

Introduction to statistics

(Day 1)

Agenda

- **Where:**
 - Mar 4: AULA GALILEO - MBC
 - Mar 5: AULA DARWIN - MBC
 - Mar 6 : Online

- **When:**
 - 14-17
 - 1 coffee break

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

- **How (to pass):**
 - Attend at least 2 lessons

How to ask questions/give feedback

- Interrupt me
- Take advantage of end/start/breaks
- Send emails alessia.visconti@unito.it
- Use the shared pad:
https://etherpad.wikimedia.org/p/intro_stats_2024_specialita (or
<https://t.ly/vRbvy>)



Introduction

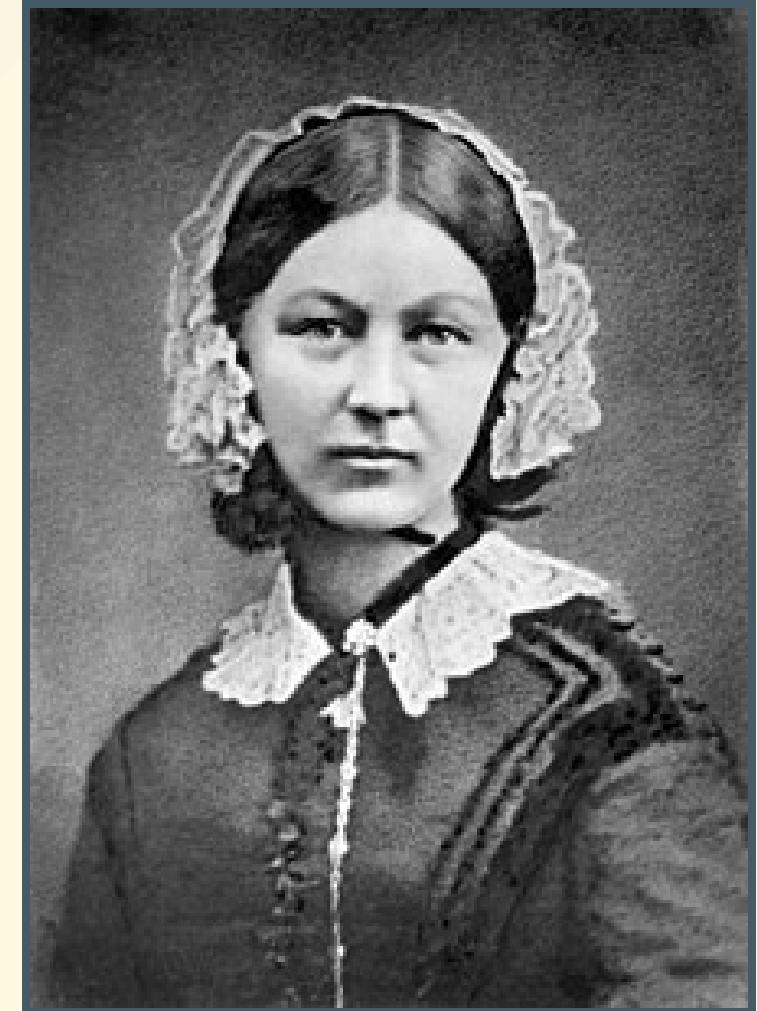


LIKE—
WHAT ARE
THE CHANCES
OF GETTING
A TAXI IN
THIS WEATHER?

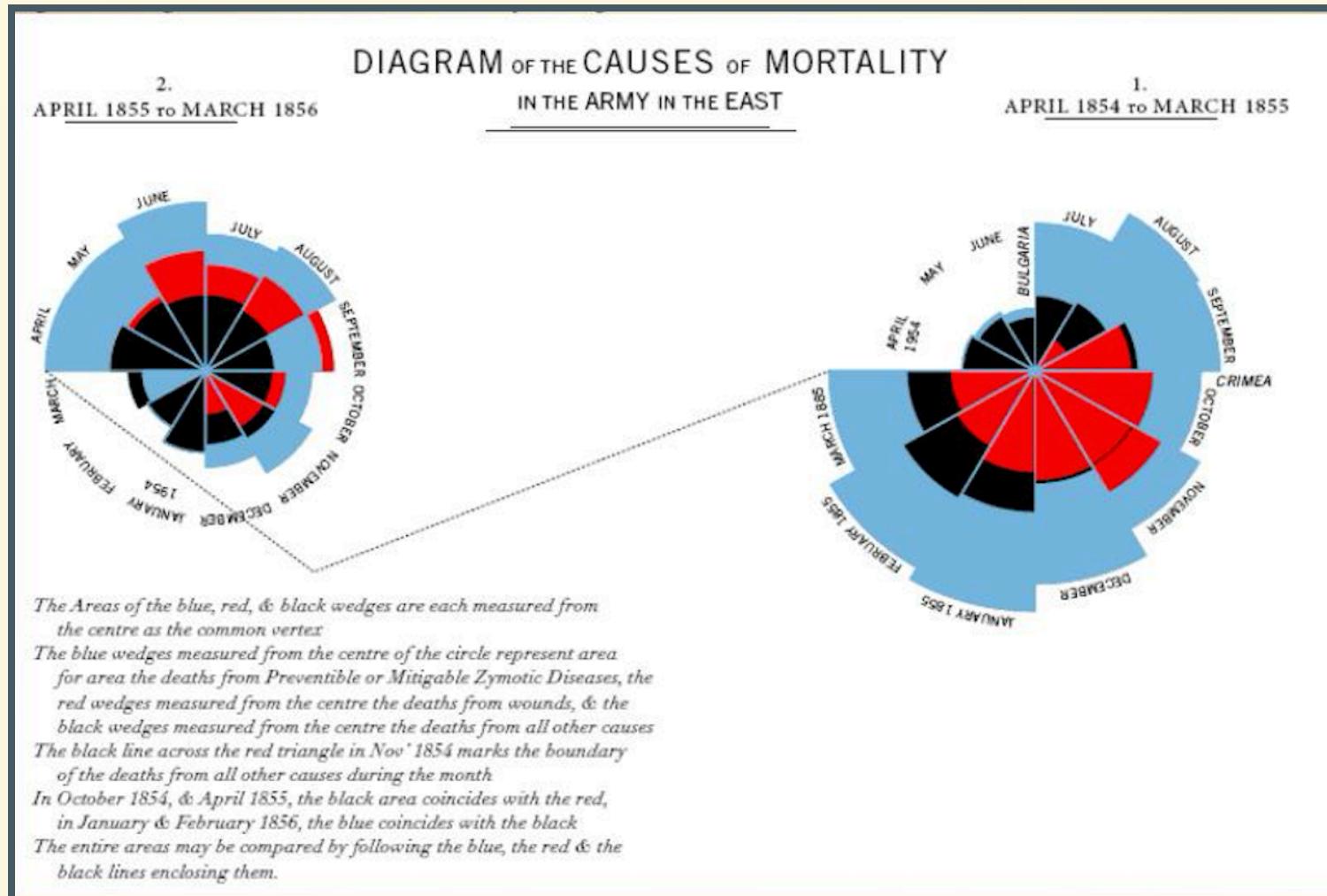
Why are we here?



Florence Nightingale



Florence Nightingale



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

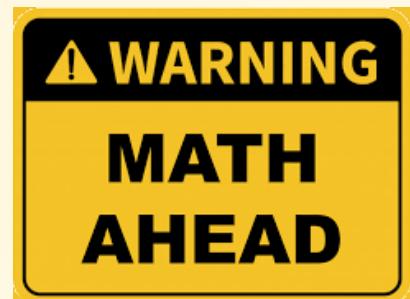
What will we learn?

Key Points

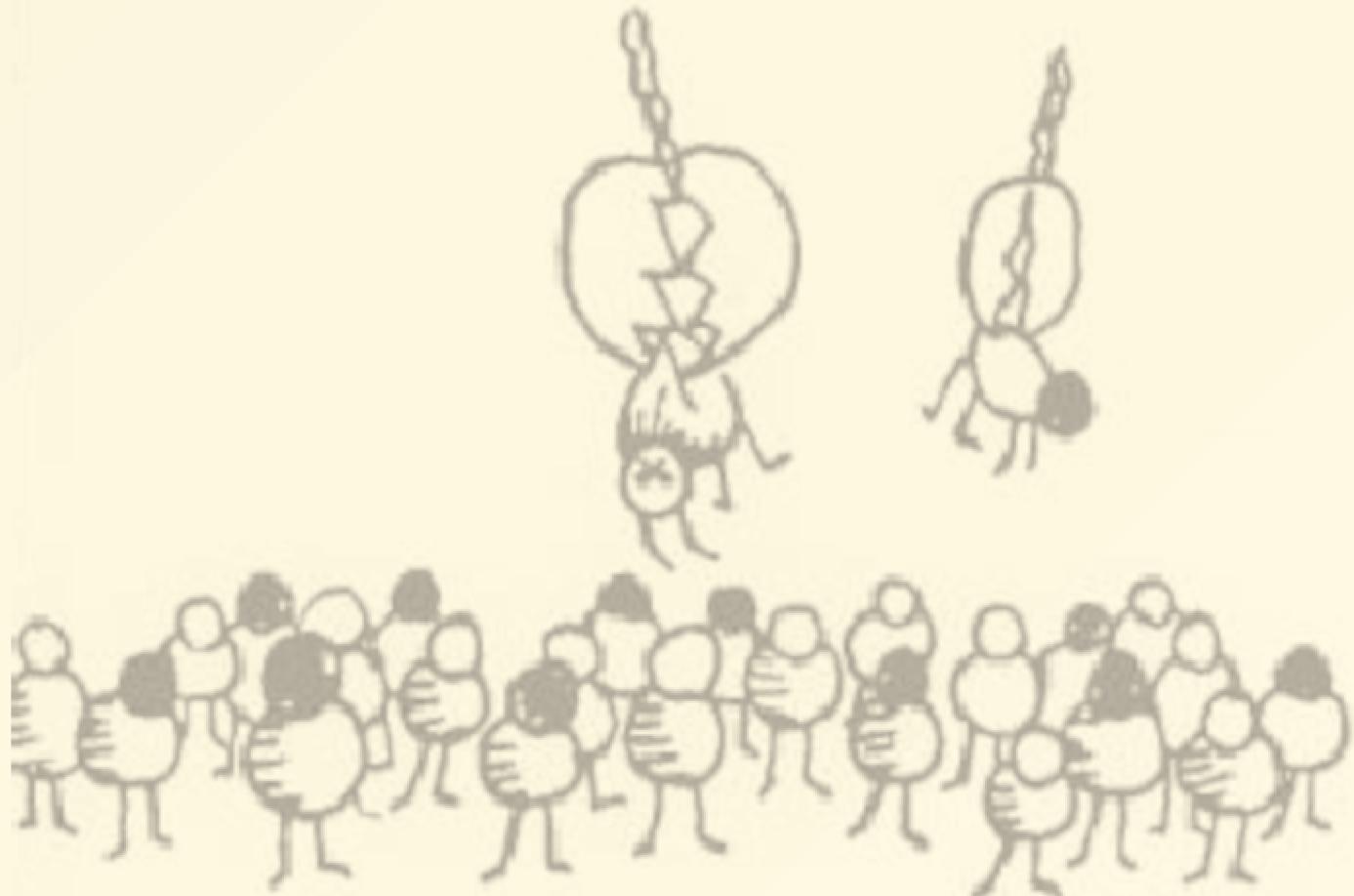
Question In patients with coronavirus disease 2019 (COVID-19) and moderate or severe acute respiratory distress syndrome (ARDS), does intravenous dexamethasone plus standard care compared with standard care alone increase the number of days alive and free from mechanical ventilation?

Topics covered

- How to sample from a population
- How to use measure of centrality and dispersion
- How to make estimation from a sample
- How to interpret confidence intervals
- How to make and test hypotheses



Sampling



Learning objectives

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

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.

CENSIMENTI PERMANENTI



PERMANENT CENSUS OF POPULATION AND HOUSING

PUBLIC INSTITUTIONS

NONPROFIT INSTITUTIONS

AGRICULTURE

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.

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



$N = 90$

$n = 10$

without replacement

La Tombola di PianetaBambini.it TABELLONE									
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



$N = 90$

$n = 10$

without replacement

49, 65, 25, 74, 18

90, 47, 24, 71, 37

TABELLONE									
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



$$N = 90$$

$$n = 10$$

$$x = 42$$

$$\text{step} = N/n = 90/10 = 9$$

La Tombola di PianetaBambini.it TABELLONE									
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



$$N = 90$$

$$n = 10$$

$$x = 42$$

$$\text{step} = N/n = 90/10 = 9$$

TABELLONE									
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 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



$$M = 90$$

$$N_{female} = 60$$

$$N_{male} = 30$$

$$n = 9$$

$$n_{female} = 6$$

$$n_{male} = 3$$

Females : 46, 20, 26,
50, 47, 3

Males : 69, 85, 87

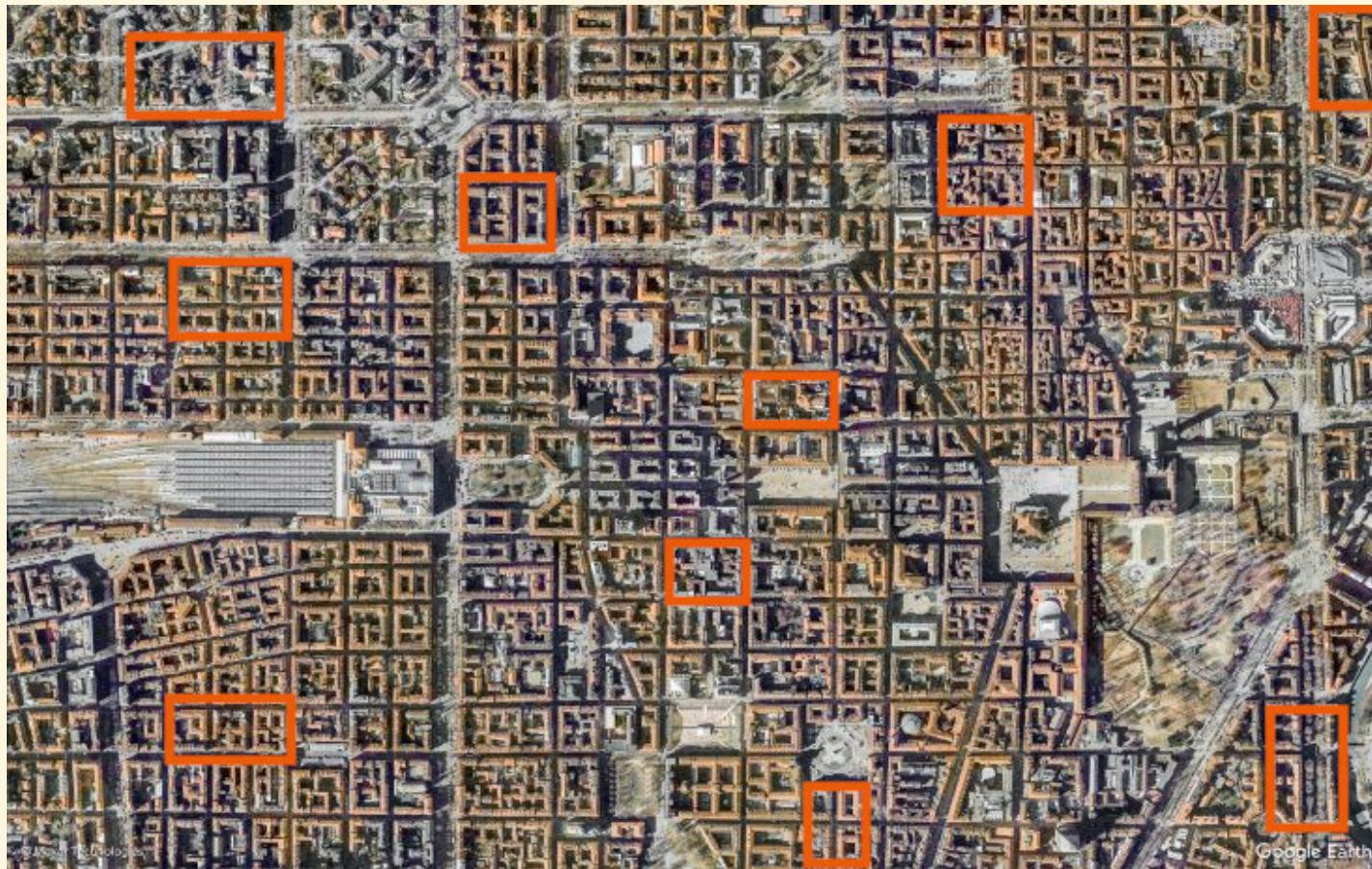
La Tombola di PianetaBambini.it										TABELLONE	www.pianetabambini.it/tombola	
1	2	3	4	5	6	7	8	9	10			
11	12	13	14	15	16	17	18	19	20			
21	22	23	24	25	26	27	28	29	30			
31	32	33	34	35	36	37	38	39	40			
41	42	43	44	45	46	47	48	49	50			
51	52	53	54	55	56	57	58	59	60			
61	62	63	64	65	66	67	68	69	70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

Strategy 4: Cluster sampling

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

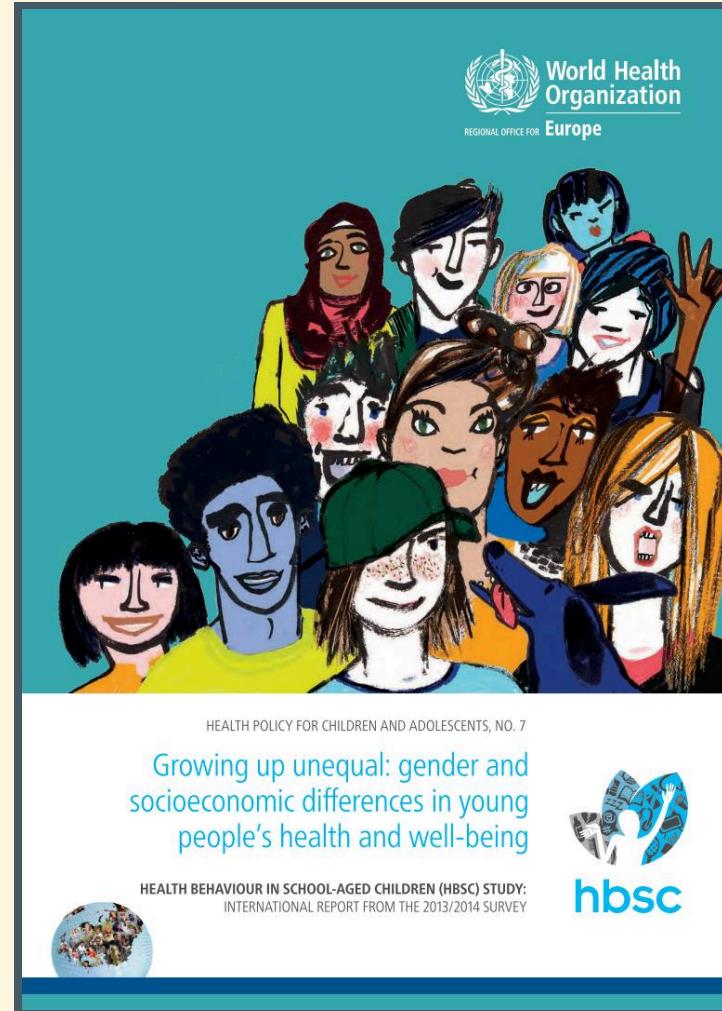
Variation: one stage (observing everything) *vs*
two stage (sampling within clusters)

Strategy 4: Cluster sampling



Sampling in the wild

<https://hbsc.org>



Exercise #1

- ? 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

Exercise #1 -- Solution

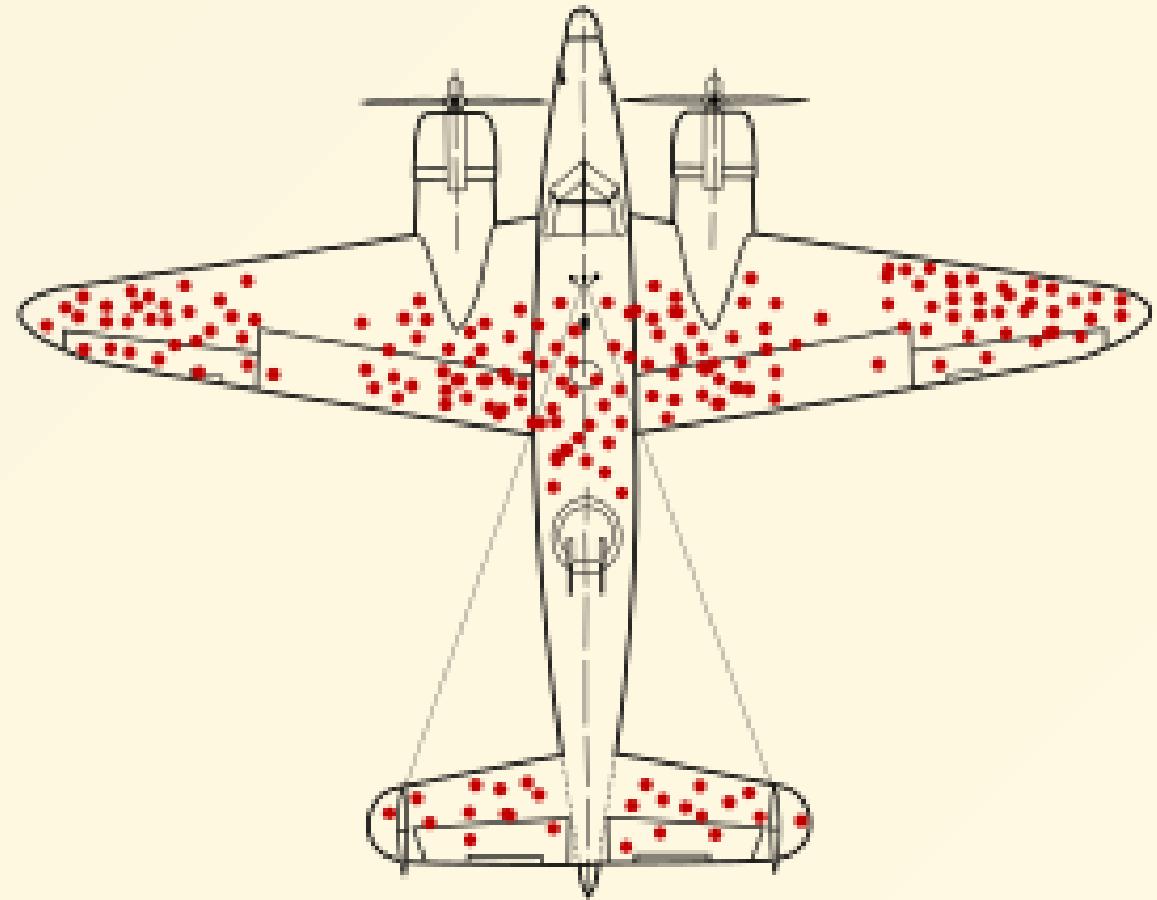
- ? A representative of a cheese factory is asking questions on cheese consumption to every 5th customer entering the supermarket

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



Selection bias in the wild

Patterns of item nonresponse behaviour to survey questionnaires are systematic and associated with genetic loci

[Gianmarco Mignogna](#), [Caitlin E. Carey](#), [Robbee Wedow](#)✉, [Nikolas Baya](#), [Mattia Cordioli](#), [Nicola Pirastu](#),
[Rino Bellococo](#), [Kathryn Fiuza Malerbi](#), [Michel G. Nivard](#), [Benjamin M. Neale](#), [Raymond K. Walters](#) &
[Andrea Ganna](#)✉

Nature Human Behaviour 7, 1371–1387 (2023) | [Cite this article](#)

Exercise #2

- ? Researchers send pensioners a snail mail to ask about their mental health after the COVID-19 lockdown. Pensioners are then asked to post their answer back

Will their study suffer any bias?

- a) No
- b) Yes, volunteer bias
- c) Yes, survivor bias
- d) Both b) and c)

Exercise #2 -- Solution

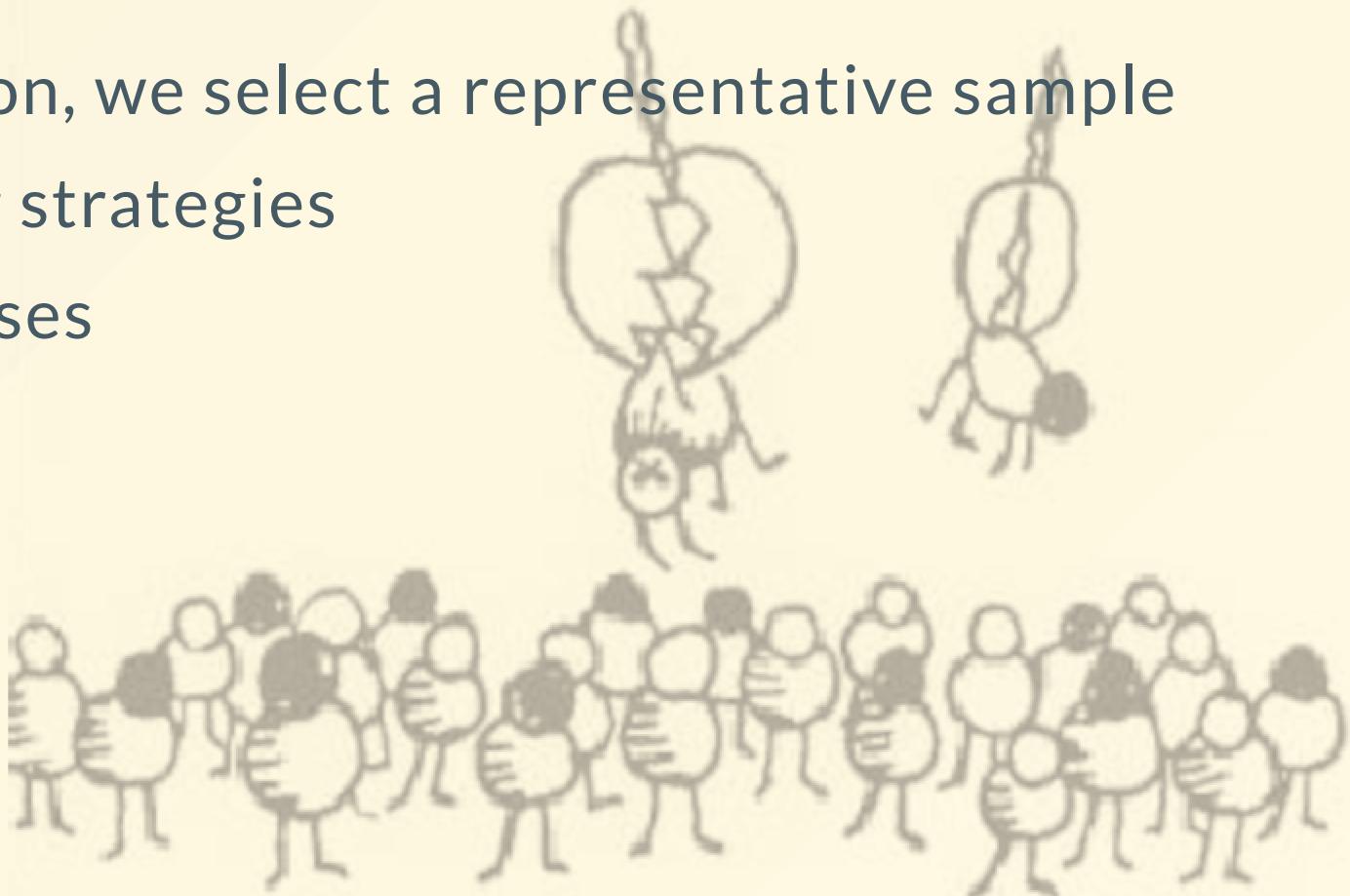
- ? Researchers send pensioners a snail mail to ask about their mental health after the COVID-19 lockdown. Pensioners are then asked to post their answer back

Will their study suffer any bias?

- a) No
- b) Yes, volunteer bias
- c) Yes, survivor bias
- d) Both b) and c)

Summary

- When can't study a population, we select a representative sample
- There are different sampling strategies
- Samples may suffer from biases



Summarise data

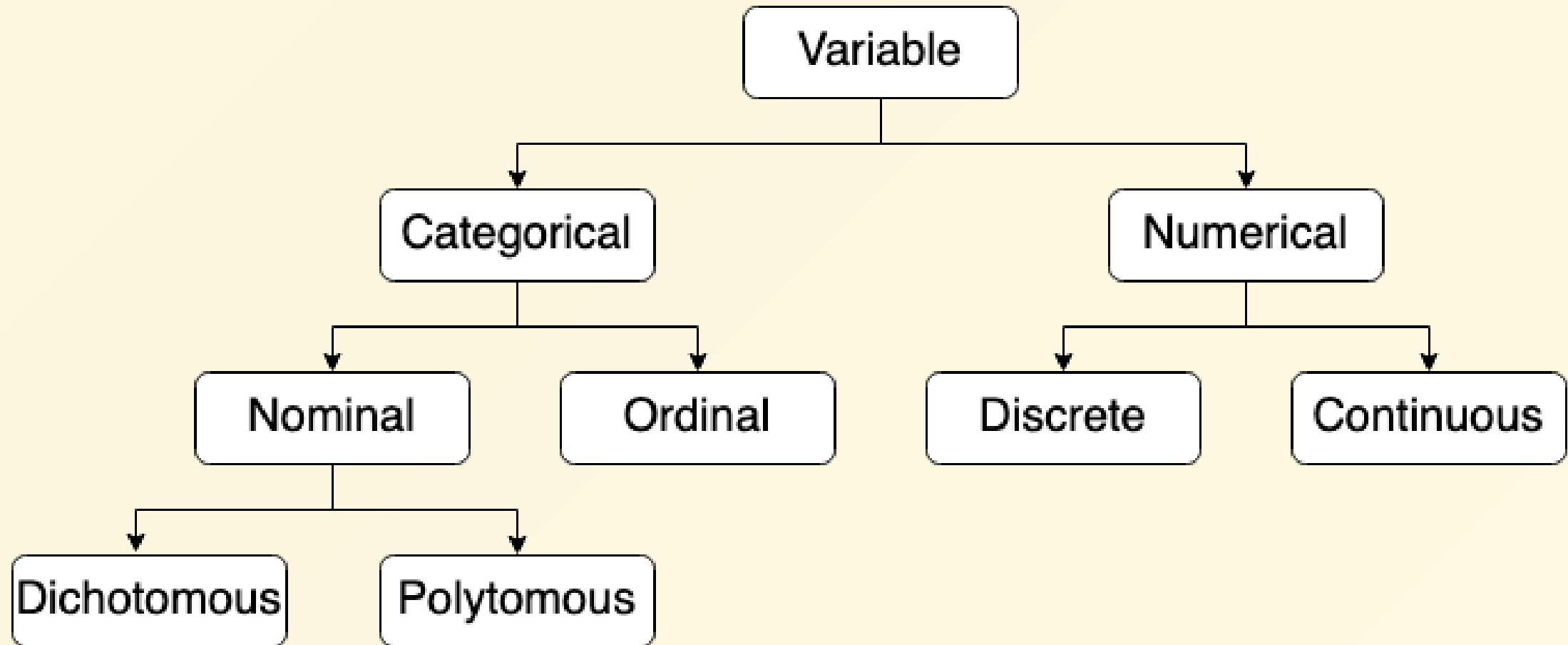
UM...
WELL... IT'S...
THEY'RE...
AHEM:
COUGH...



Learning objectives

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

Type of data



Exercise #3

? Which types of data are included in this table?

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	
<i>Male</i>	57 (64.8%)
<i>Female</i>	31 (35.2%)
Age (years)	60.5 ± 15.0
BMI (kg/m²)	28.0 ± 5.4
BRAF mutant	40 (45.5%)
LDH (\leqULN)	58 (65.9%)
Metastatic stage	
<i>Stage III unresectable</i>	2 (2.3%)
<i>M1a</i>	14 (15.9%)
<i>M1b</i>	17 (19.3%)
<i>M1c</i>	32 (36.4%)
<i>M1d</i>	23 (26.1%)
ECOG performance status	
0	47 (53.4%)
1	31 (35.2%)
2	8 (9.1%)
3	2 (2.3%)
ICI therapy	
<i>Ipilimumab</i>	1 (1.1%)
<i>Pembrolizumab</i>	20 (22.7%)
<i>Nivolumab</i>	30 (34.1%)
<i>Ipilimumab + Nivolumab</i>	37 (42.0%)

Why do we care?

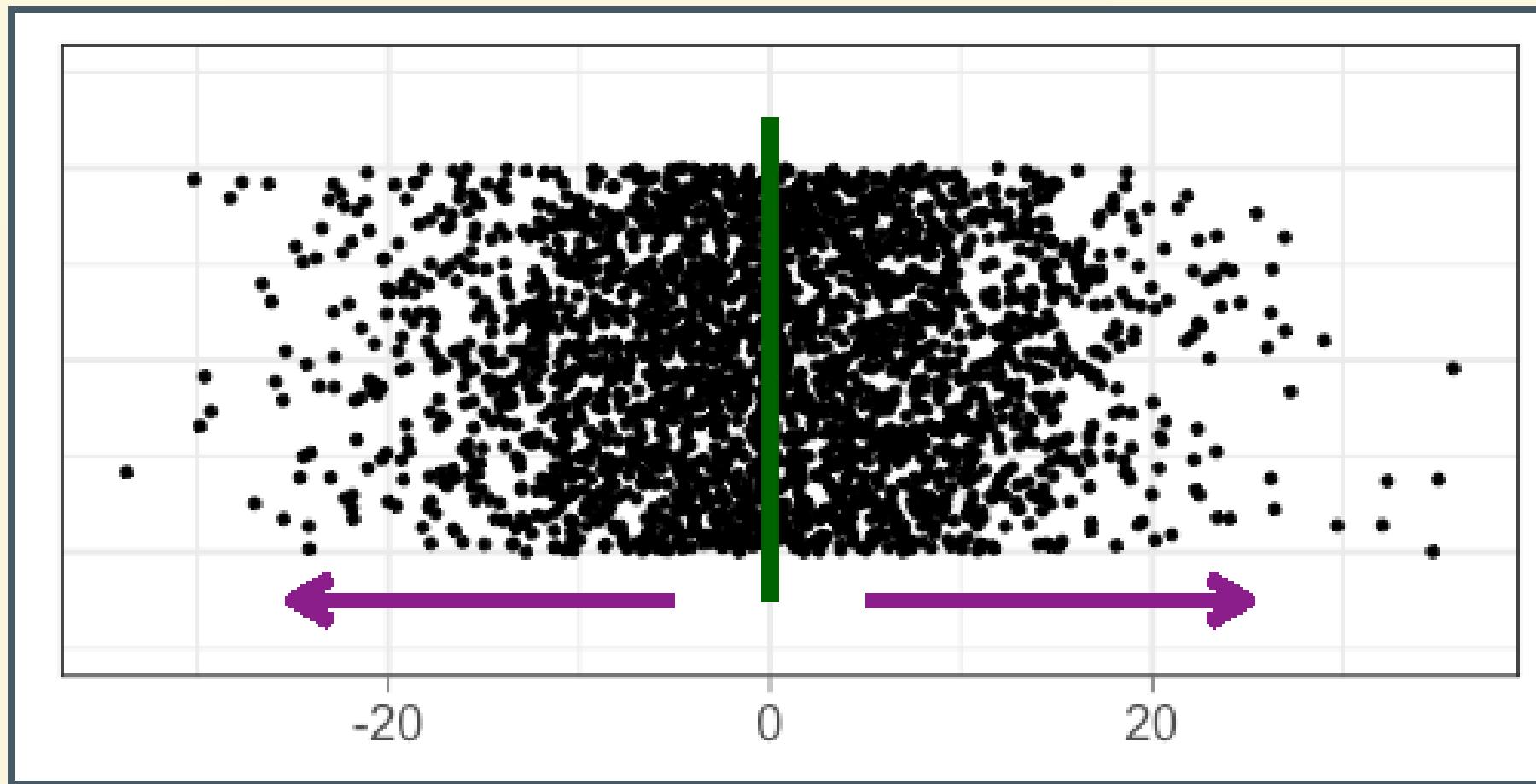
Table 1 Patient characteristics. Categorical variables are presented as number (percentage). Continuous variables are presented as mean \pm standard deviation.

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	
<i>Male</i>	57 (64.8%)
<i>Female</i>	31 (35.2%)
Age (years)	60.5 \pm 15.0
BMI (kg/m²)	28.0 \pm 5.4
BRAF mutant	40 (45.5%)
LDH (\leqULN)	58 (65.9%)
Metastatic stage	
<i>Stage III unresectable</i>	2 (2.3%)
<i>M1a</i>	14 (15.9%)
<i>M1b</i>	17 (19.3%)
<i>M1c</i>	32 (36.4%)
<i>M1d</i>	23 (26.1%)
ECOG performance status	
0	47 (53.4%)
1	31 (35.2%)
2	8 (9.1%)
3	2 (2.3%)
ICI therapy	
<i>Ipilimumab</i>	1 (1.1%)
<i>Pembrolizumab</i>	20 (22.7%)
<i>Nivolumab</i>	30 (34.1%)
<i>Ipilimumab + Nivolumab</i>	37 (42.0%)

Measures of centrality and dispersion



Measure of centrality: mode

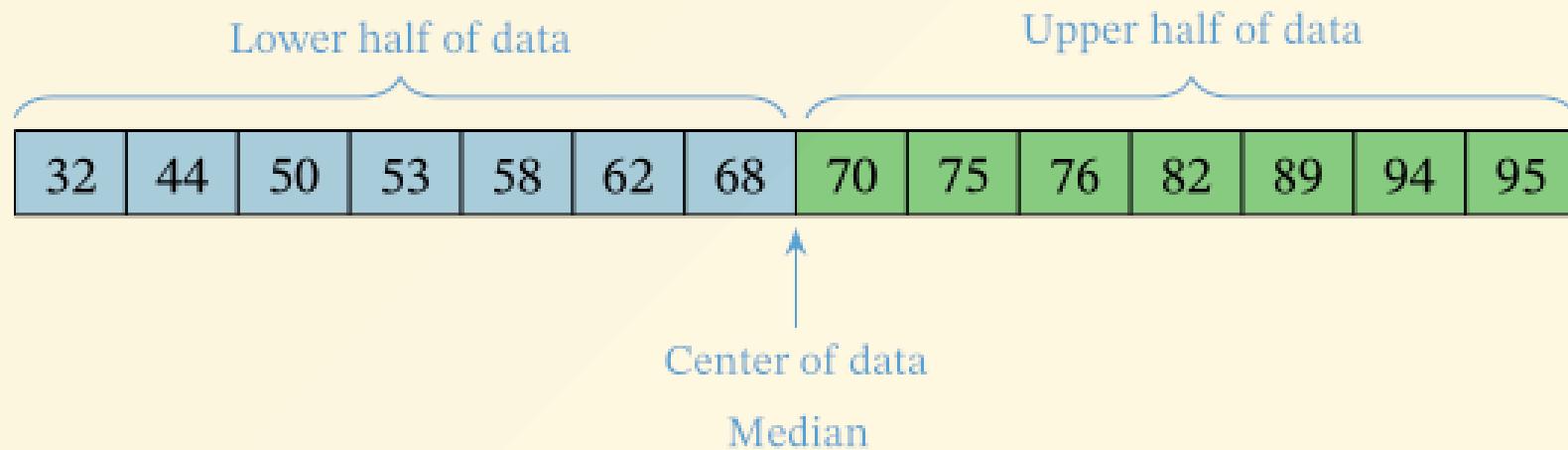
🎯 The most frequent item

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

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

Measure of centrality: median

- 🎯 The "middle" value



⚠️ Data should be sorted!

Measure of centrality: median

🎯 The "middle" value

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

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

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

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

⚠️ Data should be sorted!

Measure of centrality: median

 Robust to outliers

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

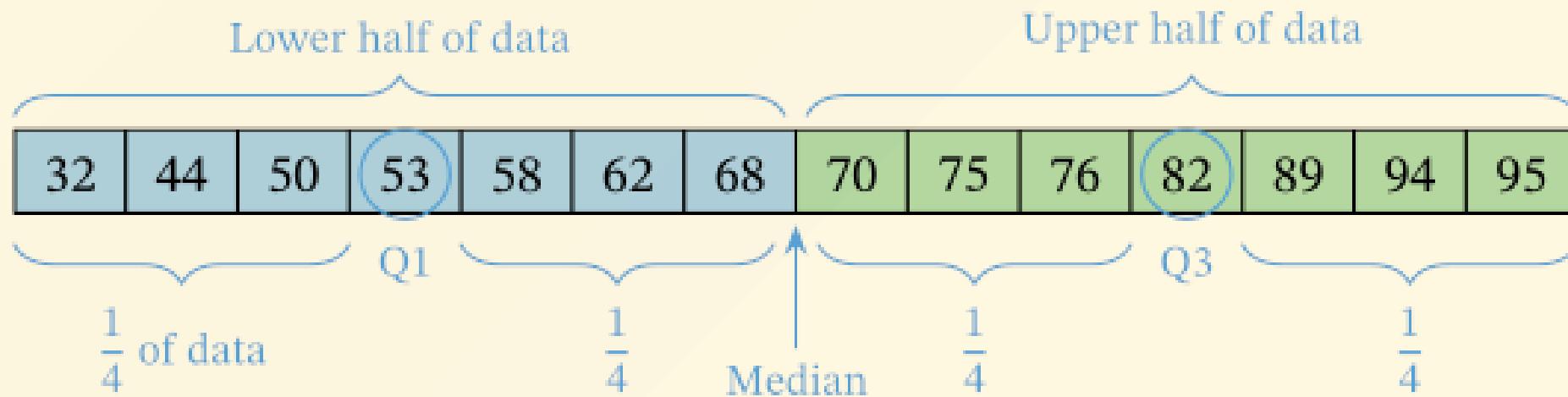
$$\text{median}(x) = x_{(n+1)/2} = x_{(7+1)/2} = x_4 = 6$$

 $n = 7, x = \{1, 3, 3, 6, 7, 8, 109\}$

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

 Data should be sorted!

Quartiles



⚠ Data should be sorted!

Quartiles

📌 $n = 11, x = \{6, 7, 15, 36, 39, \underline{40}, 41, 42, 43, 47, 49\}$

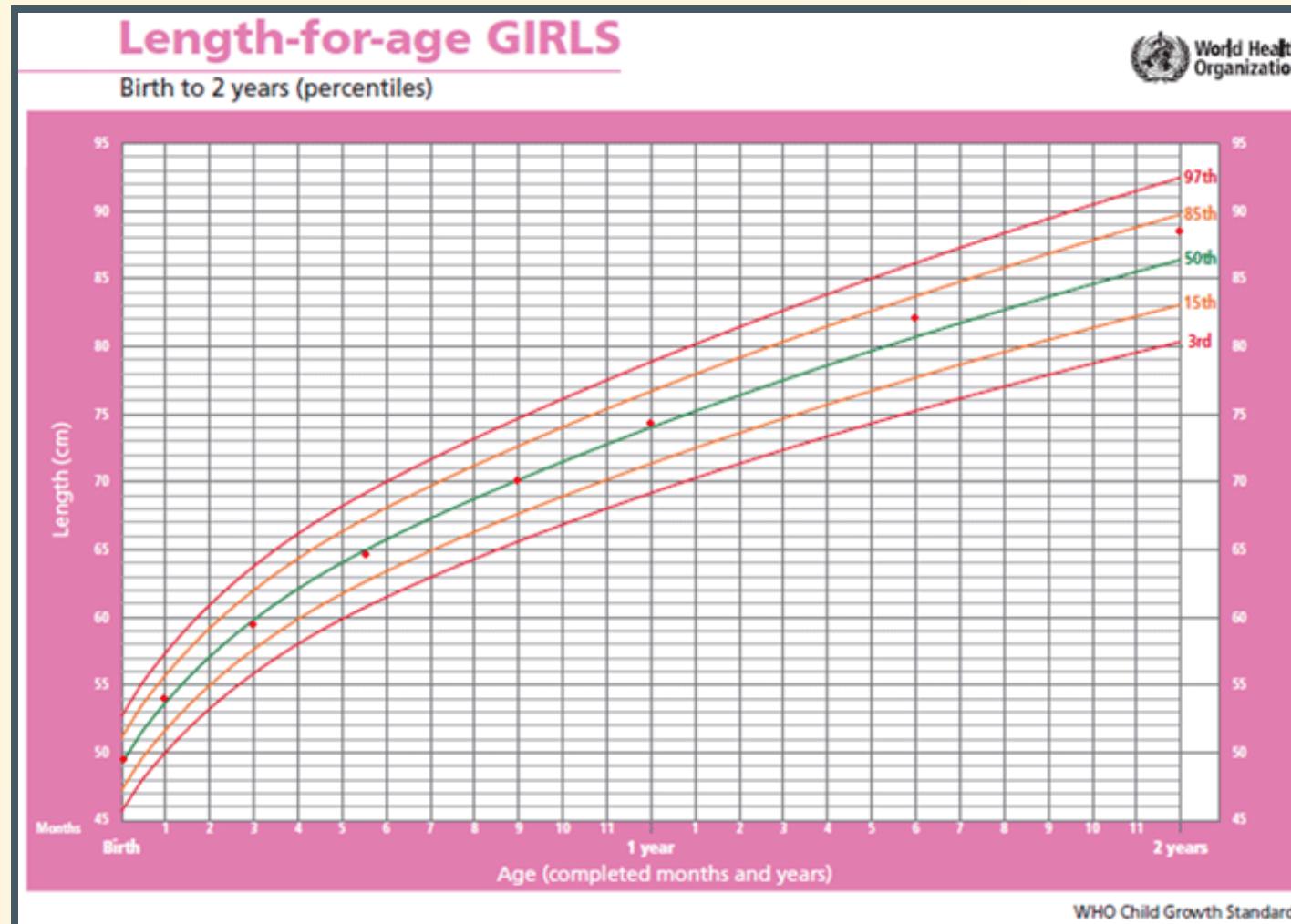
$$Q2(x) = \text{median}(x) = x_{(n+1)/2} = x_{(11+1)/2} = x_6 = 40$$

$$Q1(x_{1:5}) = x_{(5+1)/2} = x_3 = 15$$

$$Q3(x_{7:11}) = x_{(5+1)/2} = x_9 = 43$$

⚠ Data should be sorted!

Percentiles



Measure of centrality: mean

 Arithmetic mean

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

 $x = \{4, 36, 45, 50, 75\}$

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

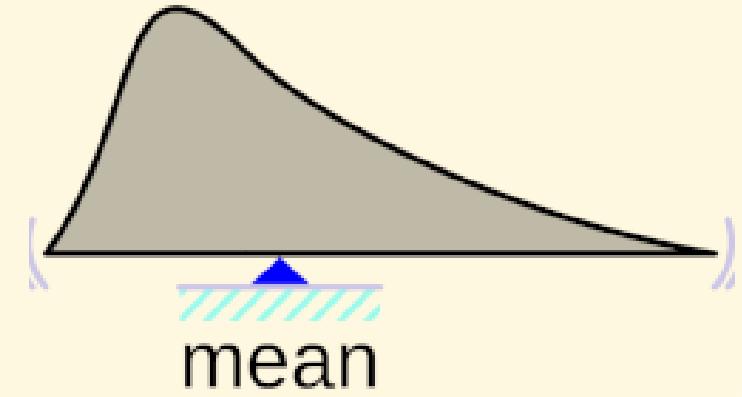
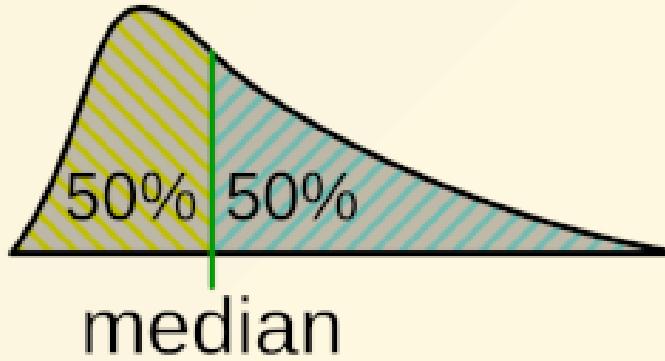
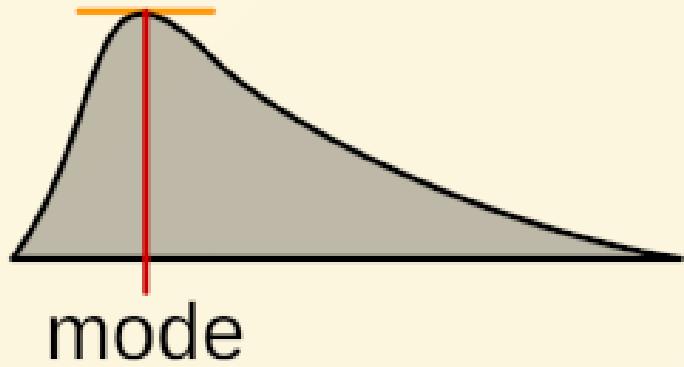
Measure of centrality: mean

🎯 Not really robust to outliers

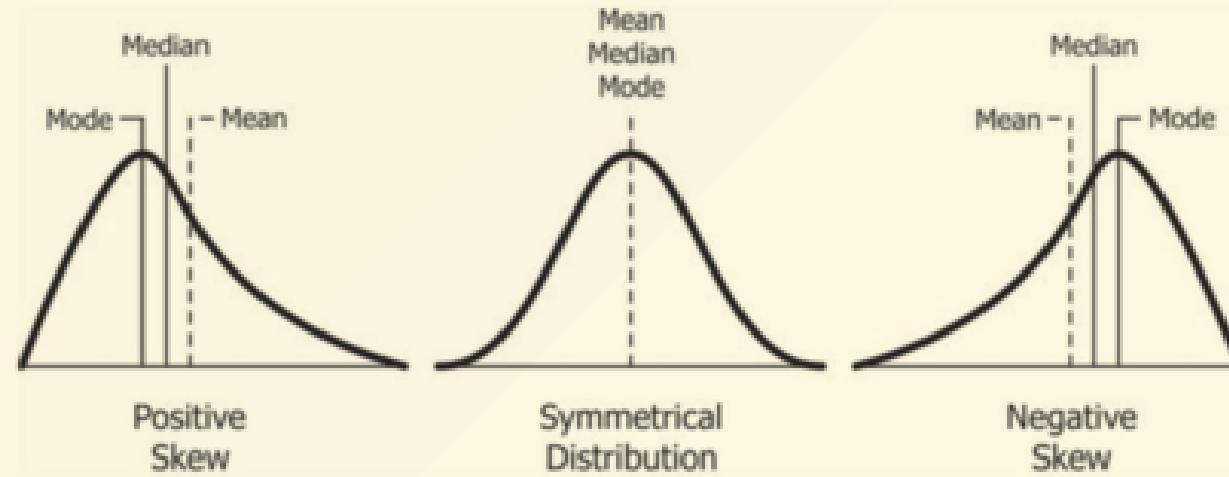
📌 $x = \{4, 36, 45, 50, 175\}$

$$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{4+36+45+50+175}{5} = 62$$

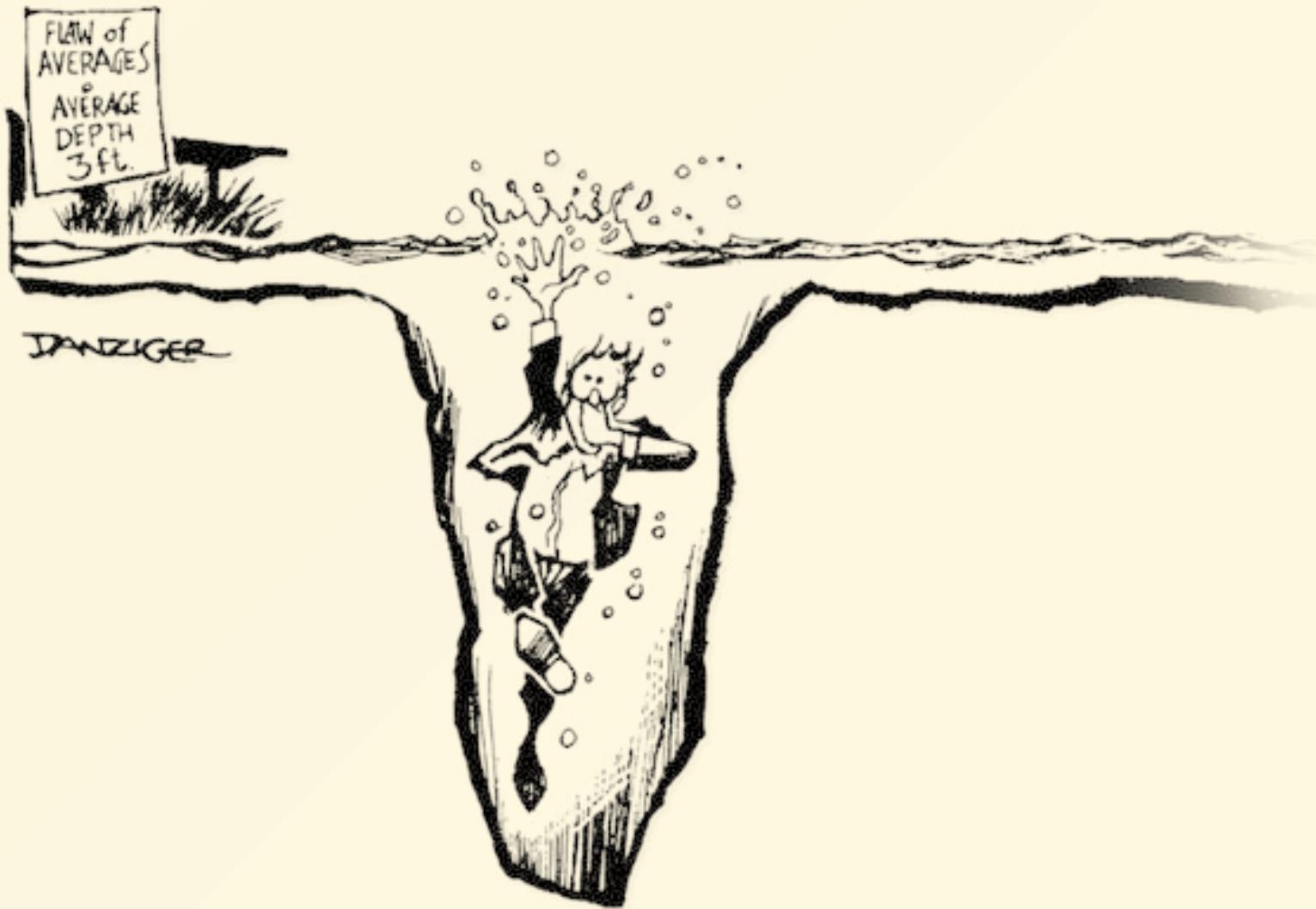
Mode vs median vs mean



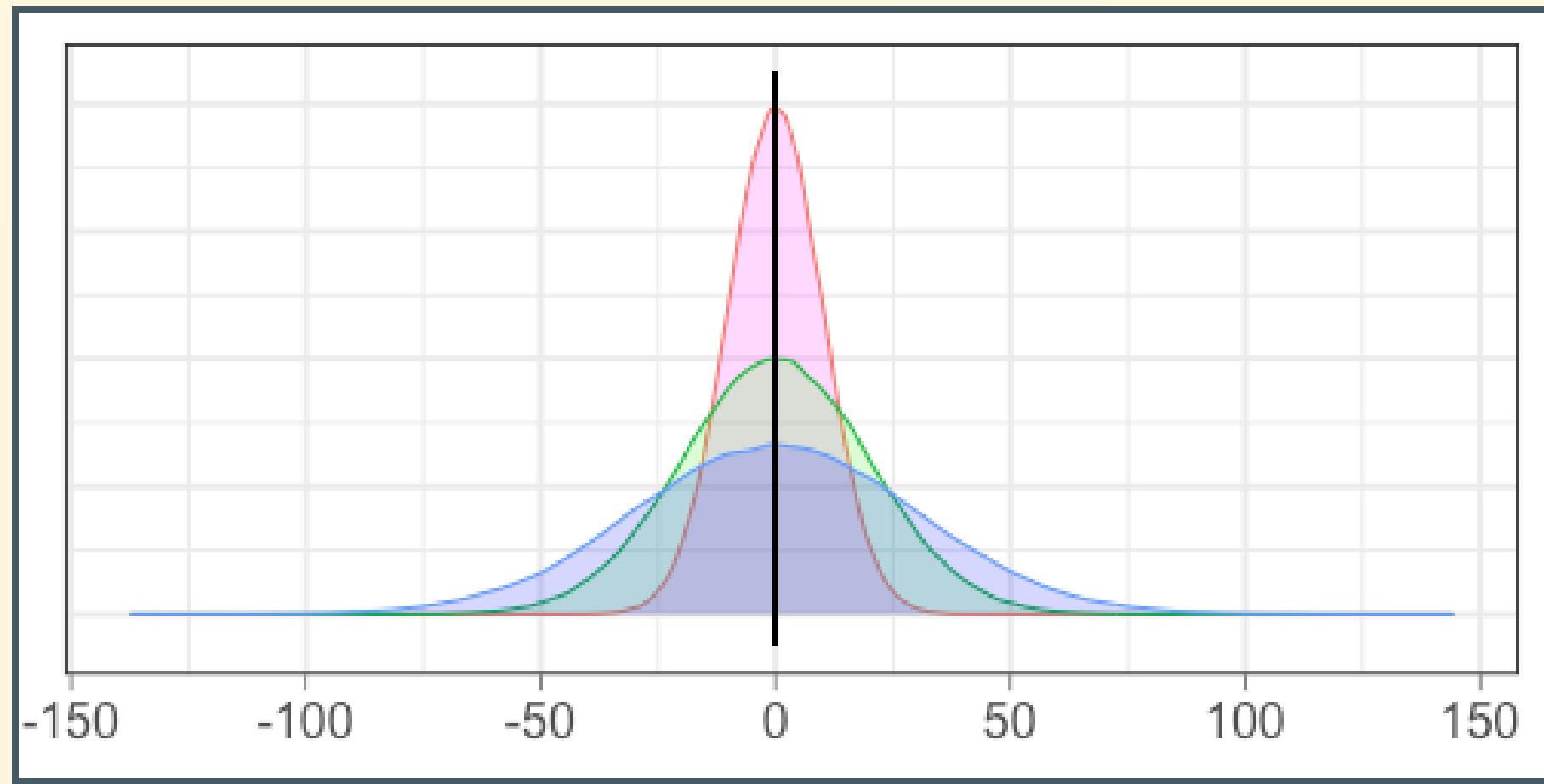
The shape of a distribution



Measures of dispersion



Measures of dispersion



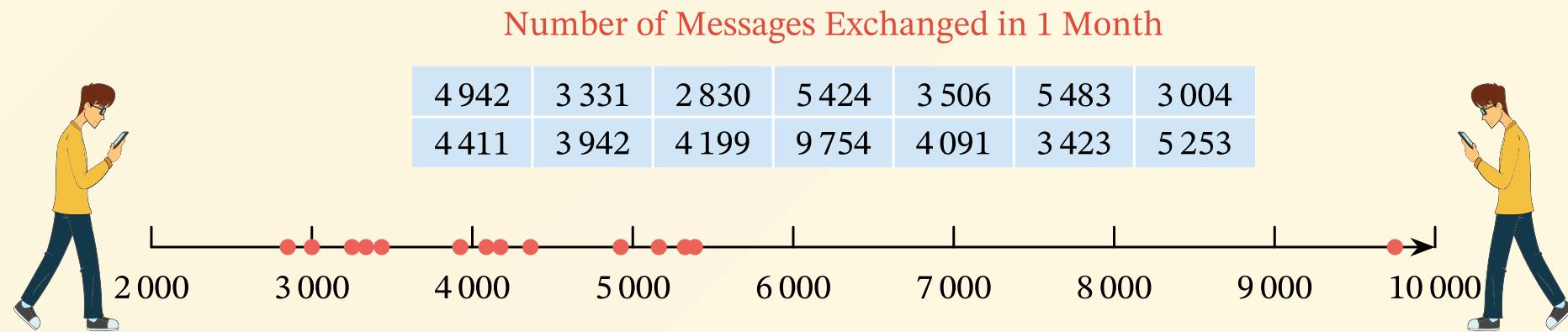
Measure of dispersion: range


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


$$x = \{6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49\}$$

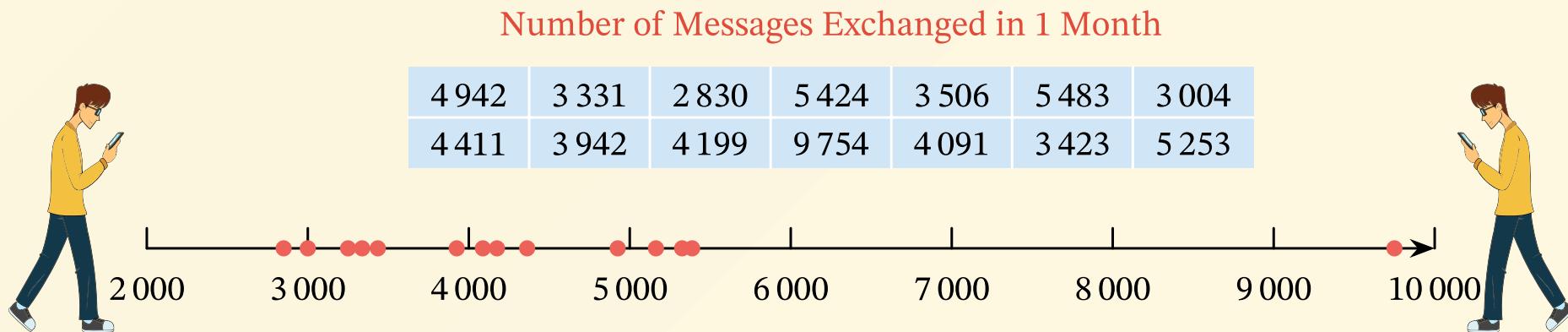
$$\text{range}(x) = \max(x) - \min(x) = 49 - 6 = 43$$

Measure of dispersion: range



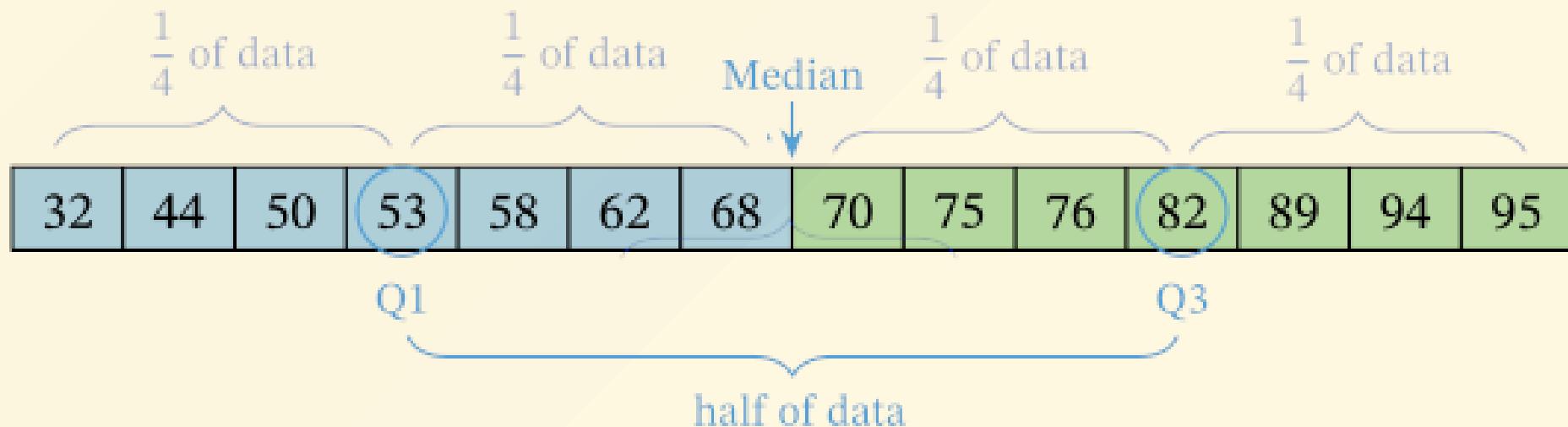
$$? \quad \text{range}(x) = ?$$

Measure of dispersion: range



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

Measure of dispersion: interquartile range



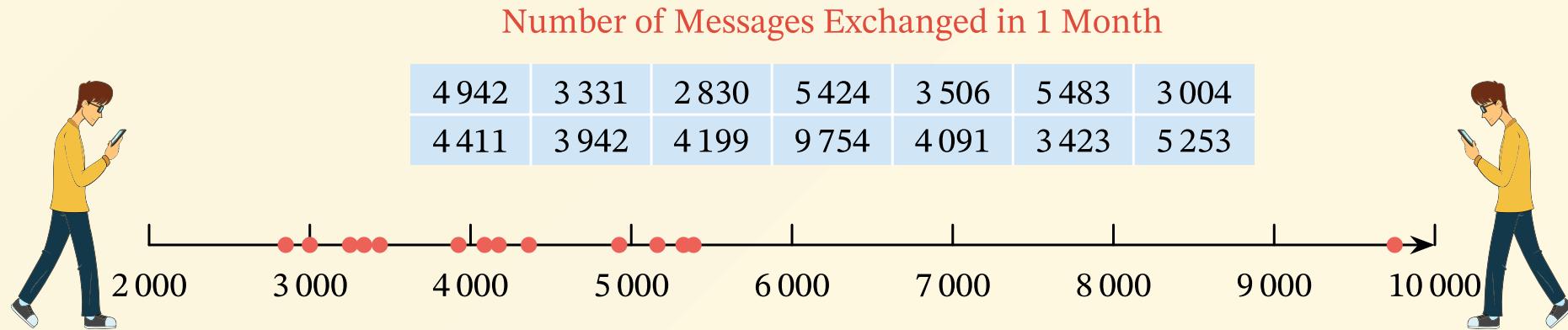
Measure of dispersion: interquartile range

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

📌 $x = \{6, 7, \underline{15}, 36, 39, \underline{40}, 41, 42, \underline{43}, 47, 49\}$

$$\text{IQR}(x) = \text{Q3}(x) - \text{Q1}(x) = 43 - 15 = 28$$

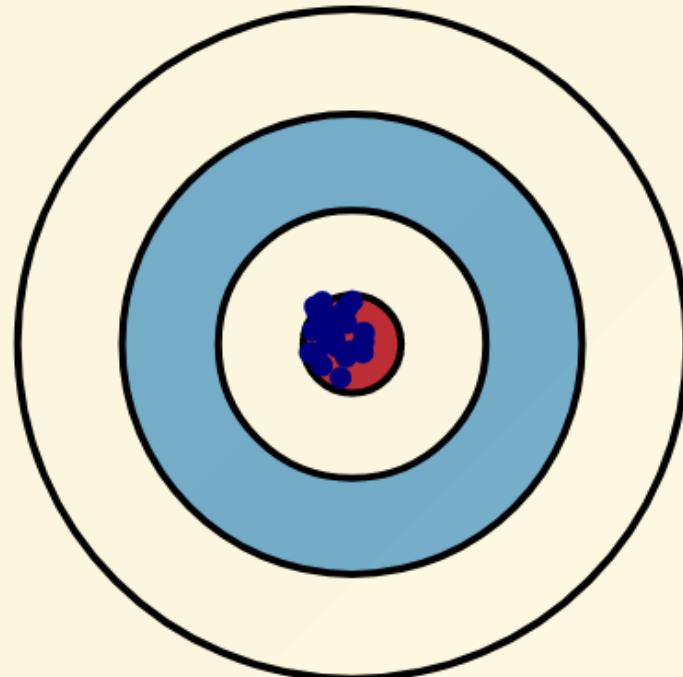
Measure of dispersion: interquartile range



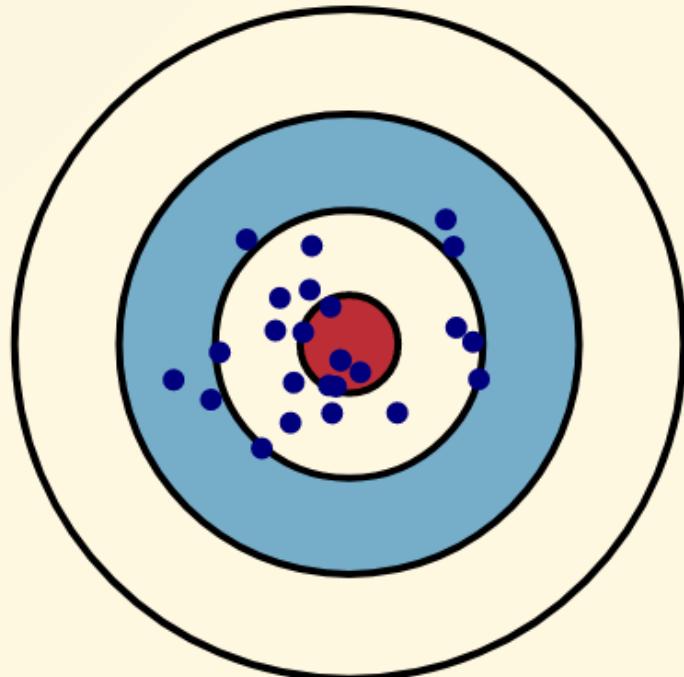
- range(x) = $\max(x) - \min(x) = 9,754 - 2,830 = 6,924$
- IQR(x) = $Q3(x) - Q1(x) = 5,253 - 3,423 = 1,830$

Measure of dispersion: variance

Low Variance



High Variance



Measure of dispersion: variance


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

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


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

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

Measure of dispersion: variance


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

why $\frac{1}{n-1}$? $\rightarrow \sum_{i=1}^n (x_i - \bar{x}) = 0$


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

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x}) &= (1 - 2) + (2 - 2) + (3 - 2) = \\ &= -1 + 0 + 1 = 0 \end{aligned}$$

Measure of dispersion: standard deviation


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

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


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

$$\begin{aligned} s &= \sqrt{\frac{1}{3-1} \times [(1-2)^2 + (2-2)^2 + (3-2)^2]} = \\ &= \sqrt{\frac{1}{2} \times [1^2 + 0^2 + 1^2]} = \sqrt{\frac{1}{2} \times 2} = \sqrt{1} = 1 \end{aligned}$$

Centrality, dispersion, and data types

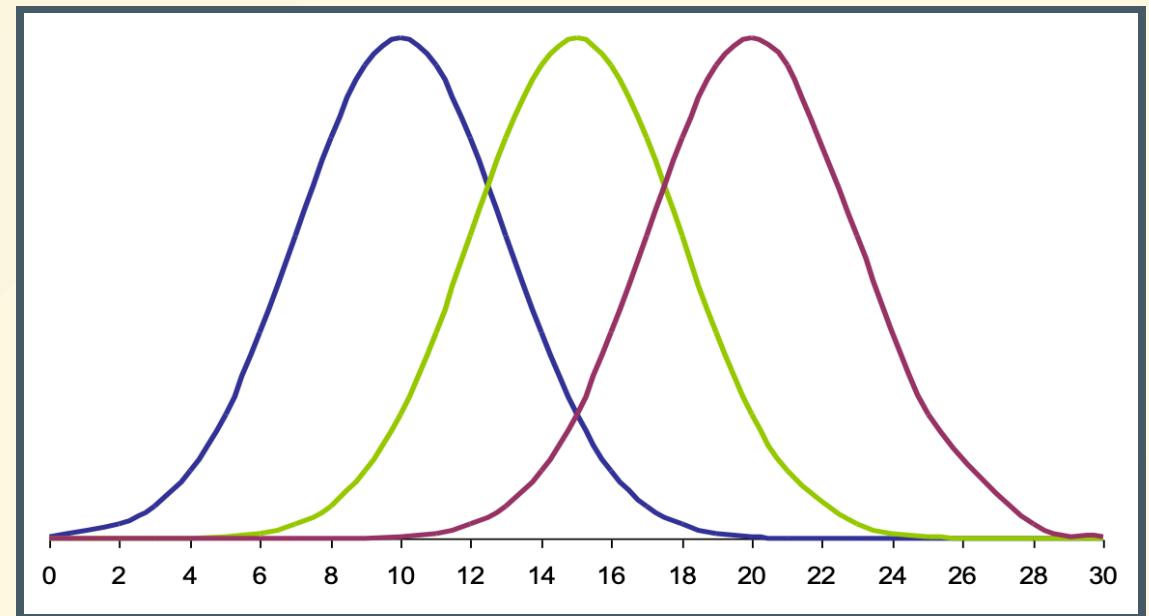
Data type	Centrality Measure	Dispersion Measure
Nominal	Mode	-
Ordinal	Median, Mode	Range, IQR
Numeric	Mean, Median, Mode	Range, IQR, standard deviation

Parameters vs statistics

- Parameters: calculated on the population (μ, σ)
- Statistics: calculated on the sample (\bar{x}, s)

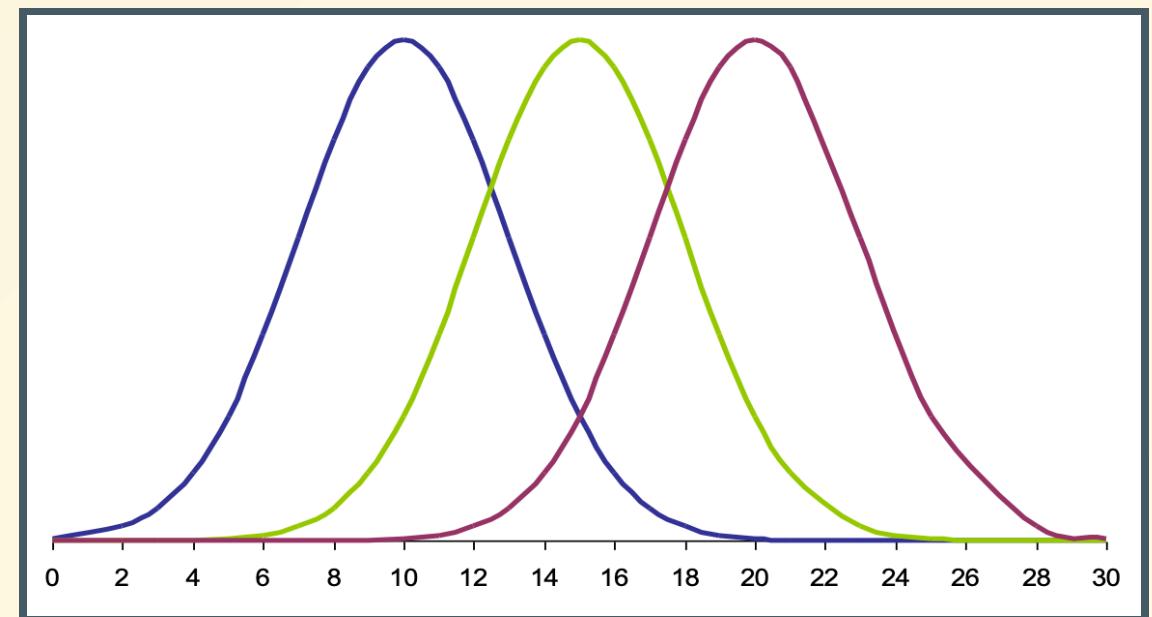
Exercise #4

- ? Which curve has the larger mean?
- a) blue
 - b) green
 - c) red



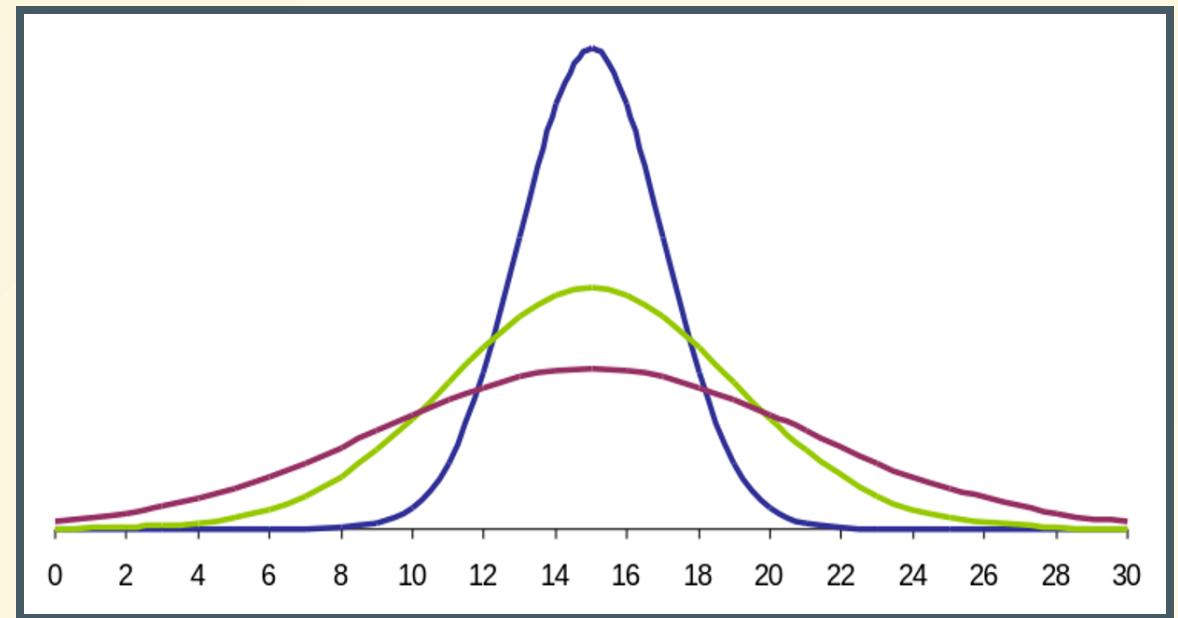
Exercise #4 -- Solution

- ? Which curve has the larger mean?
- a) blue
 - b) green
 - c) red 



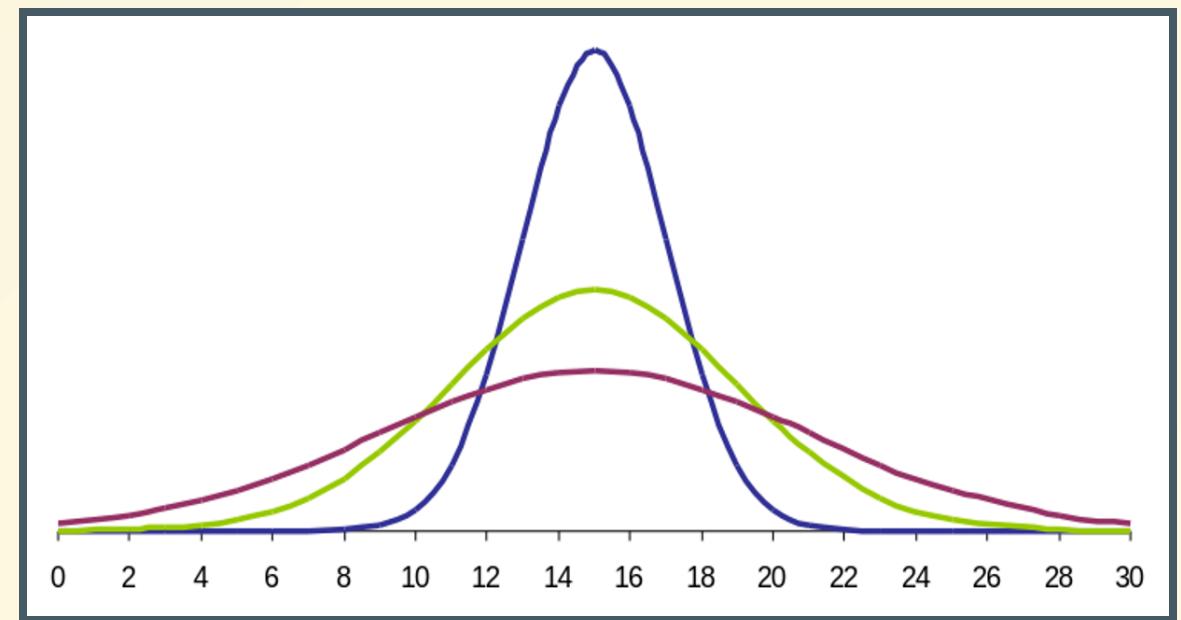
Exercise #5

- ? Which curve has the larger standard deviation?
- a) blue
 - b) green
 - c) red



Exercise #5 -- Solution

- ? Which curve has the larger standard deviation?
- a) blue
 - b) green
 - c) red 



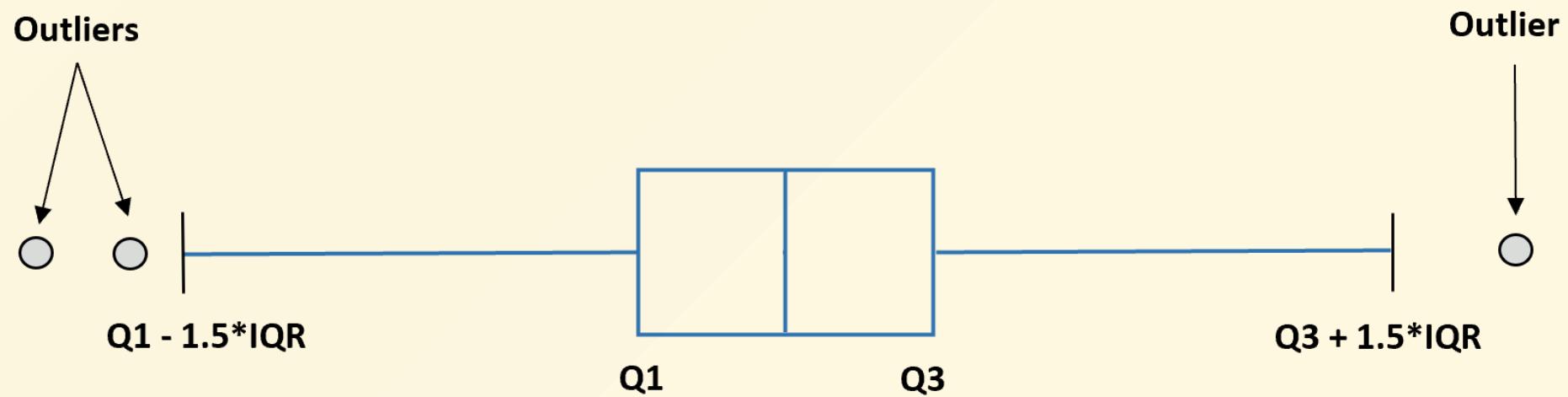
Exercise #6

- ? Researchers have collected age, sex, lipid levels
How should they summarise their data?
- a) Age: mean (SD), sex: N (%), lipid levels: mean (SD)
 - b) Age: median (IQR), sex: N (%), lipid levels: median (IQR)
 - c) Age: mean (SD), sex: mean (SD), lipid levels: mean (SD)
 - d) Either a) or b)
 - e) Either a) or c)

Exercise #6 -- Solution

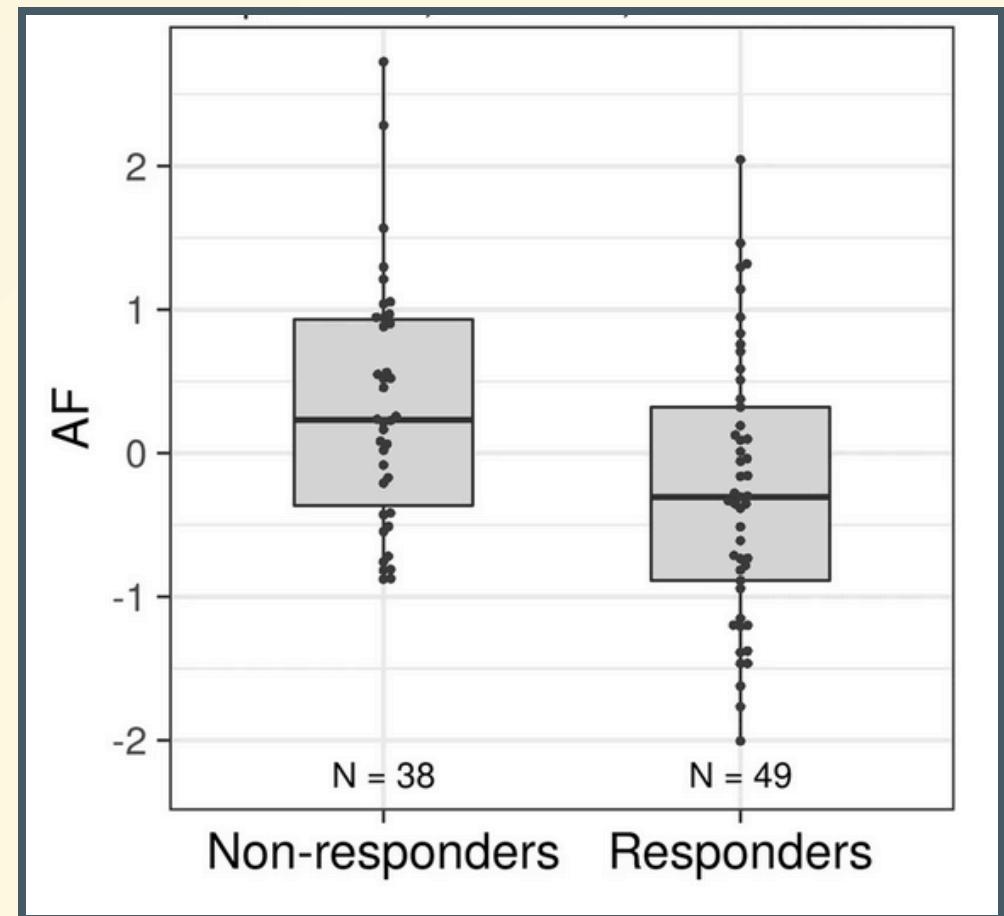
- ? Researchers have collected age, sex, lipid levels
How should they summarise their data?
- a) Age: mean (SD), sex: N (%), lipid levels: mean (SD)
 - b) Age: median (IQR), sex: N (%), lipid levels: median (IQR)
 - c) Age: mean (SD), sex: mean (SD), lipid levels: mean (SD)
 - d) Either a) or b)
 - e) Either a) or c)

Data visualisation: boxplots

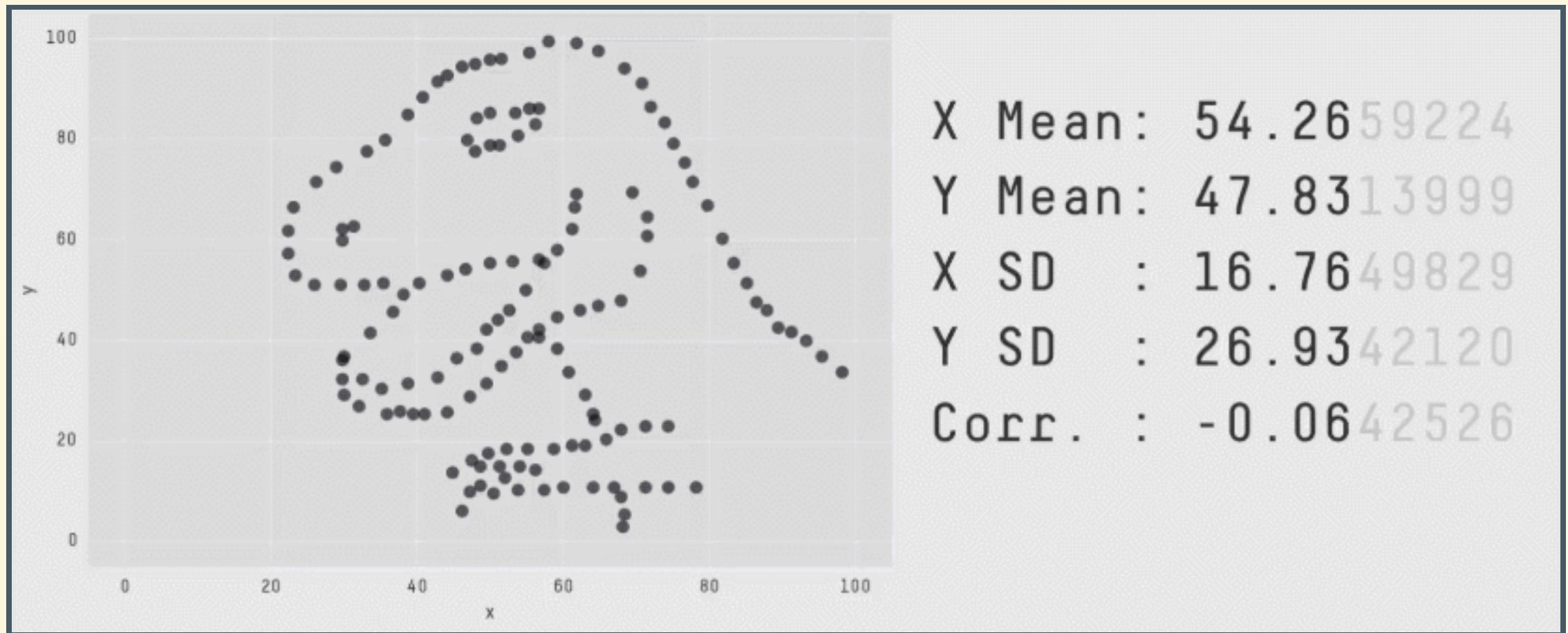


Boxplots in the wild

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
- Data are described with measures of centrality and dispersion
- Visualising your data is always a good idea

See you tomorrow

