Introduction to Medical Statistics 2026

Oxford University Clinical Research Unit
March 23-27, 2026

Ronald Geskus and the biostatistics crew
Oxford University Clinical Research Unit
Hospital for Tropical Diseases,
Ho Chi Minh City, Viet Nam

Part I

Data; Descriptive Analysis

Ronald Geskus

Contents:

1. Data characteristics
2. Variable types
3. Variable summaries: numerical and graphical
4. Introduction to R and RStudio

Outline

Outline

Some definitions

Data set: observations from a study, typically in rectangular format

Observation: set of measurements on a unit (patients, animals, farms) at one specific time (rows)

Variable: measurable characteristic; its value typically varies over observations (columns)

Example

Trial in patients with dengue shock

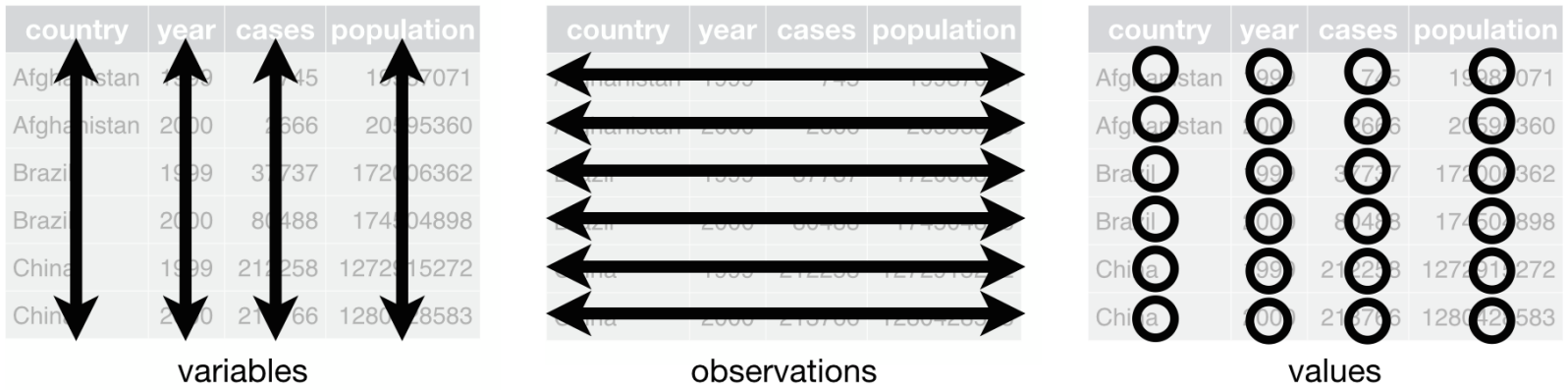
SUBJID	fluid	Age	Sex	Hct	PLT	hospdays	Reshock
01-0007	Colloid	5	F	49.0	24.0	7	No
01-0008	Blood	11	F	54.0	27.0	8	No
01-0009	Colloid	8	M	43.0	47.0	6	No
01-0010	Crystalloid	15	F	45.5	18.9	7	No
01-0011	Crystalloid	13	M	49.0	24.0	4	No
01-0012	Colloid	8	M	40.0	34.0	3	No

Rows: observations, one row per patient (each with different SUBJID)

Columns: variables

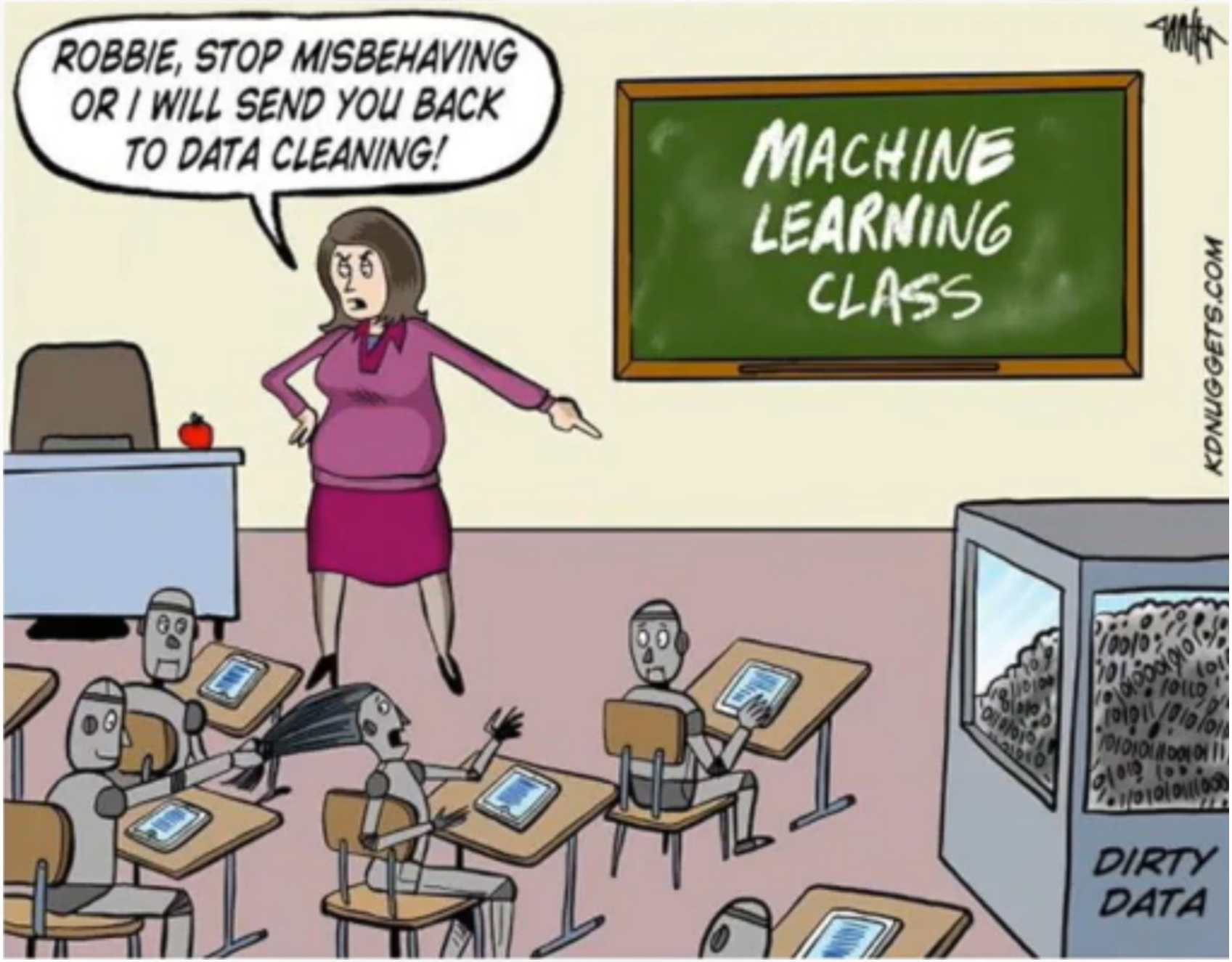
Cells: values

Data structure



Tidy data

- $\approx$ 80% of analysis time spent on data cleaning and preparation, especially when data are “messy”


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

Tidy data

- $\approx$ 80% of analysis time spent on data cleaning and preparation, especially when data are "messy"
- Tidy data: link structure with meaning
  - each variable is a column; each column is a variable
  - each observation is a row; each row is an observation
  - each value is a cell; each cell is a single value
  - different types of observation in different tables (patient characteristics, lab results, visits, adverse events, . . .)
  - if data is spread over multiple tables, then each table should include an indentifier column that allows them to be linked

See `https://r4ds.hadley.nz/data-tidy.html`

Tidy and messy data

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Compare with messy format:

```
country      year type            count
Afghanistan  1999 cases             745
Afghanistan  1999 population  19987071
Afghanistan  2000 cases            2666
Afghanistan  2000 population  20595360
Brazil       1999 cases           37737
Brazil       1999 population 172006362
  with 6 more rows
```

```
country      year rate
Afghanistan  1999 745/19987071
Afghanistan  2000 2666/20595360
Brazil       1999 37737/172006362
Brazil       2000 80488/174504898
China        1999 212258/1272915272
China        2000 213766/1280428583
```

Tidy data

- $\approx$ 80% of analysis time spent on data cleaning and preparation, especially when data are "messy"
- Tidy data: link structure with meaning
  - each variable is a column; each column is a variable
  - each observation is a row; each row is an observation
  - each value is a cell; each cell is a single value
  - different types of observation in different tables (patient characteristics, lab results, visits, adverse events, . . .)
  - if data is spread over multiple tables, then each table should include an indentifier column that allows them to be linked

  See `https://r4ds.hadley.nz/data-tidy.html`
- Suggestion for column order in data table
  - first: variables fixed by design (e.g. SUBJID)
  - next: measured variables

Tidy data: advantages

- Works smoothly with statistical software
- Makes visualization easier
- Simplifies modeling
- Reduces data cleaning errors

Dataset; further suggestions

- Variable names: informative and concise; don't use spaces in names
  - Example: `HospDays` ("CamelCase"), `hosp.days` or `hosp_days`
  - be consistent in naming and use of capitals

Example

Trial in patients with dengue shock

SUBJID	fluid	Age	Sex	Hct	PLT	hospdays	Reshock
01-0007	Colloid	5	F	49.0	24.0	7	No
01-0008	Blood	11	F	54.0	27.0	8	No
01-0009	Colloid	8	M	43.0	47.0	6	No
01-0010	Crystalloid	15	F	45.5	18.9	7	No
01-0011	Crystalloid	13	M	49.0	24.0	4	No
01-0012	Colloid	8	M	40.0	34.0	3	No

Rows: observations, one row per patient (each with different SUBJID)

Columns: variables

Cells: values

Inconsistent naming of variables!

Dataset; further suggestions

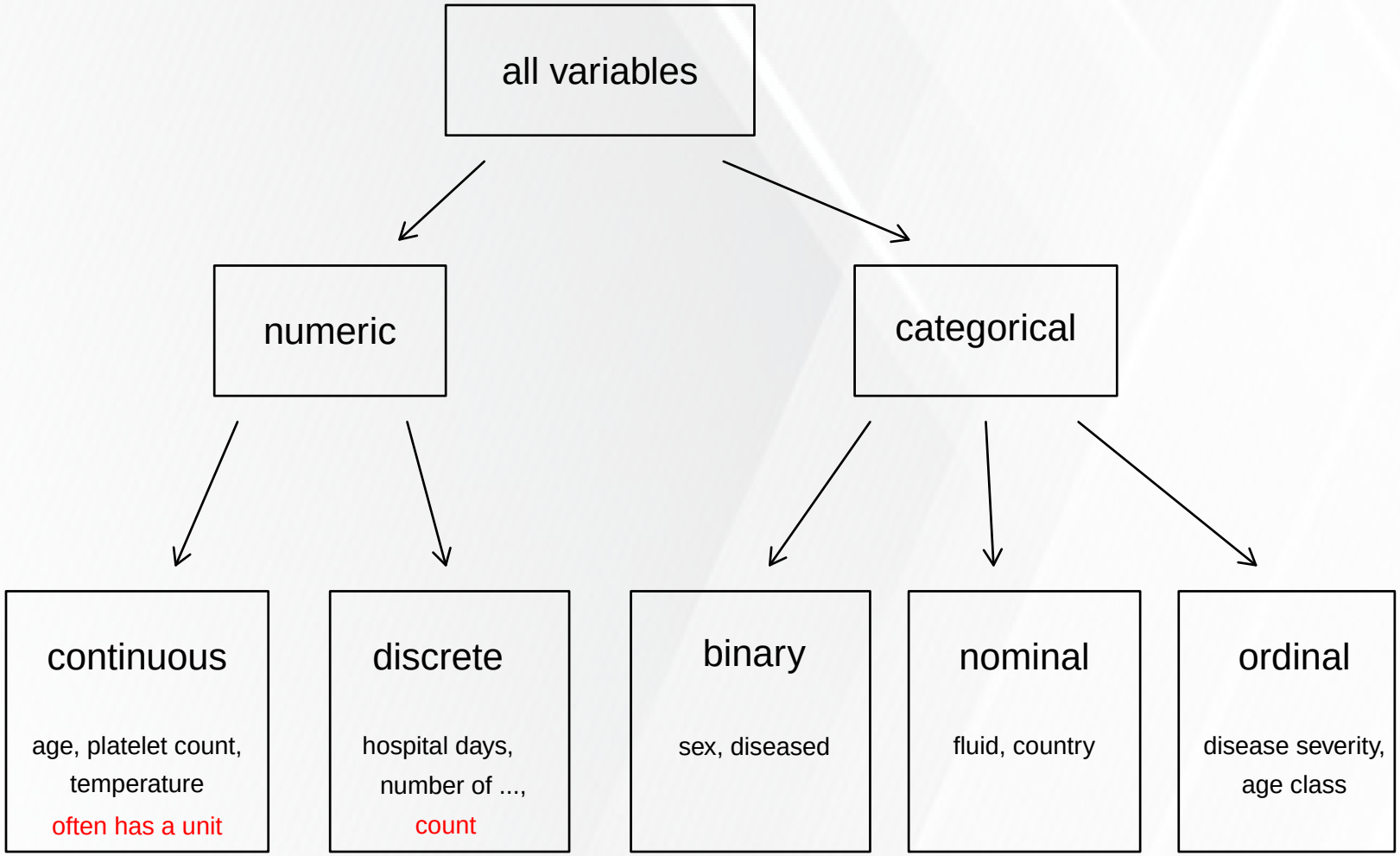
- Variable names: informative and concise; don't use spaces in names
  - Example: `HospDays` ("CamelCase"), `hosp.days` or `hosp_days`
  - be consistent in naming and use of capitals
- Getting data into statistical software
  - best: import data from database (MS Access, SQLite)
  - if data in Excel format: save as comma-separated file (.csv) before import into R.
    "*Excel is the devil - if it is used for anything else but as scratchbook or for data transfer (and even then!)*"
    (former PhD student of Ronald)
    a recent blunder due to use of Excel:
    `https://www.bbc.com/news/technology-54423988`

Outline

Types of variables; example

SUBJID	fluid	Age	Sex	Hct	PLT	hospdays	Reshock
01-0007	Colloid	5	F	49.0	24.0	7	No
01-0008	Blood	11	F	54.0	27.0	8	No
01-0009	Colloid	8	M	43.0	47.0	6	No
01-0010	Crystalloid	15	F	45.5	18.9	7	No
01-0011	Crystalloid	13	M	49.0	24.0	4	No
01-0012	Colloid	8	M	40.0	34.0	3	No

Types of variables



Types of variables; example

SUBJID	fluid	Age	Sex	Hct	PLT	hospdays	Reshock
01-0007	Colloid	5	F	49.0	24.0	7	No
01-0008	Blood	11	F	54.0	27.0	8	No
01-0009	Colloid	8	M	43.0	47.0	6	No
01-0010	Crystalloid	15	F	45.5	18.9	7	No
01-0011	Crystalloid	13	M	49.0	24.0	4	No
01-0012	Colloid	8	M	40.0	34.0	3	No

- Continuous: Age, Hct, PLT
- Discrete: hospdays (maybe Age)
- Binary: Sex, Reshock
- Nominal: fluid (few levels), SUBJID (many levels)

Coding variable values

- Categorical variables often coded as numbers
  - e.g. 1=male, 2=female
    1=Vietnam, 2=Thailand, 3=Laos
  - BUT: this does not make them numeric!
  - be aware of this when doing the analyses
    Or better: create character values from the start

A paper retracted

JAMA, 2018:

The identified programming error . . . occurred while the variable referring to the study "arm" (ie, group) assignment was recoded. The purpose of the recoding was to change the randomization assignment variable format of "1, 2" to a binary format of "0, 1." However, the assignment was made incorrectly and resulted in a reversed coding of the study groups. Even though the data analyst created and conducted some test analysis programs, they were of the type that did not show any labeling of the arm categories, only the "arm" variable in a regression.

Coding variable values

- Categorical variables often coded as numbers
  - e.g. 1=male, 2=female
    1=Vietnam, 2=Thailand, 3=Laos
  - BUT: this does not make them numeric!
  - be aware of this when doing the analyses
    Or better: create character values from the start
- Missing data
  - use special code (in R: NA “not available”).
    don’t use values like 999.
  - always report amount of missingness per variable
  - often excluded from analysis (but this may bias results)

Henri van Werkhoven
@hvwerkhoven

The strength of our research is that we didn't have any missing data after deleting the incomplete records. (Student's paper)

3:29 am · 7/10/21 · Twitter for Android

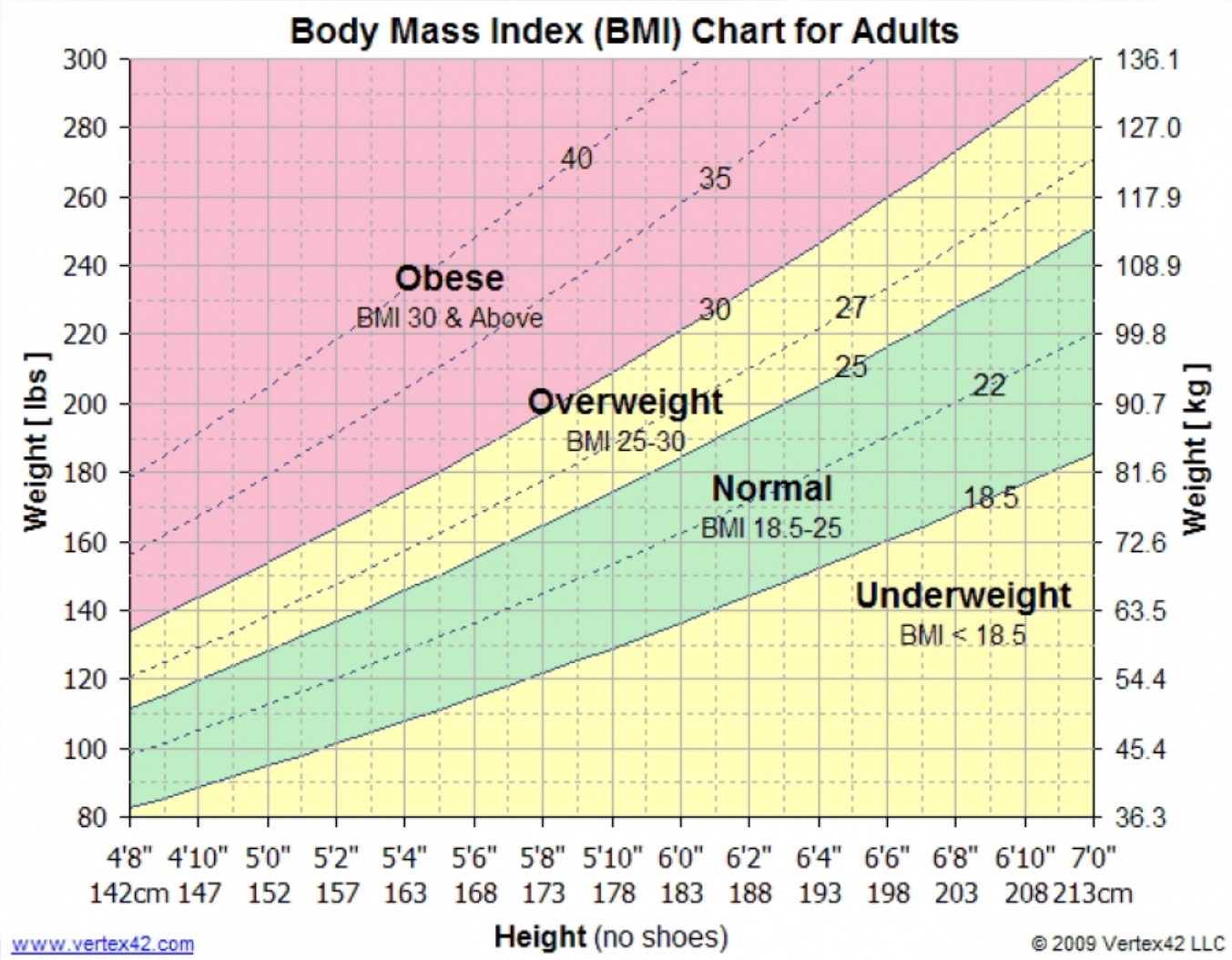
4 Retweets **1** Quote Tweet **38** Likes

Derived variables

Variables often transformed to simplify display or analysis:

- Categorise numerical variables
  - by quantiles, logical or rounded values: age groups ($<$ 18, 18-30, 30-50, $>$ 50)
  - by accepted threshold values: BMI (underweight, normal, overweight, obese), hypertension (blood pressure > 90 mmHg)

Derived variables - Body Mass Index = weight / height^2



oucru

Derived variables

Variables often transformed to simplify display or analysis:

- Categorise numerical variables
  - by quantiles, logical or rounded values: age groups ($<$ 18, 18-30, 30-50, $>$ 50)
  - by accepted threshold values: BMI (underweight, normal, overweight, obese), hypertension (blood pressure > 90 mmHg)
- Population reference standards
  - child growth curves (standard deviation scores)

Derived variables

Variables often transformed to simplify display or analysis:

- Categorise numerical variables
  - by quantiles, logical or rounded values: age groups ($<$ 18, 18-30, 30-50, $>$ 50)
  - by accepted threshold values: BMI (underweight, normal, overweight, obese), hypertension (blood pressure > 90 mmHg)
- Population reference standards
  - child growth curves (standard deviation scores)
- Transform data for statistical reasons
  - log transform skewed data

Exposure and outcome variables

Often we want to investigate relationships between variables in a specific direction: from exposure to outcome

- Outcomes are the variables we want to know more about
- Exposures are the variables we think might explain the variation in outcomes
- Statistics: quantify the strength of relationship between outcomes and exposures

Spot the exposure / outcome?

- Example 1 - Thwaites GE et al. NEJM 2004; 351: 17
  Study aim:- To determine whether adjunctive treatment with dexamethasone reduced the risk of death or severe disability after nine months of follow-up

Spot the exposure / outcome?

- Example 2 - Watson M et al. BMJ 2005; 330: 178
  Study aim:- To assess the effectiveness of safety equipment in reducing unintentional injuries for families with children aged under 5 years

Outline

ChatGPT 5.2: "Can you give me the best way to summarize variables, stratified by type of variable"

Variable Type	Best Numerical Summary	Best Graph
Continuous (Normal)	Mean ($\pm$SD)	Histogram
Continuous (Skewed)	Median (IQR)	Boxplot
Nominal	n (%)	Bar chart
Ordinal	n (%)	Ordered bar chart
Binary	n (%)	Bar chart

In research reports (Typical format "Table 1"):

- continuous normal $\rightarrow$ Mean $\pm$ SD
- continuous skewed $\rightarrow$ Median (IQR)
- caegorical $\rightarrow$ n (%)

Outline

Presentation

- Number of observations per value of variable
- Relative frequency of each value
  - percentage: between 0% and 100%
  - proportion: between 0 and 1

Presentation

- Number of observations per value of variable
- Relative frequency of each value
  - percentage: between 0% and 100%
  - proportion: between 0 and 1
- Example: fluid in patients with dengue shock

fluid	number (percent)
Crystalloid	591 (65%)
Colloid	226 (25%)
Blood	95 (10%)
Total	912 (100%)

Presentation

- Number of observations per value of variable
- Relative frequency of each value
  - percentage: between 0% and 100%
  - proportion: between 0 and 1
- Example: fluid in patients with dengue shock

fluid	number (percent)
Crystalloid	591 (65%)
Colloid	226 (25%)
Blood	95 (10%)
Total	912 (100%)

- Category values often called levels

Outline

Location and spread

Data: $x_1, x_2, \ldots, x_n$

- Location: (arithmetic) mean, geometric mean; median
- Spread (dispersion, variation): standard deviation; quartiles; (interquartile) range

Location: arithmetic mean

- Add up all the values and divide this sum by the number of values
  - e.g. 5 patients, age: 25, 63, 22, 75, 20
  - mean age: 205/5=41 years
- General formula

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Usually just called “mean”

Location: median

- Middle value after ordering
  - age: 25, 63, 22, 75, 20
  - ordered: 20, 22, 25, 63, 75
  - median age: 25 years

  Even sample size: halfway between the two “middle” observations
- Median splits sample into two halves: one half is lower, other half is larger than median

Location: mean versus median

- Mean
  - preferred if distribution $\approx$ symmetric, without long tails or extreme values (outliers)
    Classical example: normal distribution (later in course)
  - most statistical analyses are based on the mean value

Location: mean versus median

- Mean
  - preferred if distribution $\approx$ symmetric, without long tails or extreme values (outliers)
    Classical example: normal distribution (later in course)
  - most statistical analyses are based on the mean value
- Median
  - close to mean if distribution $\approx$ symmetric
  - smaller than mean if distribution "skewed to the right"
    e.g. 20, 22, 25, 63, 75 (mean 41; median 25)
  - insensitive to outliers
  - not informative for discrete variable with few different values (see handouts)

Mean versus median, clinical relevance

- Example: length of hospital stay after being admitted with SARS-CoV-2 infection
- Assume distribution skewed to the right
  Median is 8 days, mean 19 days, and 1% stays longer than 4 weeks
- Mean or median more relevant if
  - you are a patient admitted to hospital?
  - you are a hospital administrator interested in costs?

Mean versus median, clinical relevance

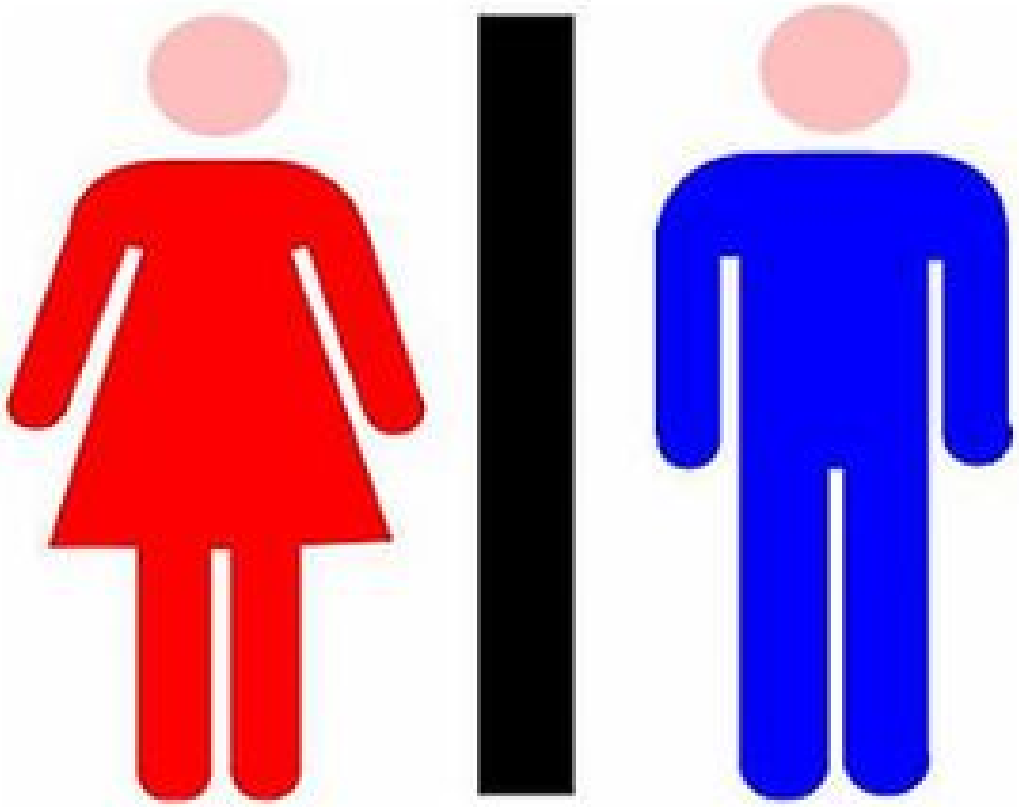
- Example: length of hospital stay after being admitted with SARS-CoV-2 infection
- Assume distribution skewed to the right
  Median is 8 days, mean 19 days, and 1% stays longer than 4 weeks
- Mean or median more relevant if
  - you are a patient admitted to hospital?
  - you are a hospital administrator interested in costs?

Answer: for a hospital administrator, the few patients that stay very long do matter because they may have a large impact on costs. A patient is more interested in the more frequent values of hospital stay and after how many days 50% is discharged.

"A statistician can have his head in an oven and his feet in ice, and he will say that on the average he feels fine."

STATISTICS

THE DISCIPLINE THAT PROVES THE AVERAGE HUMAN HAS ONE TESTICLE



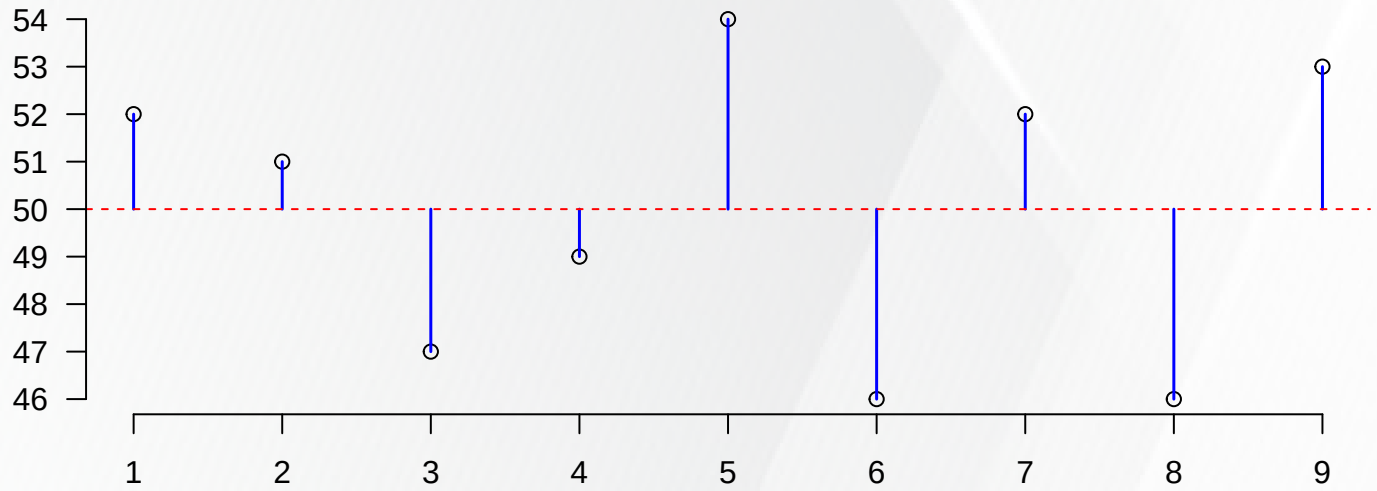
Spread (dispersion, variation)

- Example: age
  - 60, 10, 70, 20, 40, 50, 30, 80, 90
  - 52, 51, 47, 49, 54, 46, 52, 46, 53
- Both mean 50, but variation around mean larger in first row

Measure spread I: variance/standard deviation

Deviation (from mean)

value	52	51	47	49	54	46	52	46	53	mean:50
deviation	2	1	-3	-1	4	-4	2	-4	3	sum: 0



- Positive and negative deviations from mean always cancel out

Better measure: "standard deviation"

	deviation	absolute deviation	squared deviation
	2	2	4
	1	1	1
	-3	3	9
	-1	1	1
	4	4	16
	-4	4	16
	2	2	4
	-4	4	16
	3	3	9
sum	0	24	76

$$\textbf{variance} = \frac{\text{sum of squared deviations}}{n-1} = \frac{76}{9-1} = 9.5$$

$$\textbf{sd} = \sqrt{9.5} \qquad \textbf{sd}\text{: standard deviation}$$

Measure spread I: variance/standard deviation

- Variance: square each deviation from mean, and average

$$\text{variance} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

(n.b. use of “n-1” is for statistical properties)

- Standard deviation (sd): square root of variance

$$\text{sd} = \sqrt{\text{variance}}$$

Measure spread II: range, interquartile range

- Range: distance from minimum to maximum value

Measure spread II: range, interquartile range

- Range: distance from minimum to maximum value
- x-quantile: value below which fraction x of the values is located
  x-percentile: value below which x% of the values is located
- Special case: quartiles
  - first quartile q_1 : 0.25-quantile/25-percentile
  - third quartile q_3: 0.75-quantile/75-percentile
  - note: median is second quartile
- Interquartile range (IQR): $q_3 - q_1$
  - note: (q_1, q_3) often called IQR, but formally not correct

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20
   same median and IQR

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20
   same median and IQR
2. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 8, 9

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20
   same median and IQR
2. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 8, 9
   same median, (b) larger IQR

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20
   same median and IQR
2. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 8, 9
   same median, (b) larger IQR
3. (a) 1, 2, 3, 4, 5
   (b) 6, 7, 8, 9, 10

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20
   same median and IQR
2. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 8, 9
   same median, (b) larger IQR
3. (a) 1, 2, 3, 4, 5
   (b) 6, 7, 8, 9, 10
   (b) larger median, same IQR

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20
   same median and IQR
2. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 8, 9
   same median, (b) larger IQR
3. (a) 1, 2, 3, 4, 5
   (b) 6, 7, 8, 9, 10
   (b) larger median, same IQR
4. (a) 0, 10, 50, 60, 100
   (b) 0, 100, 500, 600, 1000

Exercise: medians and IQRs

Compare distributions (a) and (b) based on their medians and IQRs.

1. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 7, 20
   same median and IQR
2. (a) 3, 5, 6, 7, 9
   (b) 3, 5, 6, 8, 9
   same median, (b) larger IQR
3. (a) 1, 2, 3, 4, 5
   (b) 6, 7, 8, 9, 10
   (b) larger median, same IQR
4. (a) 0, 10, 50, 60, 100
   (b) 0, 100, 500, 600, 1000
   (b) both median and IQR larger

STATISTICIANS ARE MEAN AND SLIGHTLY DEVIANT

Measures of spread: comparison

- Standard deviation (sd)
  - Useful if data $\approx$ symmetric. If $\approx$ normal distribution then
    - $\approx$ 68% of observations lie between $\overline{x} \pm \text{sd}$
    - $\approx$ 95% of observations lie between $\overline{x} \pm 2 \times \text{sd}$
  - difficult to interpret for skewed data

	mean	sd	median	(q_1, q_3)	(min, max)
hospdays	5.1	3.2	4	(3, 6)	(1, 43)

$\overline{x} - 2 \times \text{sd}$ gives negative value: 5.1-6.4=-1.3

  - sensitive to outliers
- Quartiles or IQR
  - can always be used to quantify spread

oucru

UNIVERSITY OF OXFORD

wellcome

Measures of spread: comparison

- Standard deviation (sd)
  - Useful if data $\approx$ symmetric. If $\approx$ normal distribution then
    - $\approx$ 68% of observations lie between $\overline{x} \pm \text{sd}$
    - $\approx$ 95% of observations lie between $\overline{x} \pm 2 \times \text{sd}$
  - difficult to interpret for skewed data

	mean	sd	median	(q_1, q_3)	(min, max)
hospdays	5.1	3.2	4	(3, 6)	(1, 43)

$\overline{x} - 2 \times \text{sd}$ gives negative value: 5.1-6.4=-1.3

  - sensitive to outliers
- Quartiles or IQR
  - can always be used to quantify spread
- Minimum, maximum, range
  - not best measure; range increases with sample size
  - only used as additional information

Outline

ChatGPT 5.2: "Can you give me the best way to summarize variables, stratified by type of variable"

Variable Type	Best Numerical Summary	Best Graph
Continuous (Normal)	Mean ($\pm$SD)	Histogram
Continuous (Skewed)	Median (IQR)	Boxplot
Nominal	n (%)	Bar chart
Ordinal	n (%)	Ordered bar chart
Binary	n (%)	Bar chart

In research reports (Typical format "Table 1"):

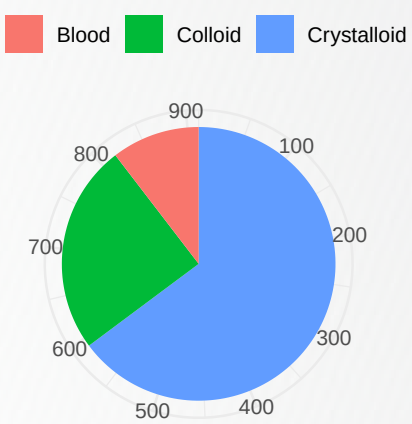
- continuous normal $\rightarrow$ Mean $\pm$ SD
- continuous skewed $\rightarrow$ Median (IQR)
- caegorical $\rightarrow$ n (%)

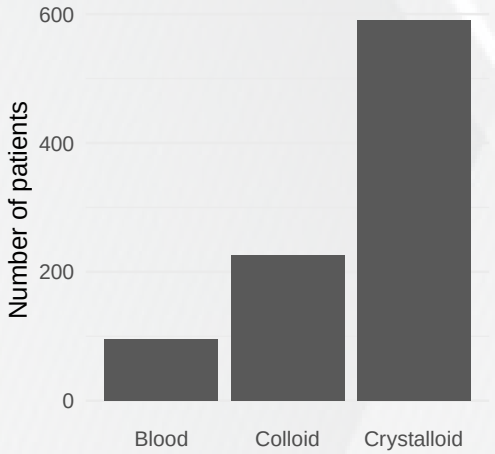
Main graph types for a single variable

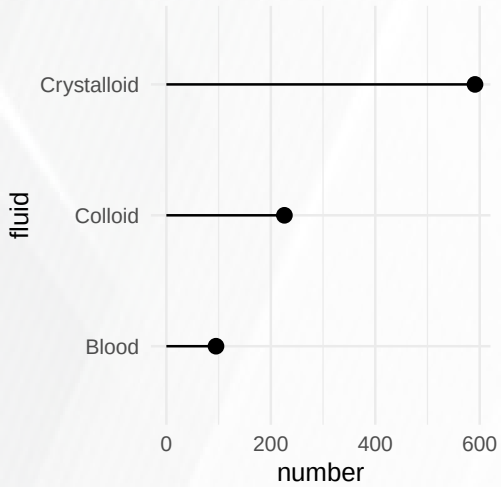
- Categorical (percentage/frequency)
  - Pie chart
  - Bar chart
  - Dotplot

  Often little added value compared to numerical summary (for single variable)

Pie, bar and dotplot







Main graph types for a single variable

- Categorical (percentage/frequency)
  - Pie chart (not recommended)
  - Bar chart
  - Dotplot (often preferred over bar chart)

  Often little added value compared to numerical summary (for single variable)

Main graph types for a single variable

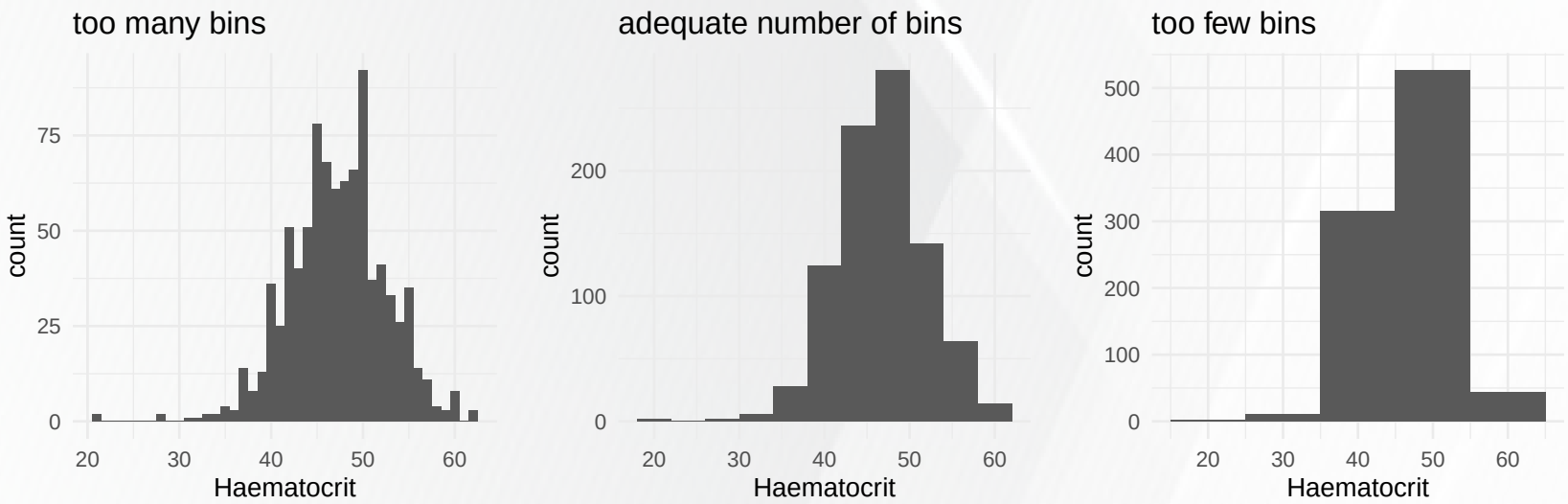
- Categorical (percentage/frequency)
  - Pie chart (not recommended)
  - Bar chart
  - Dotplot (often preferred over bar chart)

  Often little added value compared to numerical summary (for single variable)
- Continuous
  - Histogram, frequency polygon, density, ridgeplot
  - Boxplot, violin plot, raincloud plot
  - Cumulative frequency ("empirical cumulative distribution function", ECDF)

Histogram

- Group values of a variable into bins of equal width; plot number in each bin as barchart
- Problem: visual appearance may depend on the chosen number and location of bins
  - choosing too many bins shows noise, choosing too few hides relevant details
  - try several choices of the number of bins

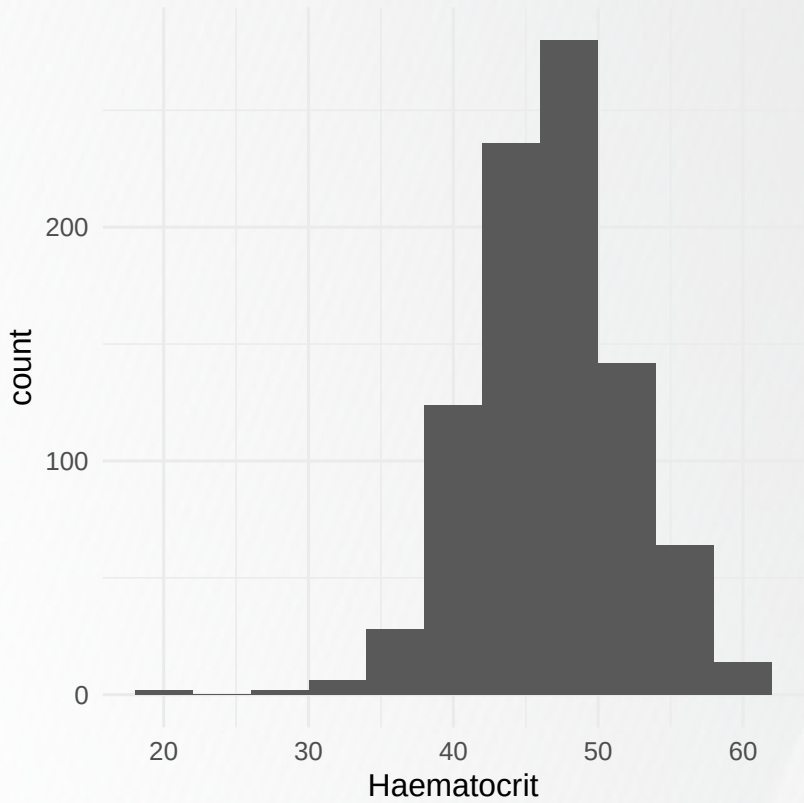
Histograms: different chosen binwidths

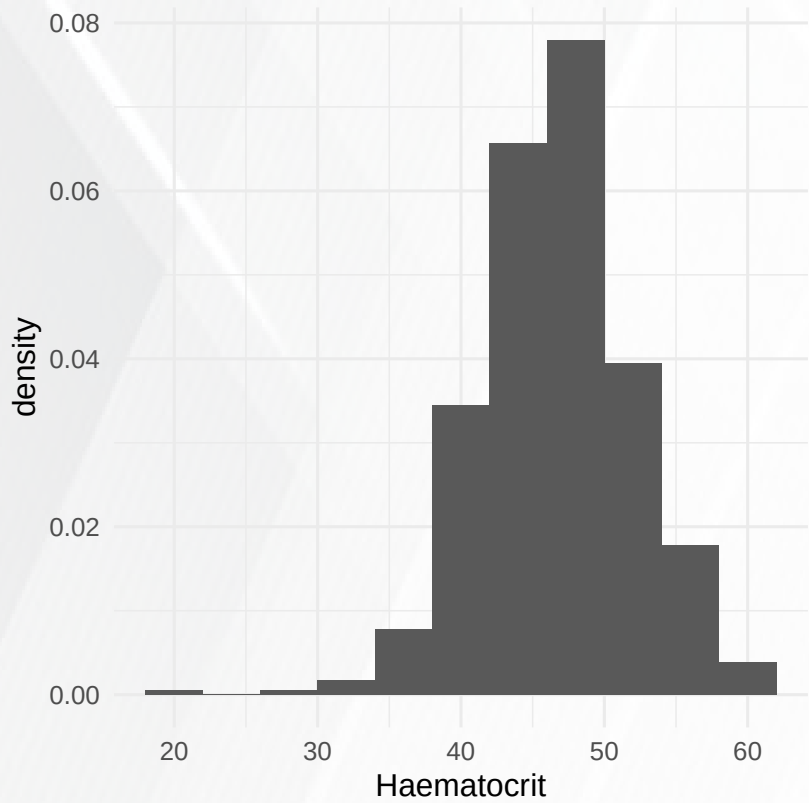


Histogram

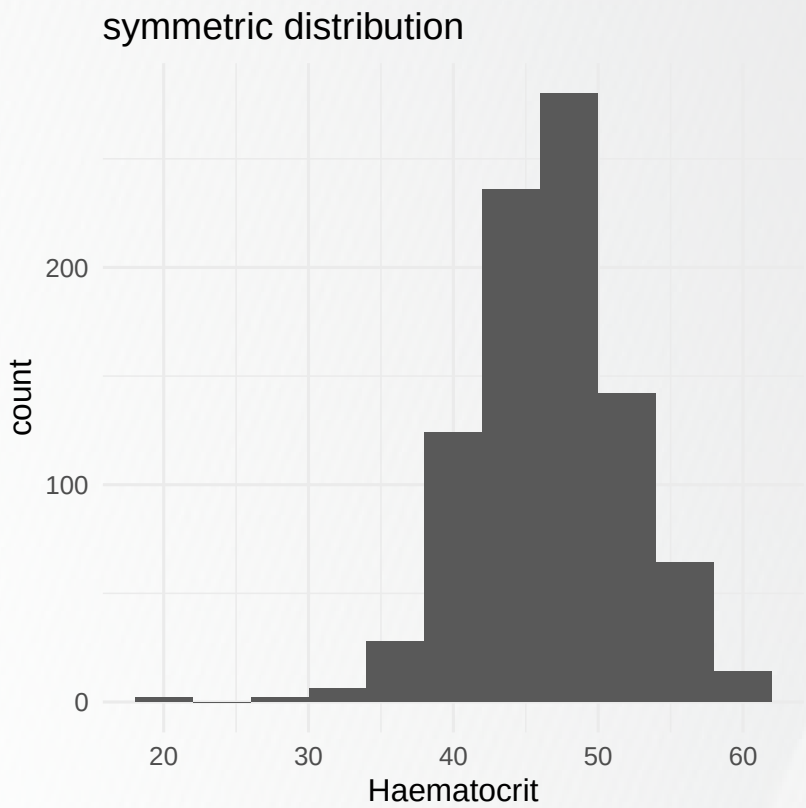
- Group values of a variable into bins of equal width; plot number in each bin as barchart
- Problem: visual appearance may depend on the chosen number and location of bins
  - choosing too many bins shows noise, choosing too few hides relevant details
  - try several choices of the number of bins
- Alternative along y-axis: use standardized value instead of count
  - height of bar$\times$binwidth (=bar area) represents relative frequency of that grouped value
    areas sum up to 1 (or 100%)

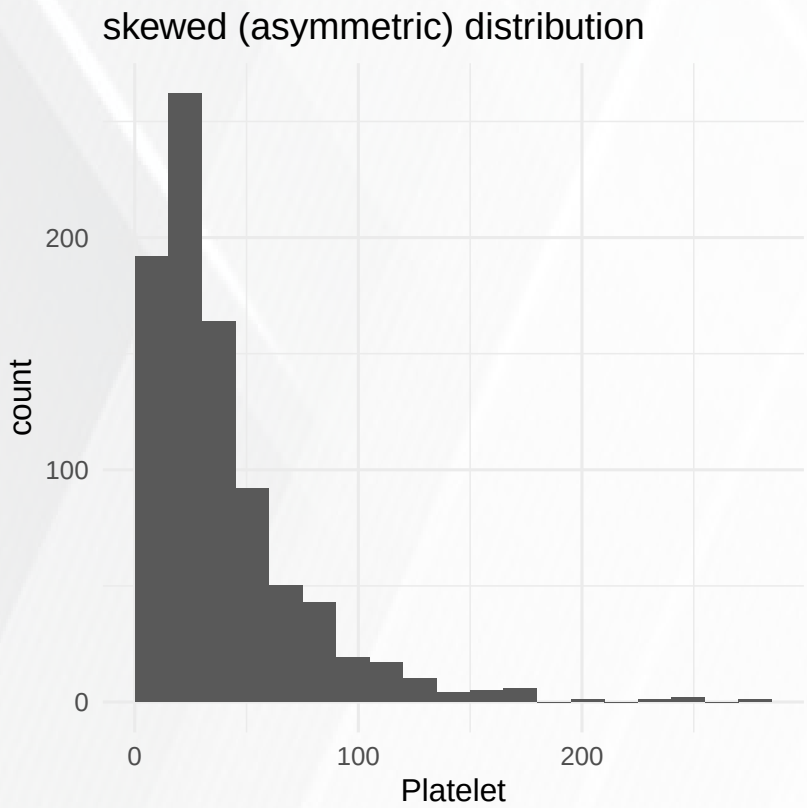
Histogram: count versus frequency/area



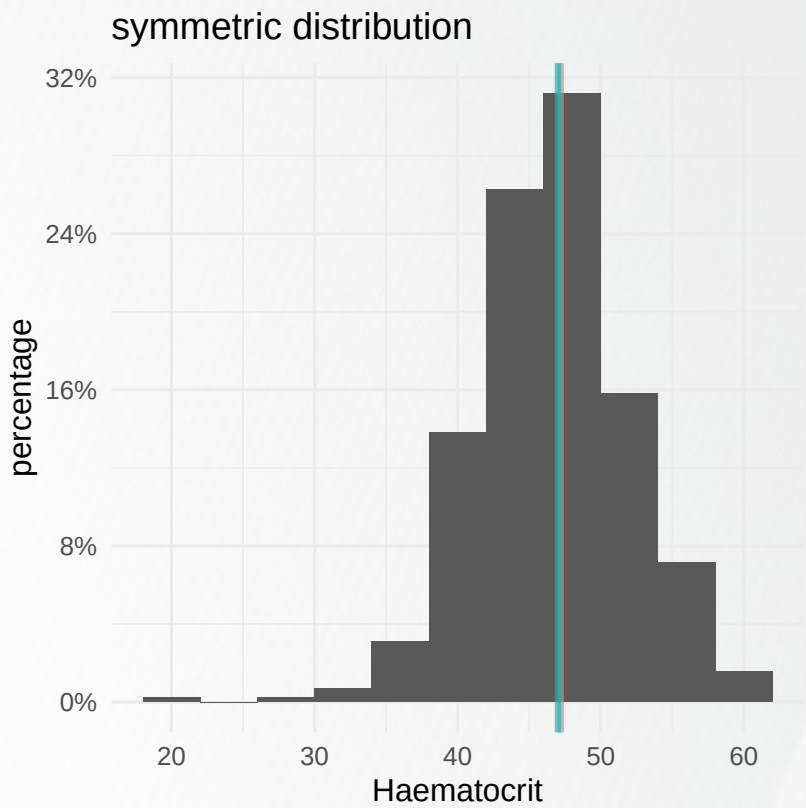


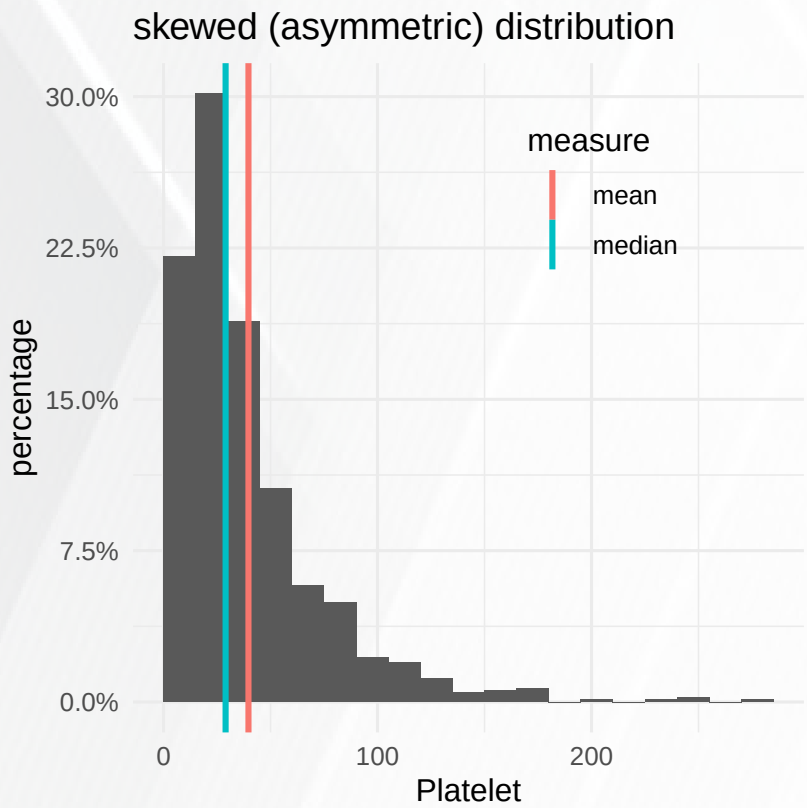
Histogram: distributions with different shape





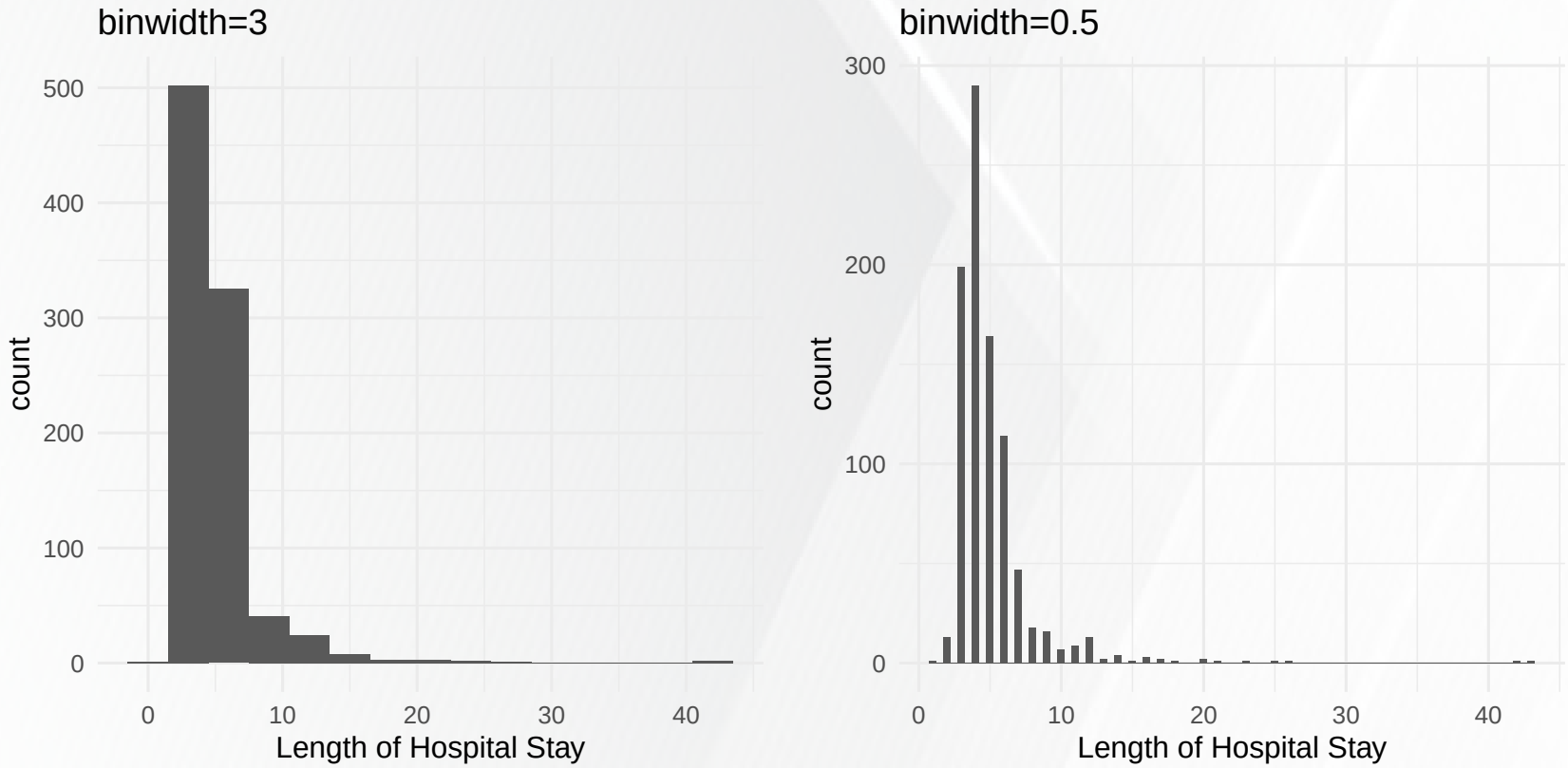
Histogram: relation with location measures



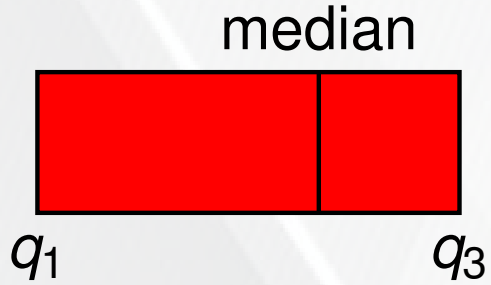


Histogram: discrete numeric variable

Make each value a separate bar

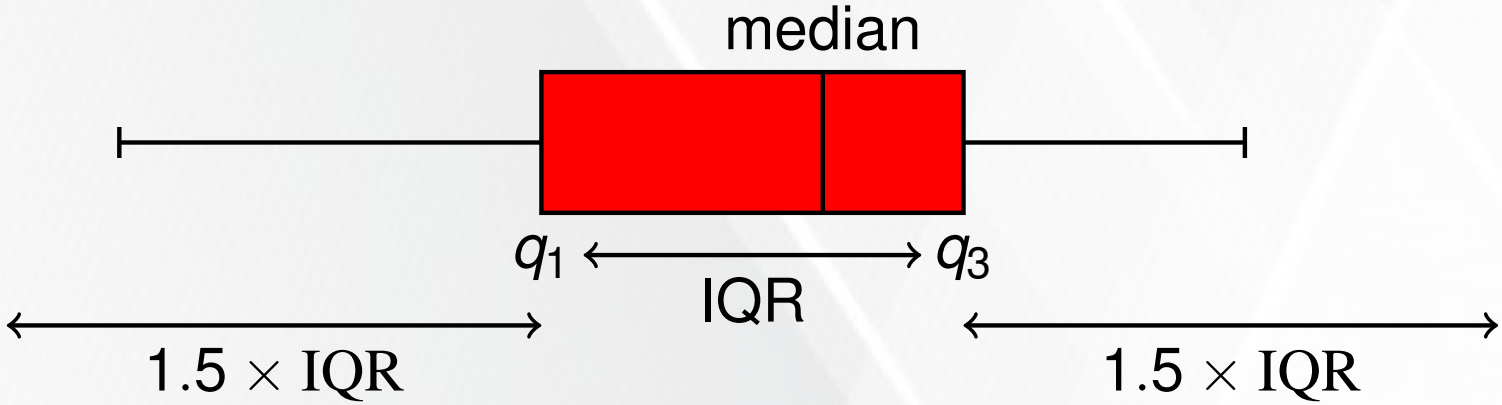


Boxplot



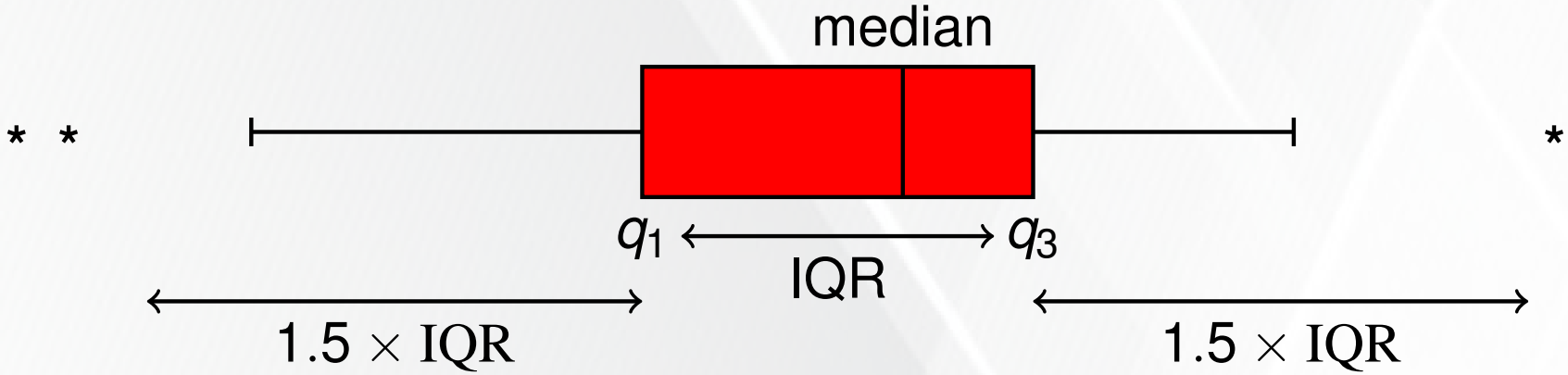
- Formal name: box-and-whisker plot
- Box (red area): most common observations

Boxplot



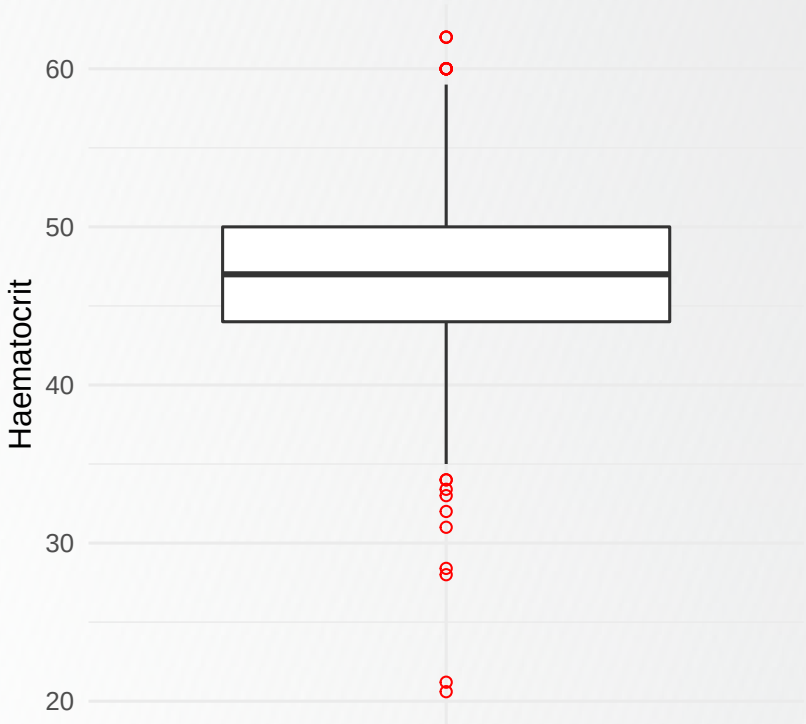
- Formal name: box-and-whisker plot
- Box (red area): most common observations
- Whiskers: less common but still typical
  - from 1st and 3rd quartile to the furthest away observation, but still
    $\geq q_1 - 1.5 \times \text{IQR}$ and $\leq q_3 + 1.5 \times \text{IQR}$

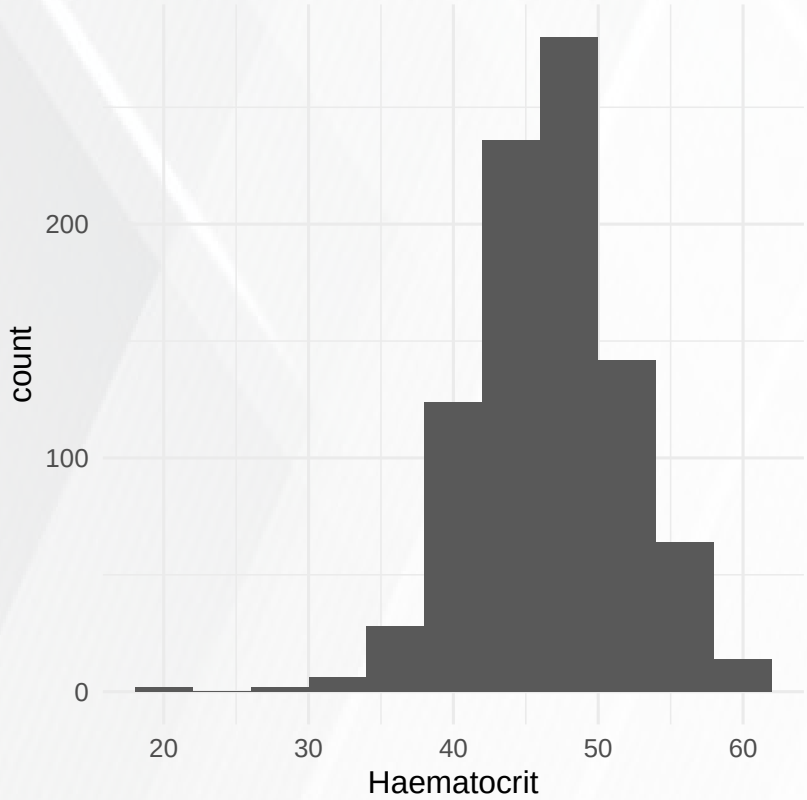
Boxplot



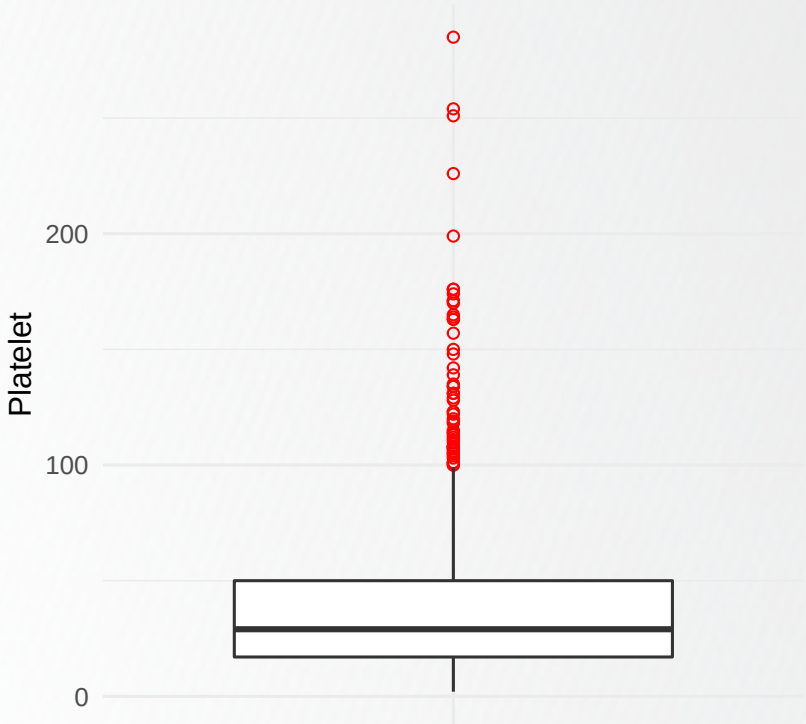
- Formal name: box-and-whisker plot
- Box (red area): most common observations
- Whiskers: less common but still typical
  - from 1st and 3rd quartile to the furthest away observation, but still
    $\geq q_1 - 1.5 \times \text{IQR}$ and $\leq q_3 + 1.5 \times \text{IQR}$
- Outliers
  - All points outside of the whiskers

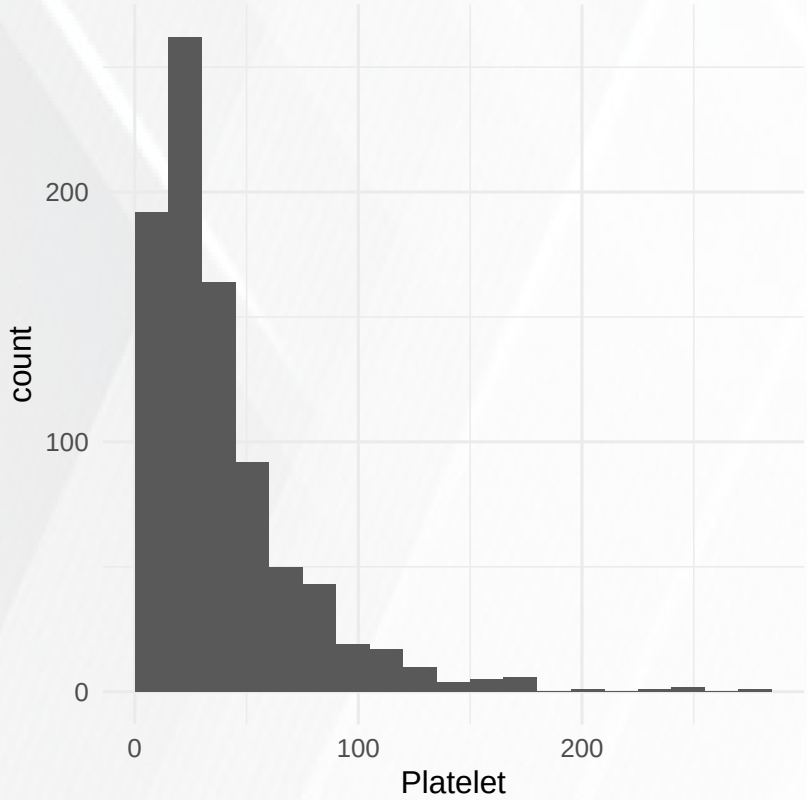
Boxplot versus histogram: symmetric distribution



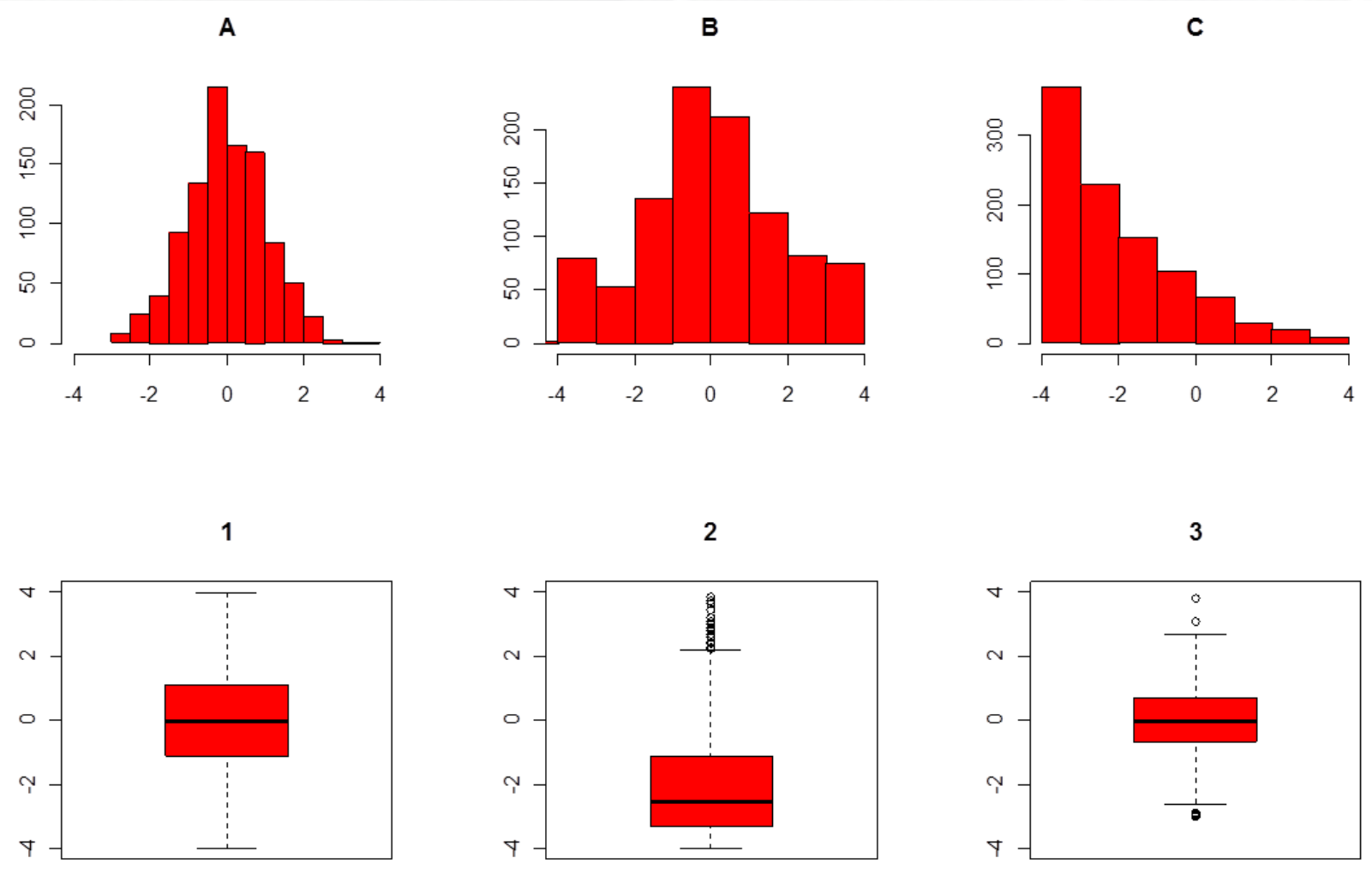


Boxplot versus histogram: asymmetric distribution

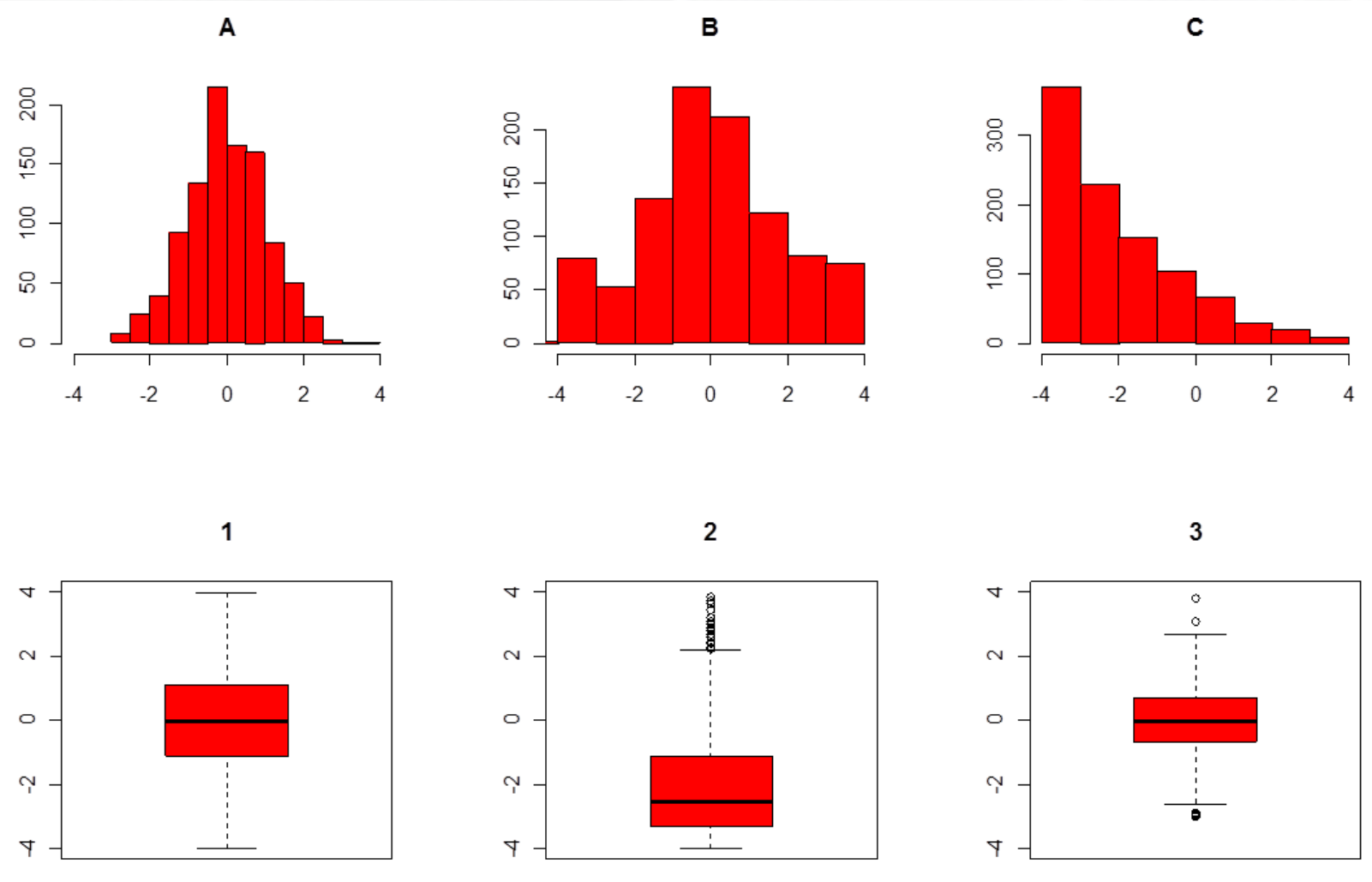




Quiz: match histogram and boxplots



Quiz: match histogram and boxplots

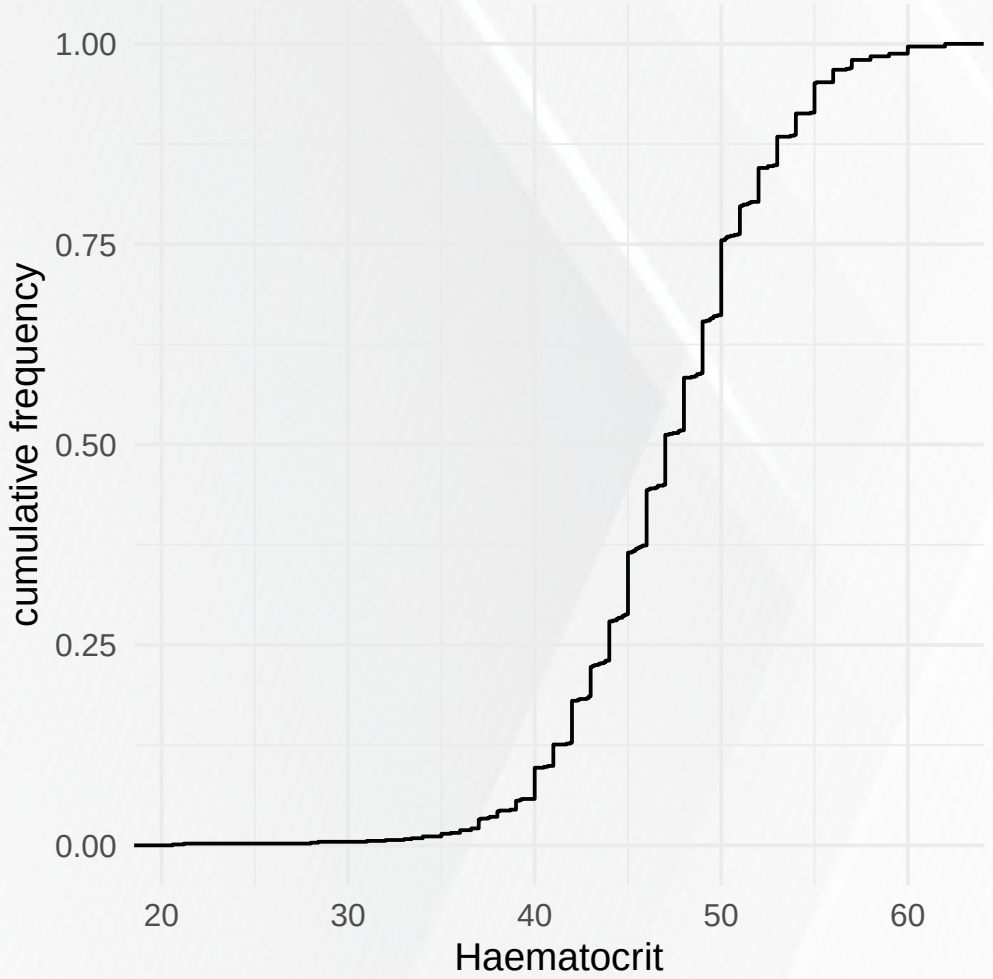


C-2: skewed distribution

A-3: peaked distribution, small IQR allocates some as outliers

B-1: larger IQR, nothing seen as outlier

Empirical Cumulative Distribution Function (ECDF)



Gives full description of values (most are integers), but harder to see specific pattern and skewness

Outline

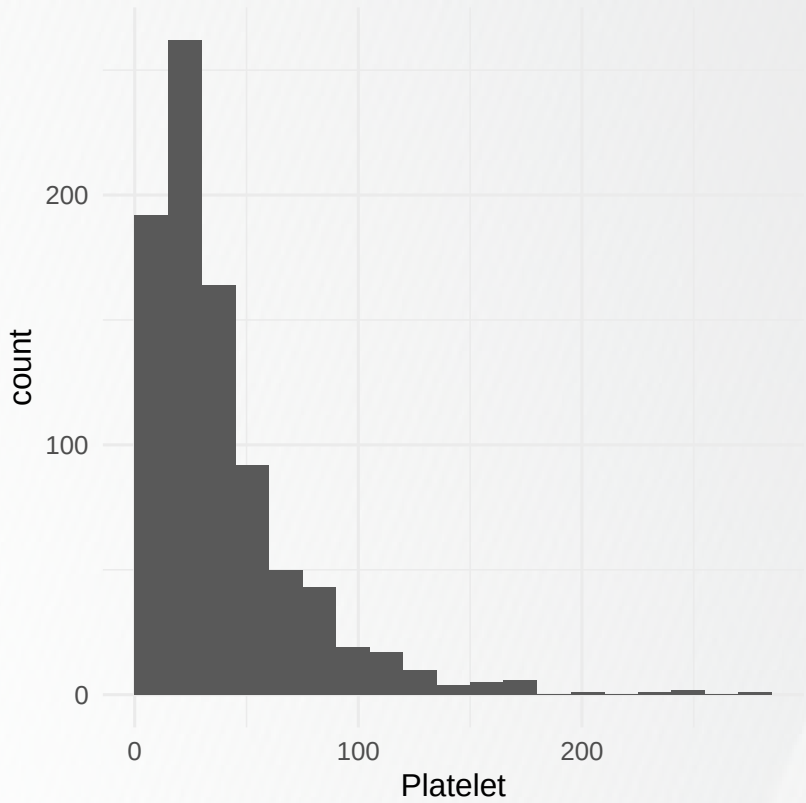
Variable transformation: logarithm

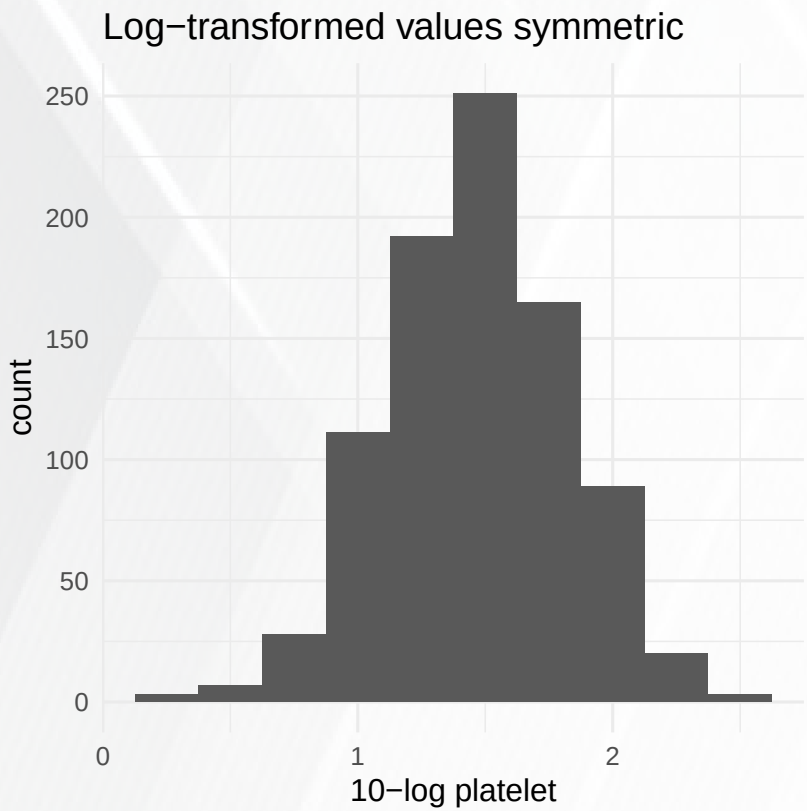
- Choose the base: natural (e), 10, 2
  - dengue viremia: range < 60 to $100,000,000$ copies/mL
    $\log_{10}(x) : 10 \to 1; 100 \to 2; 10,000 \to 4; 100,000,000 \to 8$
  - concentration dilutions: range 2-64
    $\log_2(x) : 2 \to 1; 4 \to 2; 8 \to 3; 64 \to 6$
  - natural base $e = 2.718\ldots$ $\log_e(2.718) = 1; \exp(1) = 2.718$

Variable transformation: logarithm

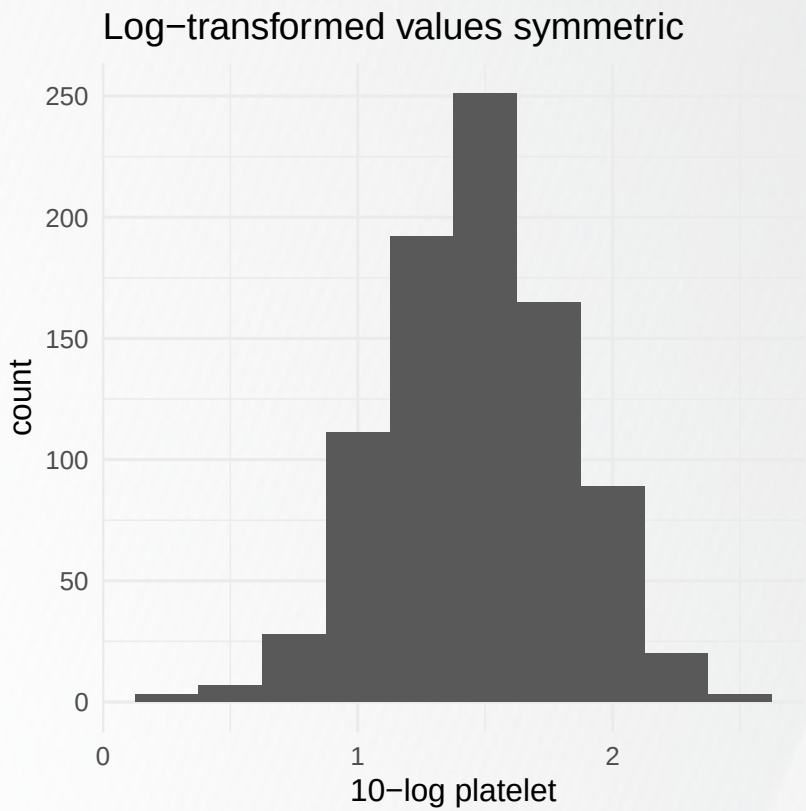
- Choose the base: natural (e), 10, 2
  - dengue viremia: range < 60 to $100,000,000$ copies/mL
    $\log_{10}(x) : 10 \to 1; 100 \to 2; 10,000 \to 4; 100,000,000 \to 8$
  - concentration dilutions: range 2-64
    $\log_2(x) : 2 \to 1; 4 \to 2; 8 \to 3; 64 \to 6$
  - natural base $e = 2.718\ldots$ $\log_e(2.718) = 1; \exp(1) = 2.718$
- You can use any base, but report the one you choose
- Warning: logarithm of 0 does not exist, would cause missing values, use for example $\log_{10}(x+1)$ or $\log_{10}(x+10)$

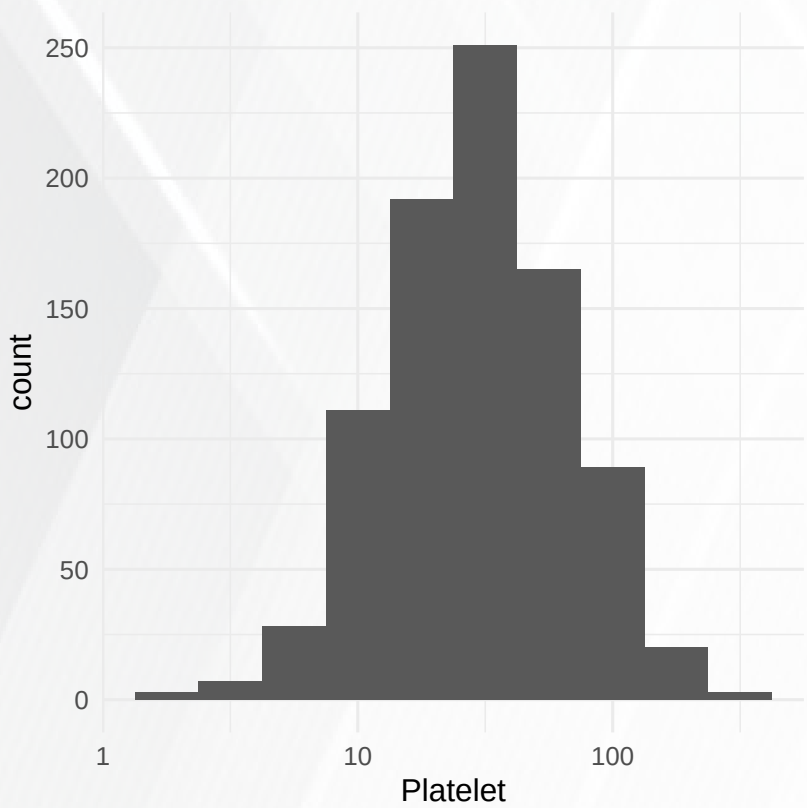
Data skewed to the right (long tail)





May use original values in labels





Summary descriptive analysis

- Tidy data: observations, variables, values
- Variables
  - binary, nominal, ordinal
  - discrete, continuous

  Other terms: derived variables, exposure, outcome
- Numerical summary
  - categorical: count/proportions
  - numerical: location (mean or median); spread (sd or IQR/quartiles)
- Graphics: histogram, boxplot, ecdf...
- Transformation of skewed numerical variable?
- Trying different summary options may give new insights

References

- Van den Broeck J, Cunningham SA, Eeckels R, Herbst K (2005) *Data cleaning: detecting, diagnosing, and editing data abnormalities.* PLoS Med 2:e267
- Data Management in Large-Scale Education Research
- You've just received your first dataset to analyse. Now what?

Outline

Data

- Data: structure
- Variables: type

Variables: numerical summaries

- Categorical variables
- Numerical variables

Variables: graphical summaries

- Transformation of variables

Practicals; Introduction to R

How to make the exercises

All materials at
`https://tranhung93.github.io/Introduction-to-Medical-Statistics/`

- Two versions of the exercises
  - those with little R experience: `Web-R version`
    R code can be written and run directly in the web browser (Chrome preferred); some code adaptation needed
  - those with some experience in R (e.g. via RStudio)
    Use `RStudio version`; code in separate R `Script file`
- Answers will be uploaded after the class

The R program

- Basic R Program is not user friendly; hardly any graphical user interface
- RStudio, a very neat "integrated development environment"
  - many user friendly options
  - uses the R program under the skin
  - great integration with Markdown and Quarto, which facilitates reproducible research
- R with Graphical User Interface, such as Blue Sky Statistics and Jamovi

R: functions

- All actions are performed via functions
  - what do I want R to do for me?
  - what does it need to know from me in order to do that?

R: functions

- All actions are performed via functions
  - what do I want R to do for me? goal
  - what does it need to know from me in order to do that? arguments
- Basic structure: `goal(arg1= , arg2= , ...)`
  example: `summary(object=cmTbm)`

R: functions

- All actions are performed via functions
  - what do I want R to do for me? goal
  - what does it need to know from me in order to do that? arguments
- Basic structure: `goal(arg1= , arg2= , ...)`
  example: `summary(object=cmTbm)`
- Many functions have formula structure
  `goal(y~x, data=..., ...)`
  example: `plot(bldwcc~age, data=cmTbm, ...)`

R: functions

- All actions are performed via functions
  - what do I want R to do for me? goal
  - what does it need to know from me in order to do that? arguments
- Basic structure: `goal(arg1= , arg2= , ...)`
  example: `summary(object=cmTbm)`
- Many functions have formula structure
  `goal(y~x, data=..., ...)`
  example: `plot(bldwcc~age, data=cmTbm, ...)`
- Argument names can be left out if there is no risk of ambiguity, for example

```
summary(cmTbm)
plot(bldwcc~age, cmTbm, ...)
```

R: extensions

- Package: collection of functions and/or data sets
- Some come with R program
- 1000's additional written by R users
  Need to be downloaded and installed on computer (easy in RStudio)
- Packages need to be loaded into R session via R code `library`
  example: `library(ggplot2)`

R: objects

- Everything is an object ($\approx$ files in operating system):
  Most important ones: data sets; functions
- You decide about the name of objects you create
  RStudio: overview of objects created in Environment tab

R: objects

- Everything is an object (≈ files in operating system):
  Most important ones: data sets; functions
- You decide about the name of objects you create
  RStudio: overview of objects created in Environment tab
- Output of a function can be assigned to an object or to an element of an object:

```
DataColloid <- subset(myData, fluid=="Colloid")
```

- We can also assign to a column in data set
  Example: transform values 1 and 2 into categorical:

```
cmTbm$sex <- factor(cmTbm$sex, labels=c("M","F"))
```

R: objects

- Everything is an object (≈ files in operating system):
  Most important ones: data sets; functions
- You decide about the name of objects you create
  RStudio: overview of objects created in Environment tab
- Output of a function can be assigned to an object or to an element of an object:

```
DataColloid <- subset(myData, fluid=="Colloid")
```

- We can also assign to a column in data set
  Example: transform values 1 and 2 into categorical:

```
cmTbm$sex <- factor(cmTbm$sex, labels=c("M","F"))
```

- Objects created can be used for further analyses

Writing code: use R script or R Markdown file

- Records everything that you have done
- Easy to check & reproduce
- Saves a lot of your time
- R script
  - collection of R code
  - comments start with #
- R Markdown and Quarto
  - combines R code with text
  - in RStudio, R code is run via button `Knit` (Markdown) or `Render` (Quarto)
  - produces formatted report in MS Word, html or pdf format

Further information

Ask LLM, or some suggestions:

- Peter D.R. Higgins. Reproducible Medical Research with R
- Grolemund G, Wickham H. R for Data Science
- Xie Y, Allaire JJ, Grolemund G. R Markdown: The Definitive Guide
- Easy data cleaning with the janitor package
- Your Data's Untold Secrets: An Introduction to Descriptive Stats with R
- YouTube videos BlueSky statistics