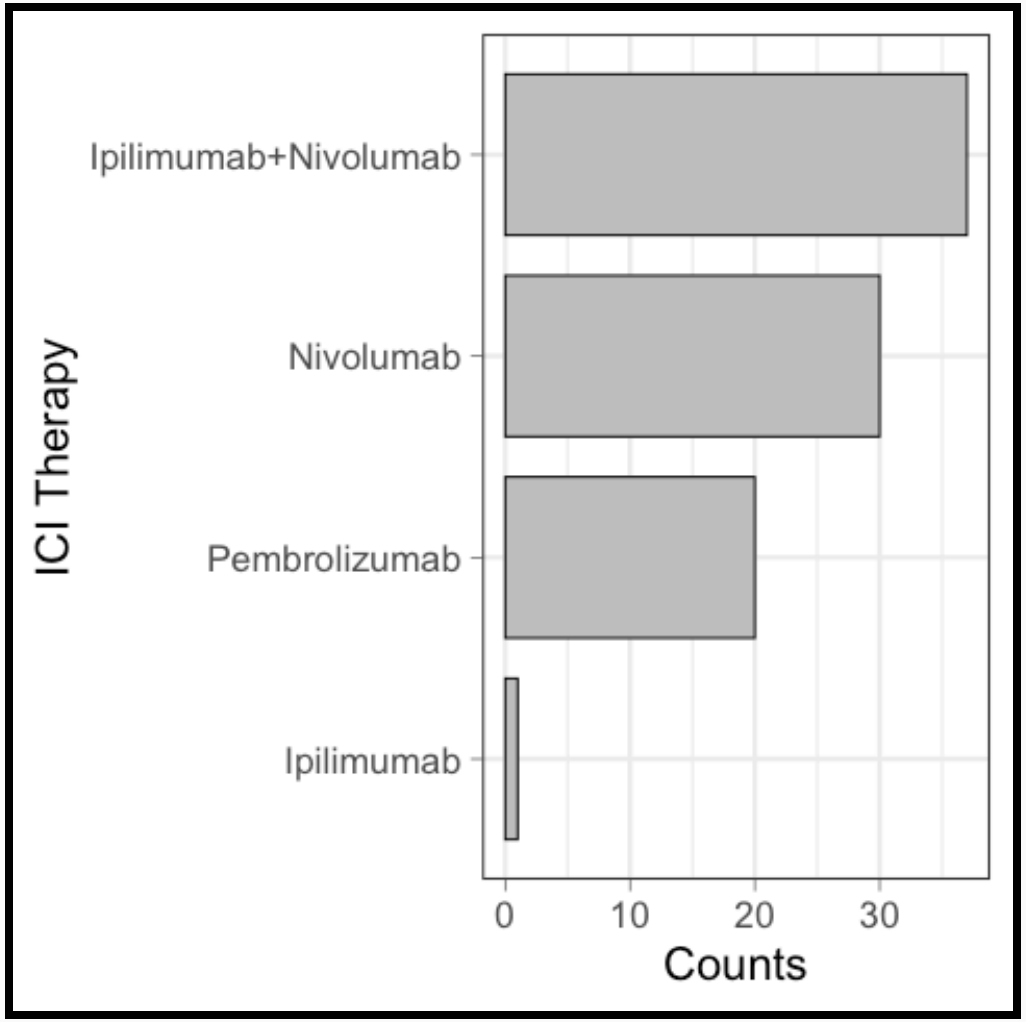
- categorical data
  - absolute frequency
  - relative frequency



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Horizontal bar chart

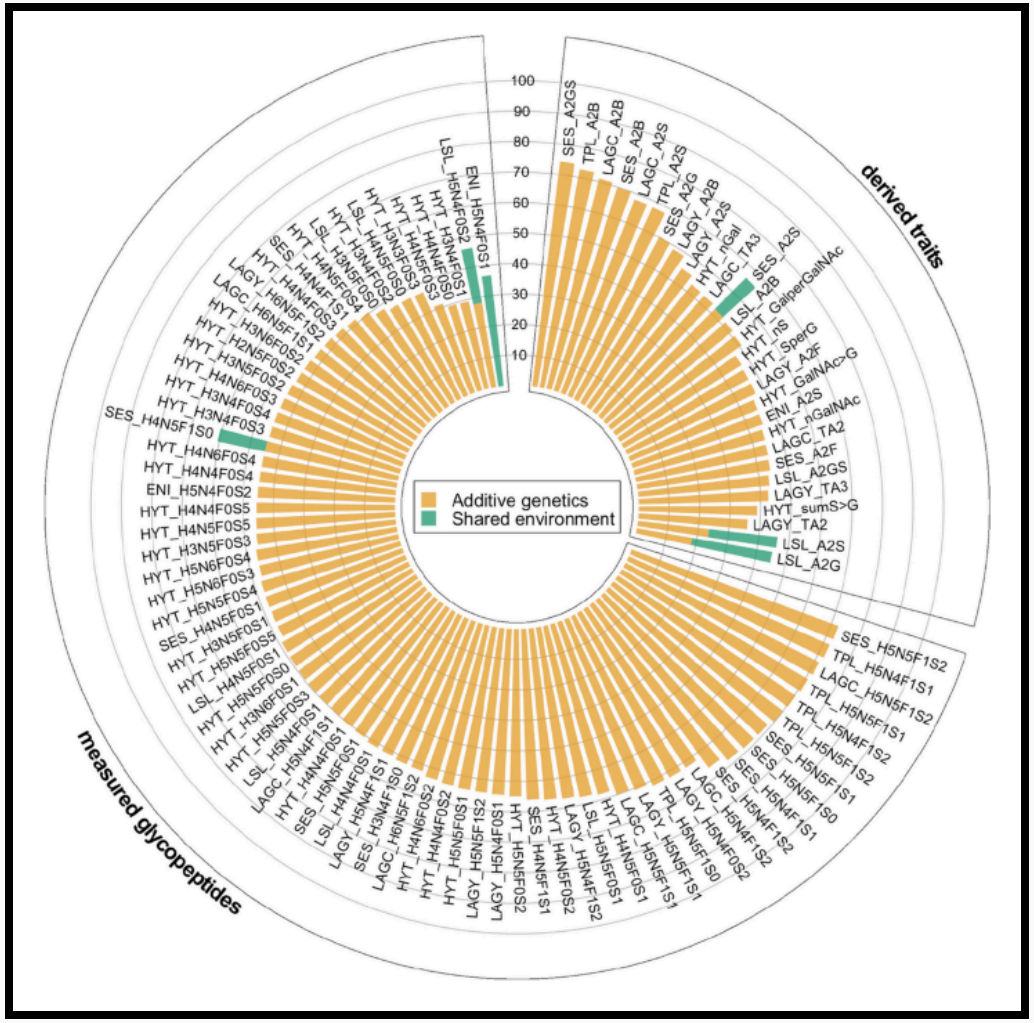
- categorical data
  - absolute frequency
  - relative frequency



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Circular bar chart

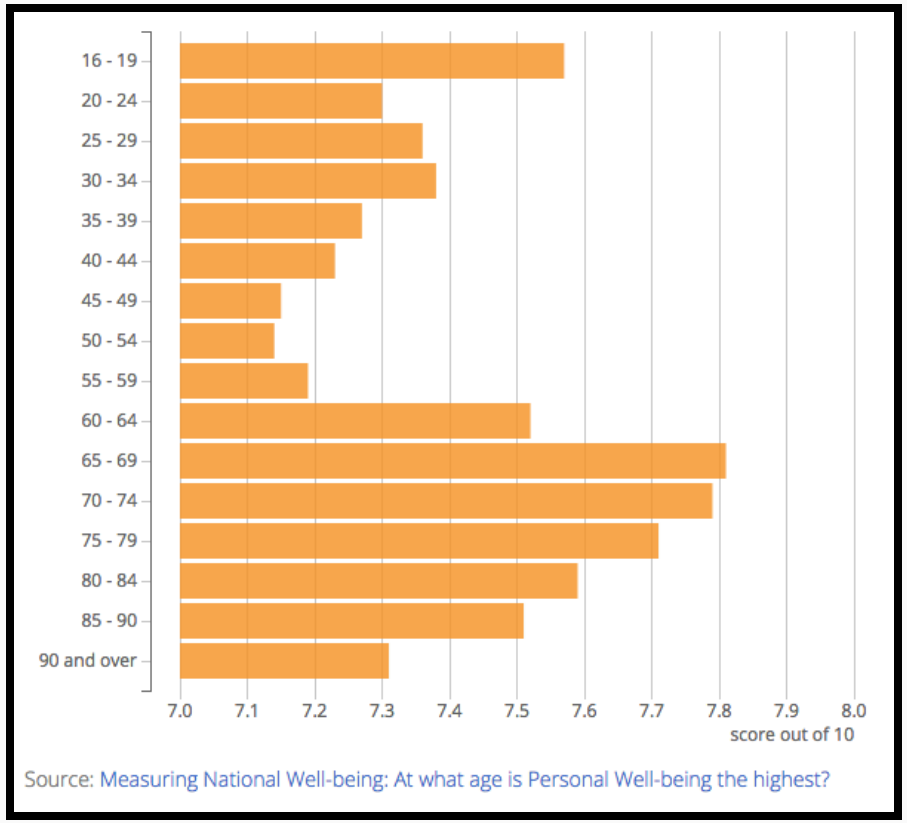
- categorical data
  - absolute frequency
  - relative frequency



Visconti, A., *et al.*. *The genetics and epidemiology of N-and O-immunoglobulin A glycomics.*, 2024, doi:10.1186/s13073-024-01369-6

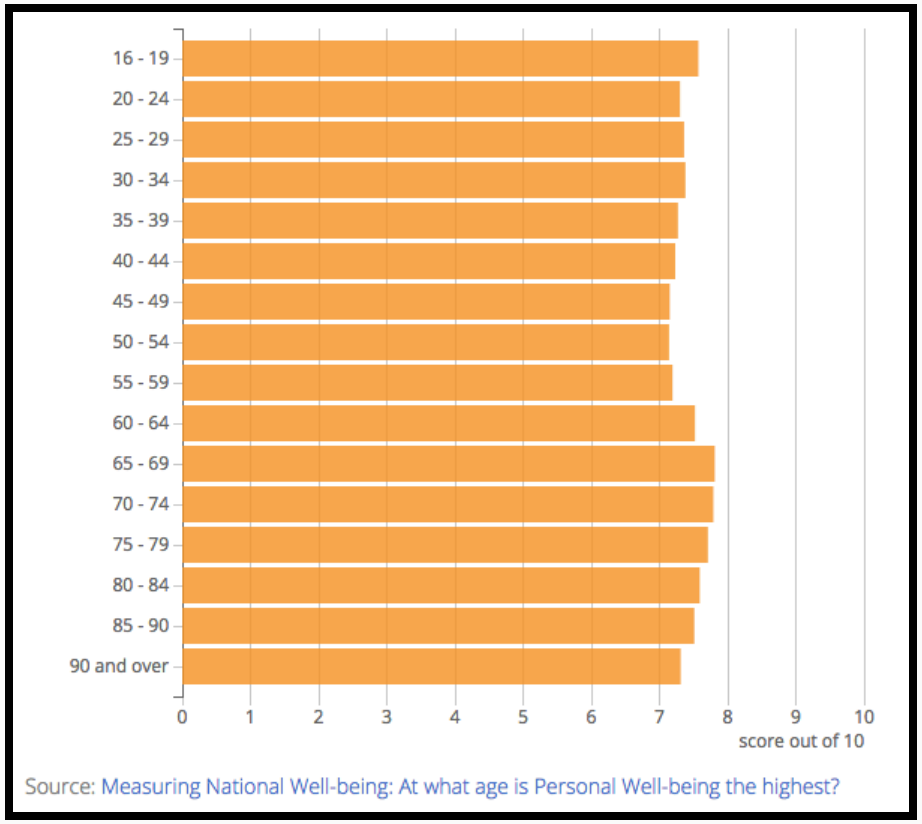
What does this graph tell us?

- From 1 to 10, how happy were you yesterday?



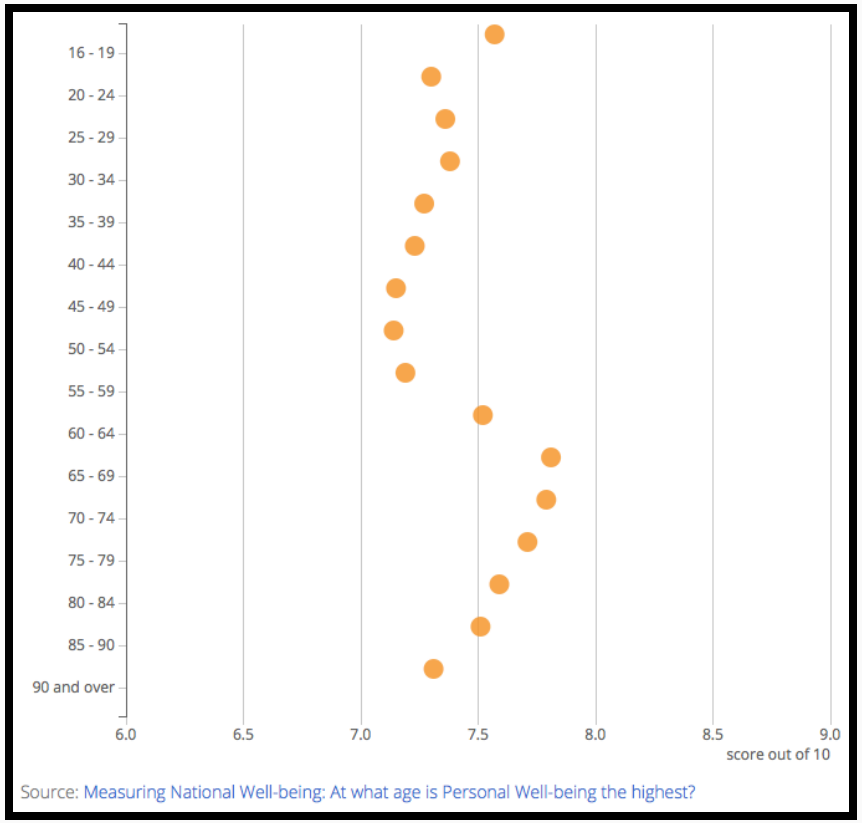
What does this graph tell us?

- From 1 to 10, how happy were you yesterday?



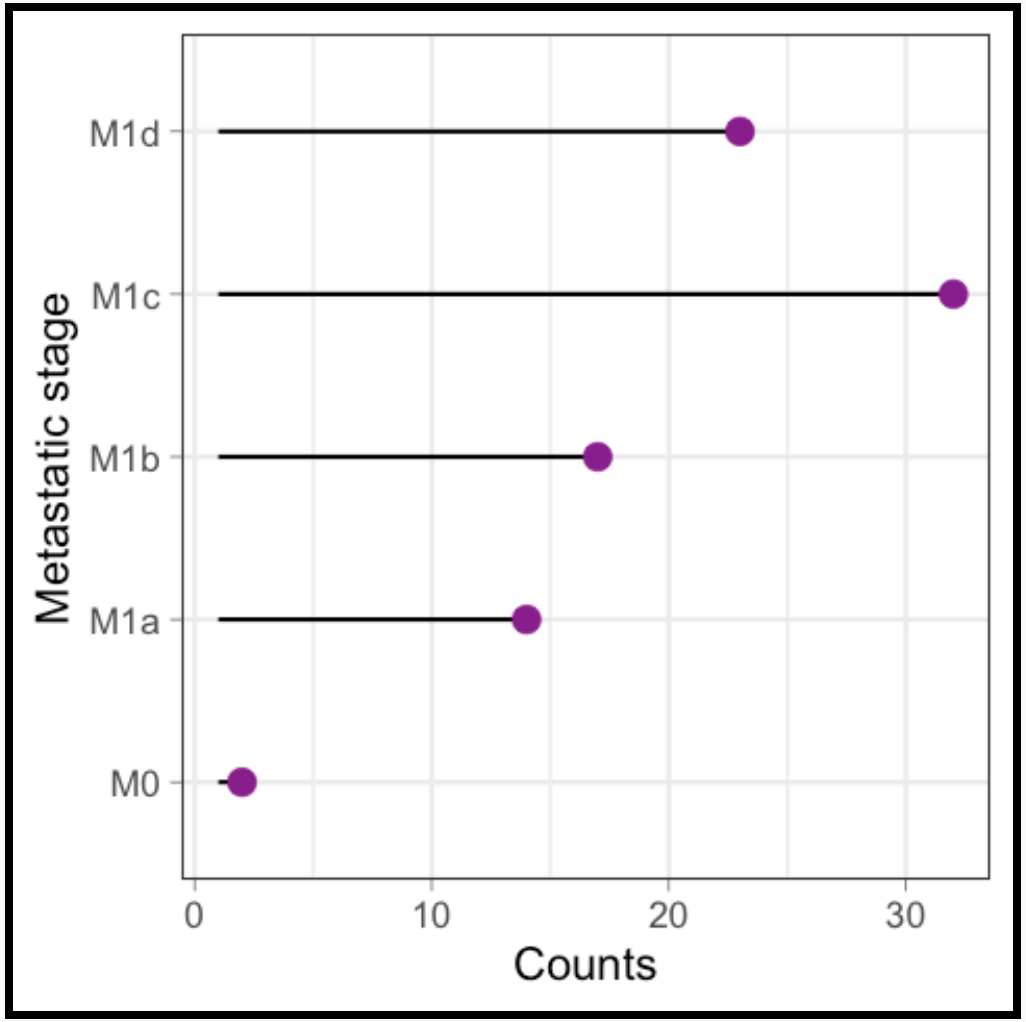
What does this graph tell us?

- From 1 to 10, how happy were you yesterday?



Lollipop chart

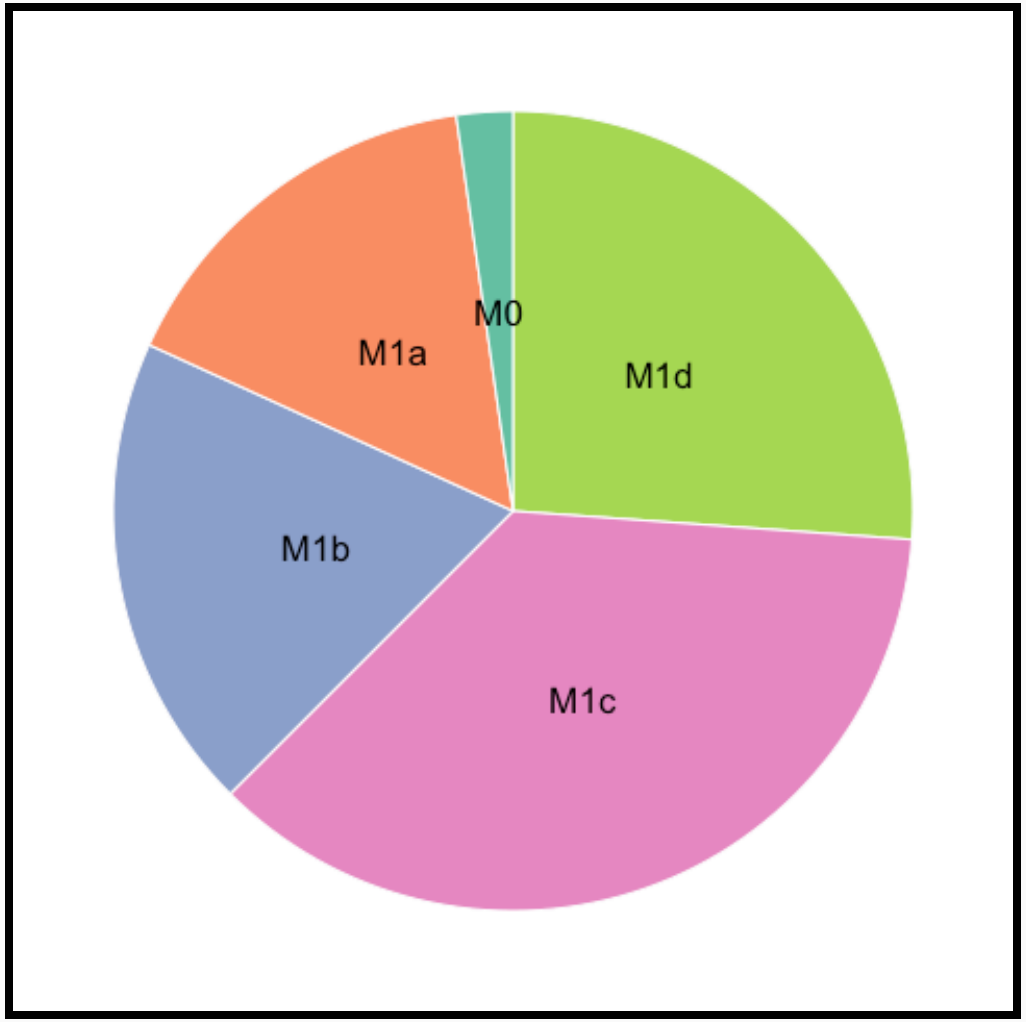
- categorical data
  - absolute frequency
  - relative frequency



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Pie chart

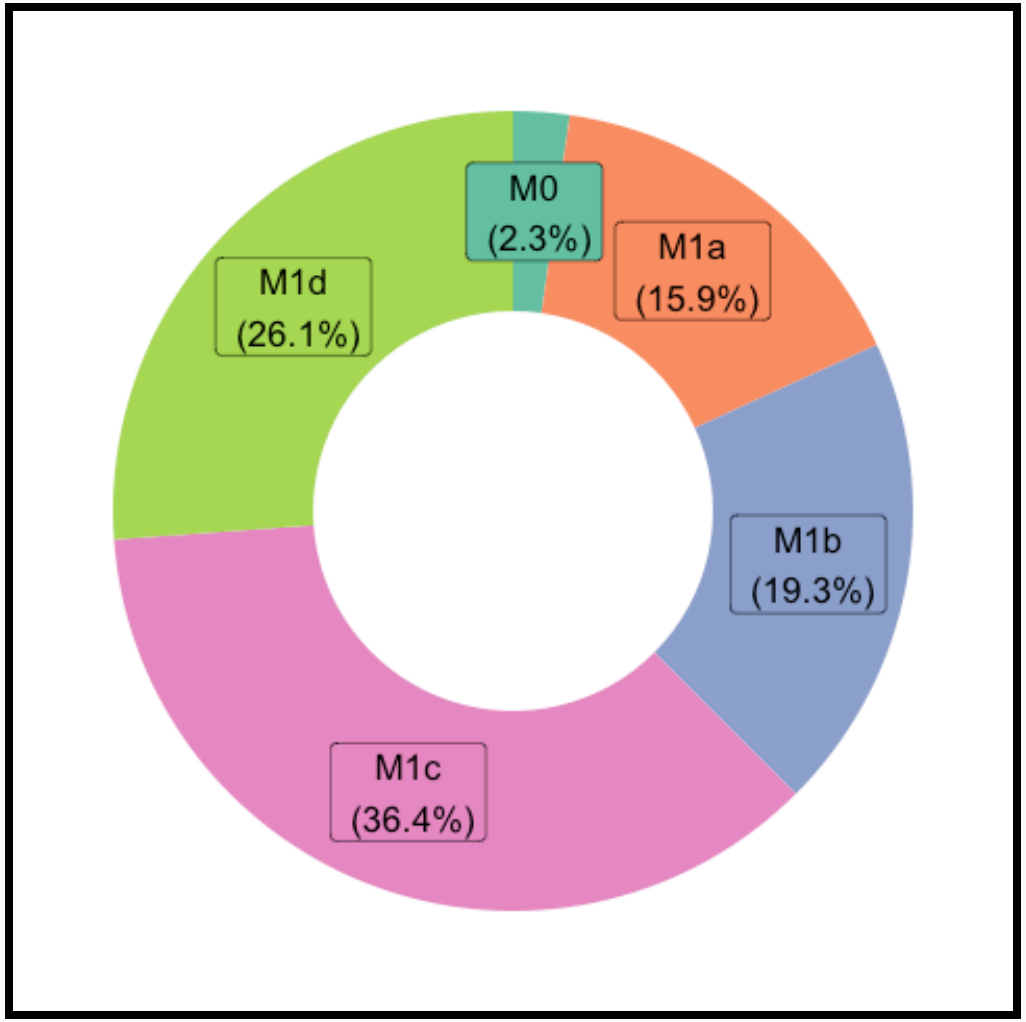
- categorical data
  - relative frequency



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Donut chart

- categorical data
  - relative frequency



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Waffle chart

- categorical data
  - relative frequency

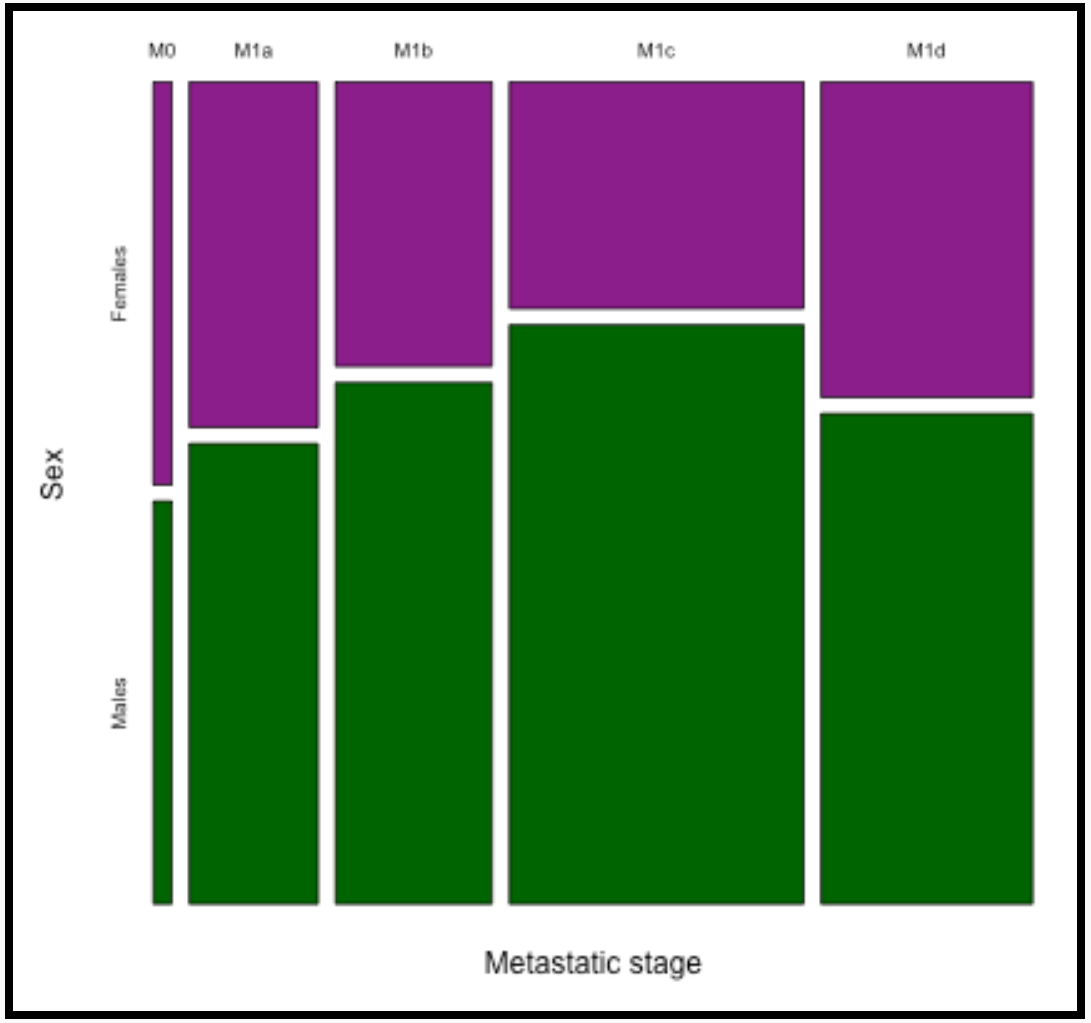
Visconti A., et al., Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Infographics

Spiegelhalter, D., *The Art of Statistics: Learning From Data*, Pelican, 2019

Mosaic plot

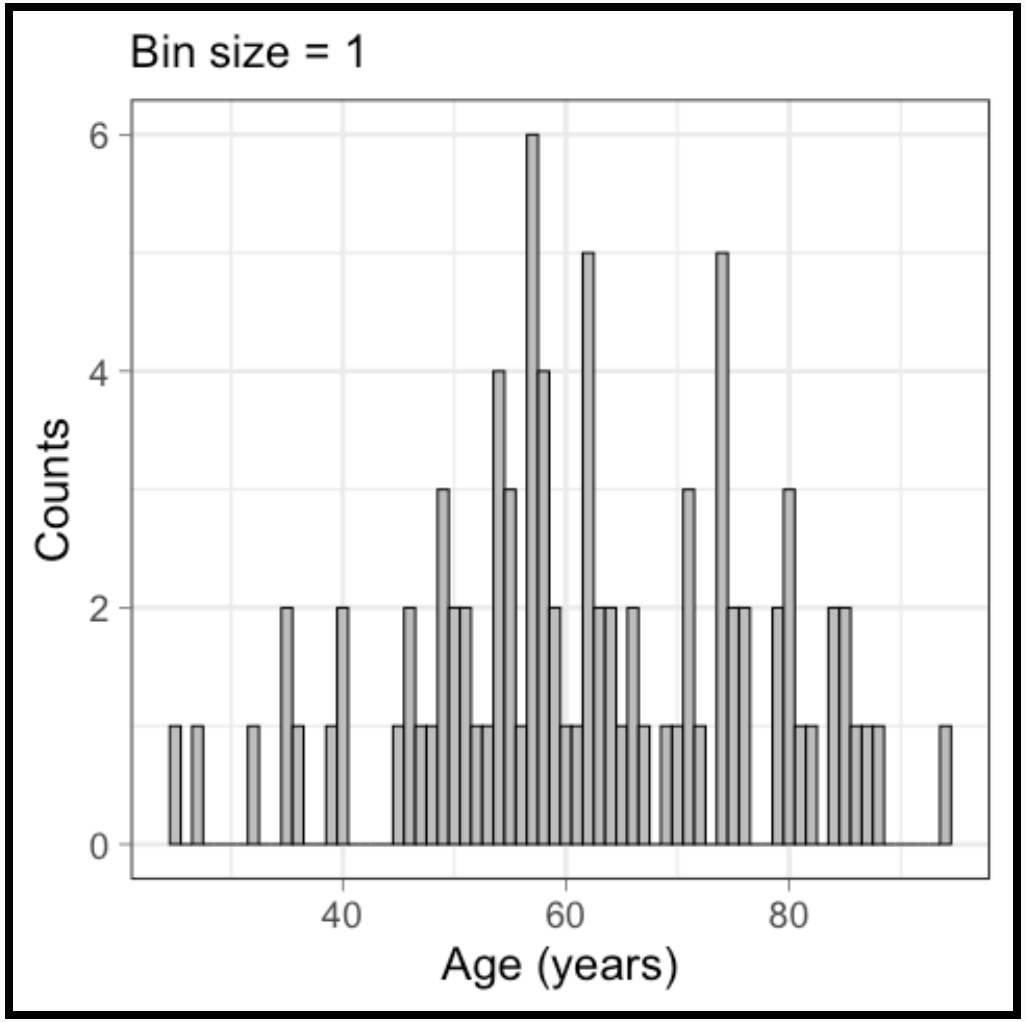
- categorical data
  - relative frequency



Visconti A., et al., Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Histogram

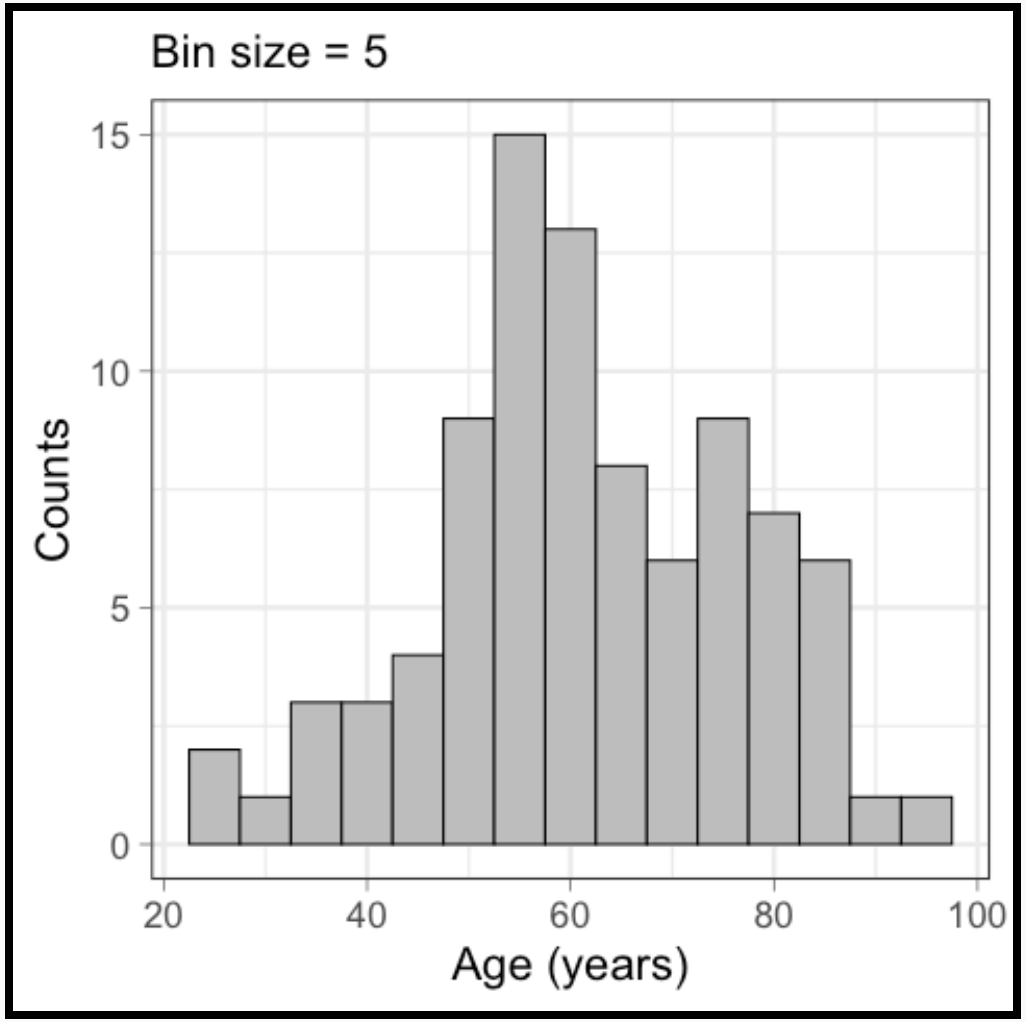
- numerical data



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Histogram

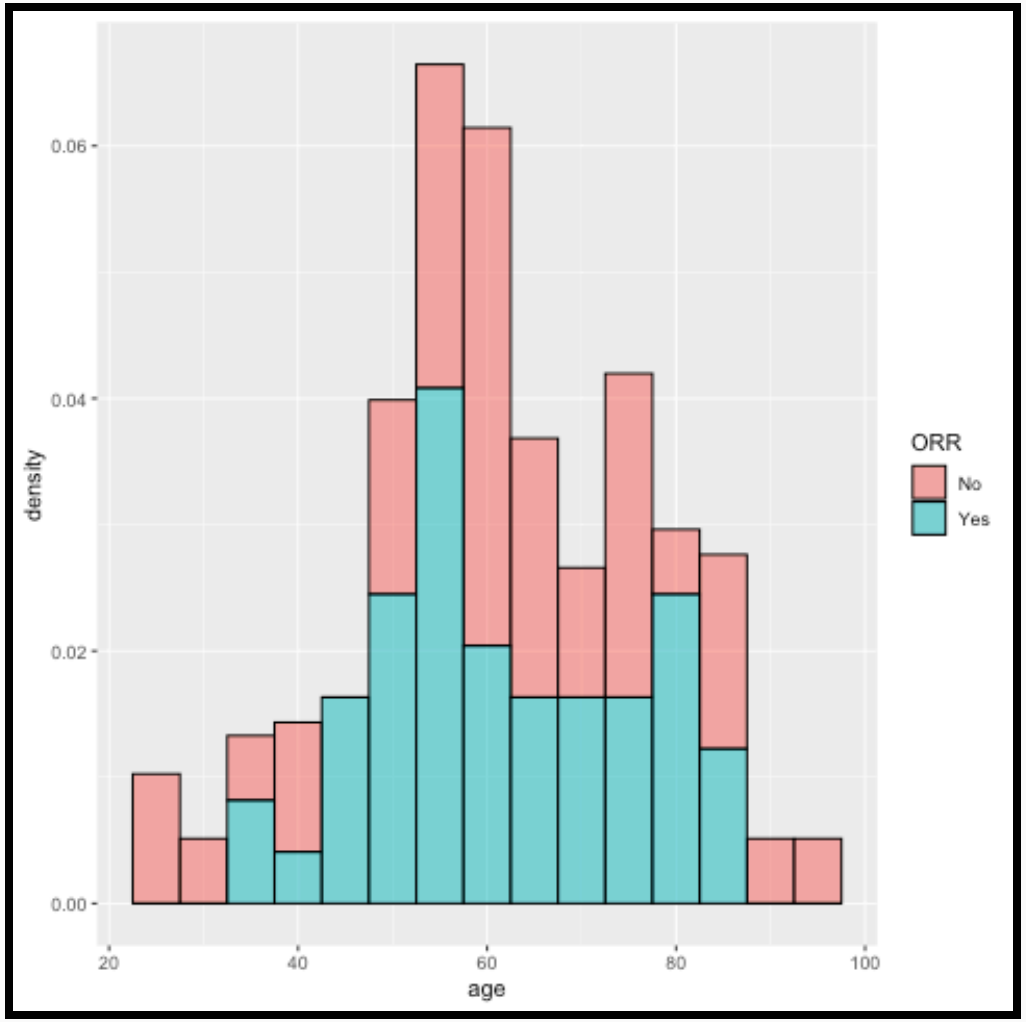
- numerical data



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Histogram

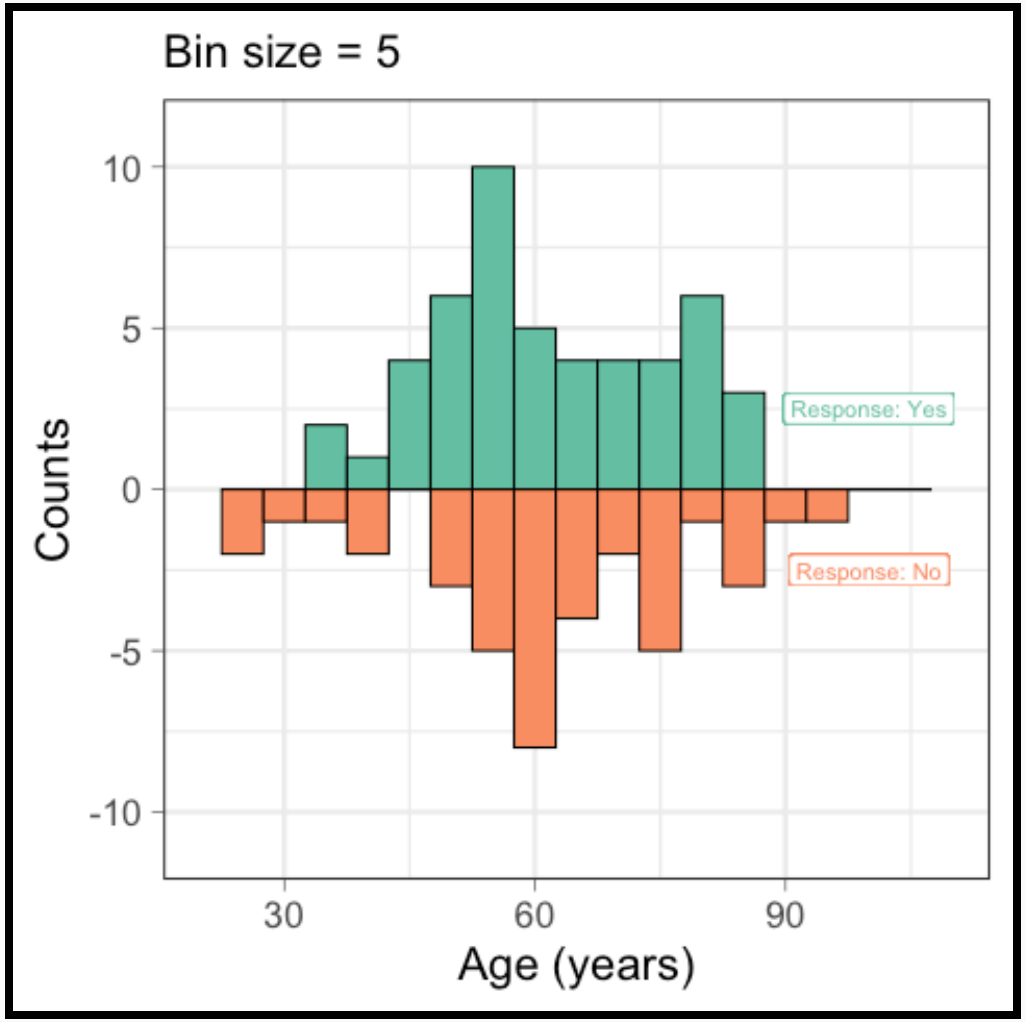
- numerical data



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

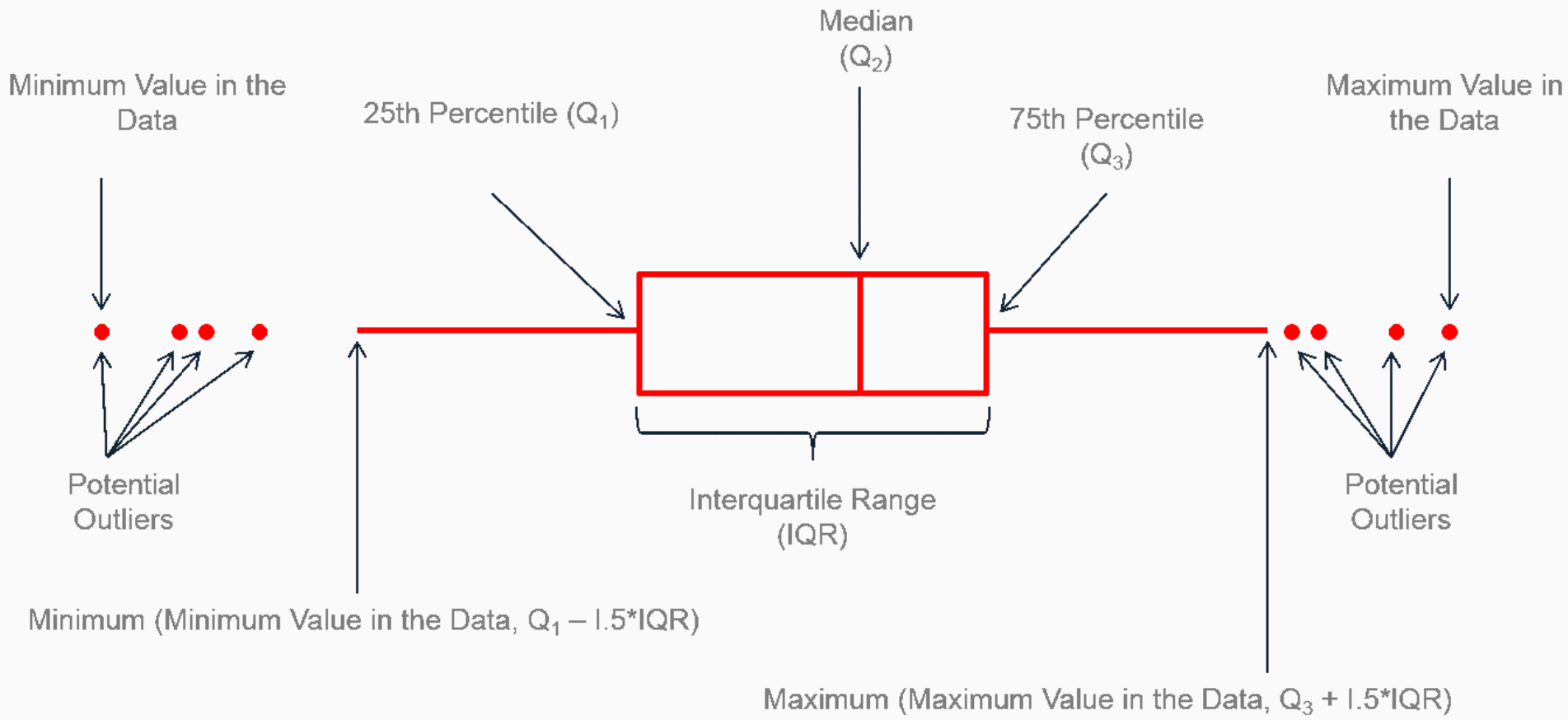
Miami plot/Mirror histogram

- numerical data



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

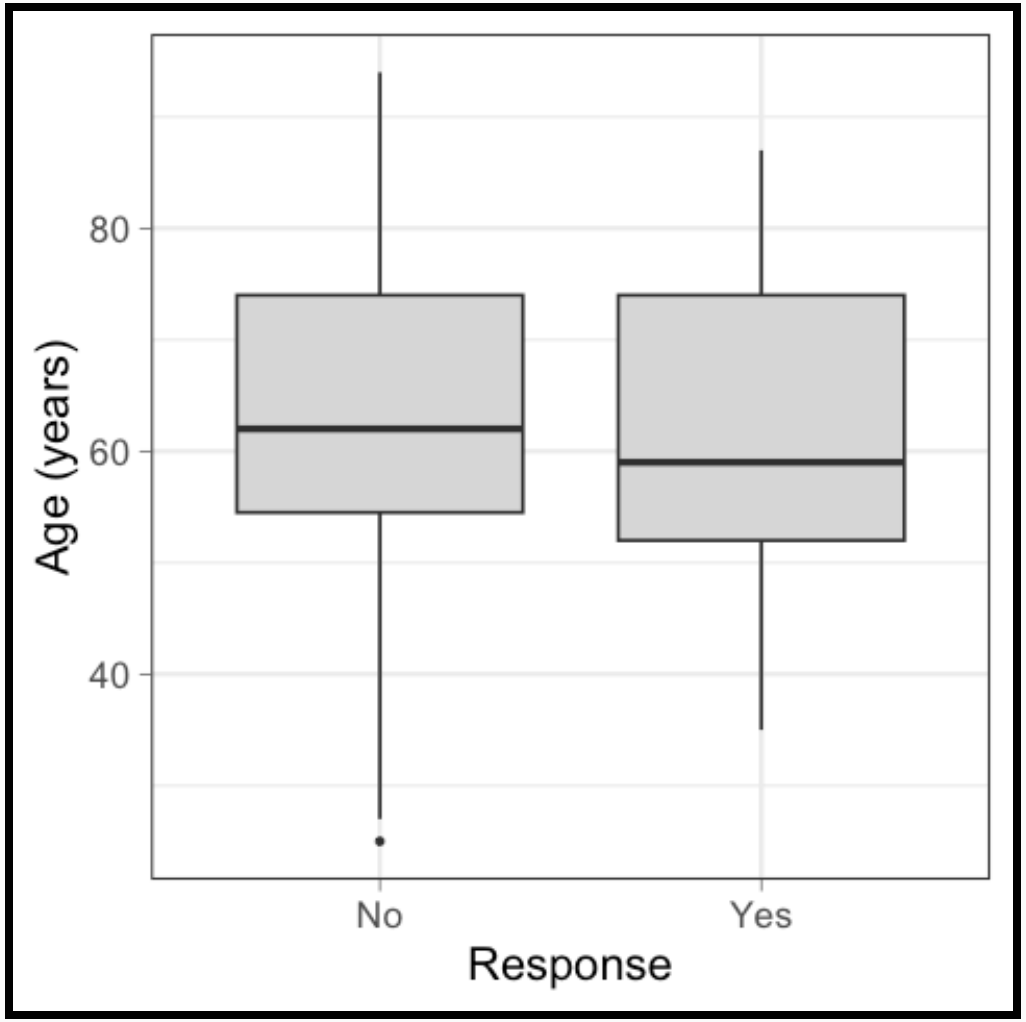
Boxplot



https://r-graph-gallery.com/boxplot

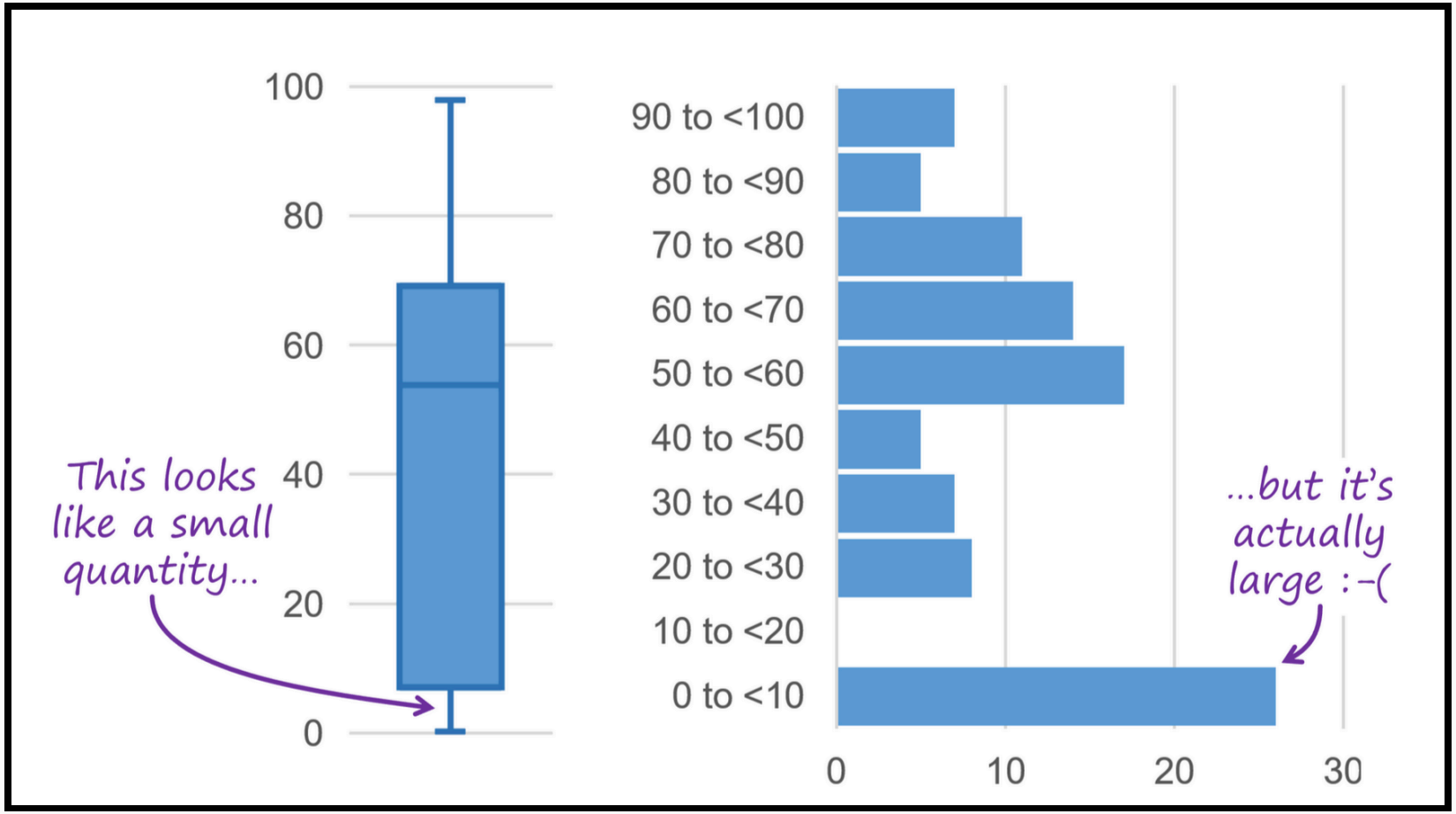
Boxplot

- numerical data



Visconti A., *et al.*, Total serum *N*-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

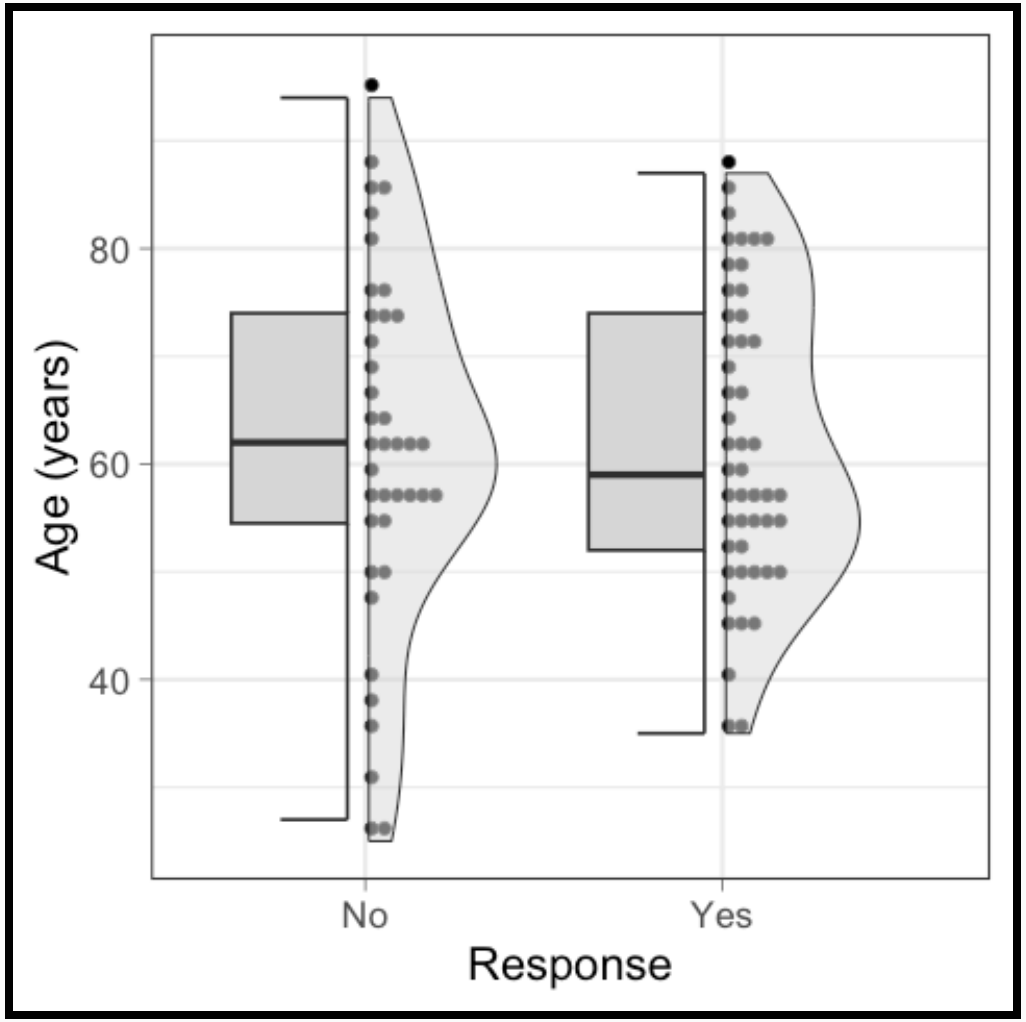
Boxplot



https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/

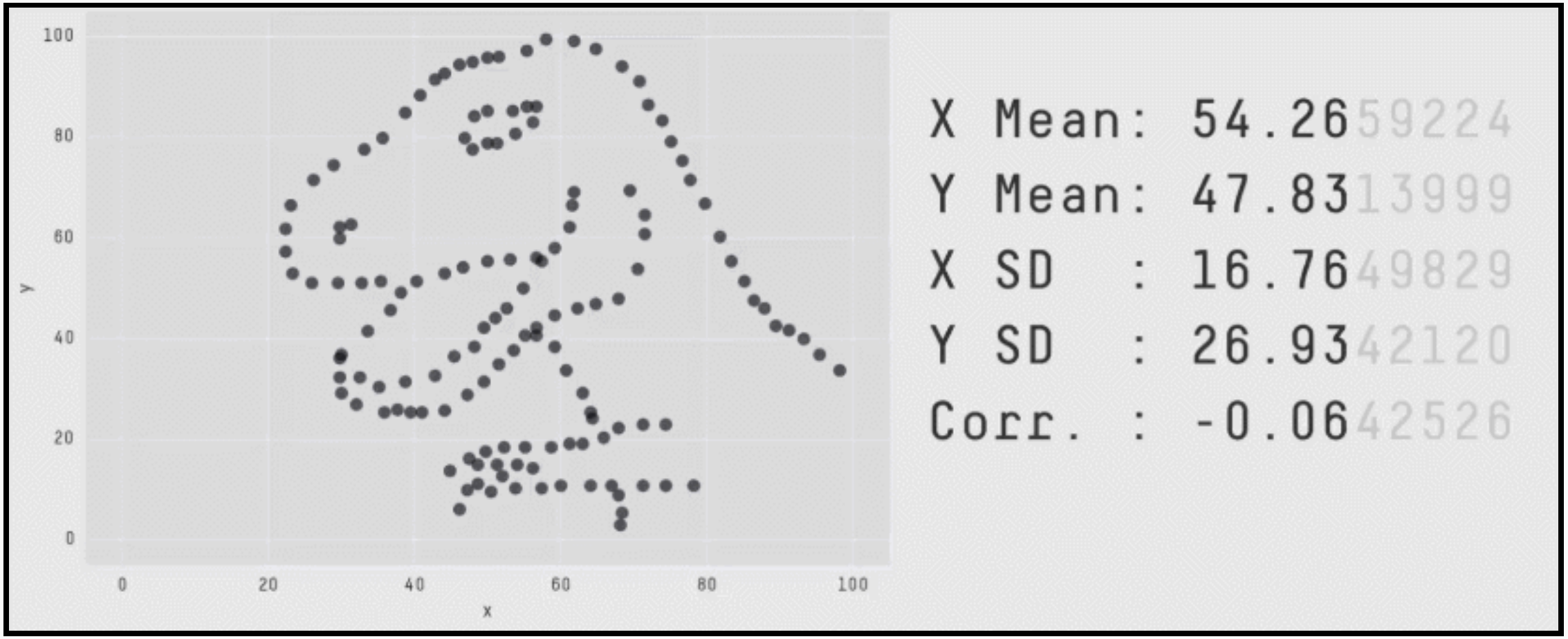
Boxplot

- numerical data



Visconti A., *et al.*, *Total serum N-glycans associate with response to immune checkpoint inhibition therapy and survival in patients with advanced melanoma*, BMC Cancer, 2023 doi:10.1186/s12885-023-10511-3

Data visualisation: DataSaurus Dozen



Summary

- Data come in different types
- Categorical variables can be summarized with absolute and relative frequencies
- Numerical variables can be summarized with measures of central tendency and dispersion, remembering that some of these measures are influenced by asimmetrical distribution and/or outliers
- Variables can be summarised with multiple graphical representations (but some are better than others), and visualising your data is always a good idea
- Samples are summarised by statistics, populations by parameters

See you tomorrow