Introduction to Biostatistics & Descriptive Statistics

Phùng Khánh Lâm, MD, PhD
Department of Epidemiology, Faculty of Public Health
University of Medicine and Pharmacy at Ho Chi Minh City
05/08/2020

Outline

1. Introduction to biostatistics
2. Descriptive statistics
3. Making effective graphs and tables

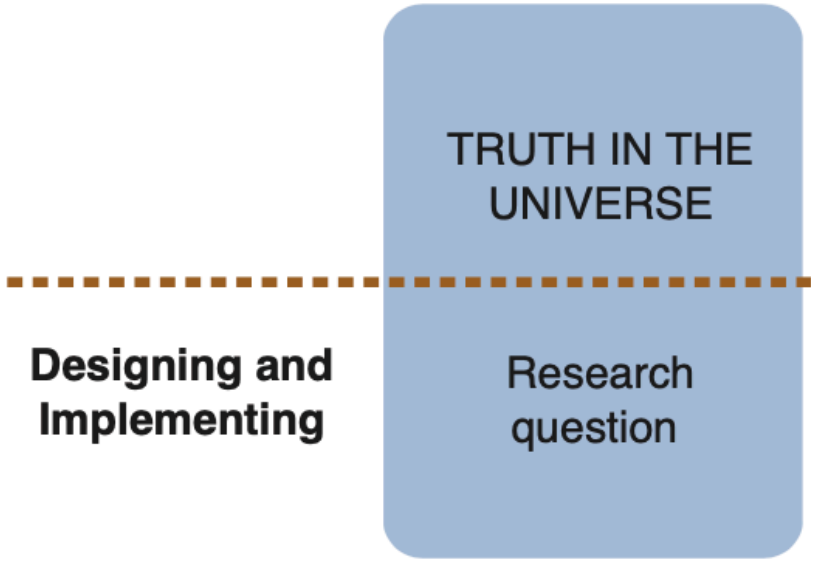
These slides were based on learning materials originally developed by Dr Marcel Wolbers, Prof Ronald Geskus and other members of the Biostatistics group at OUCRU

Introduction to biostatistics

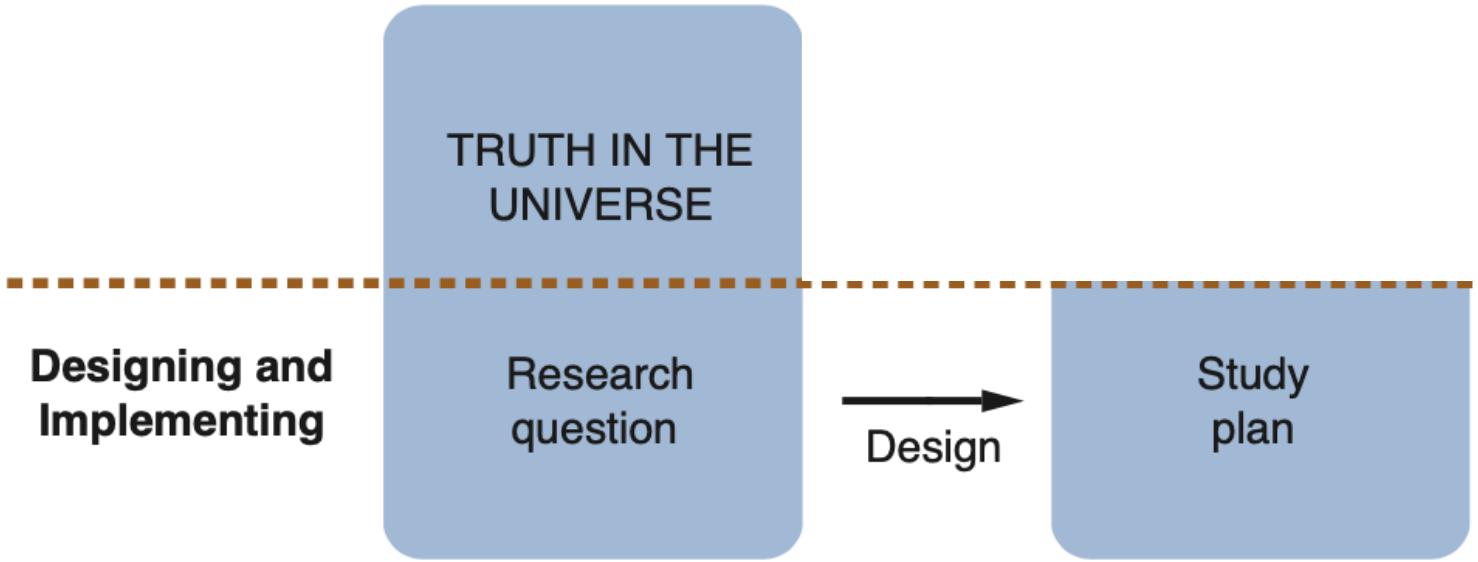
The research process

The research process

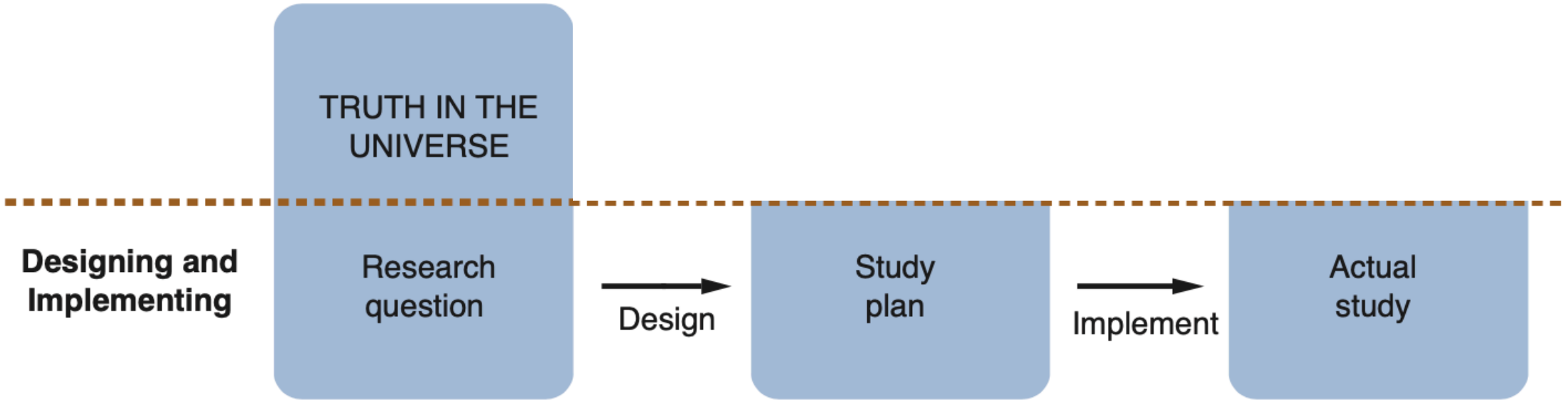
The research process



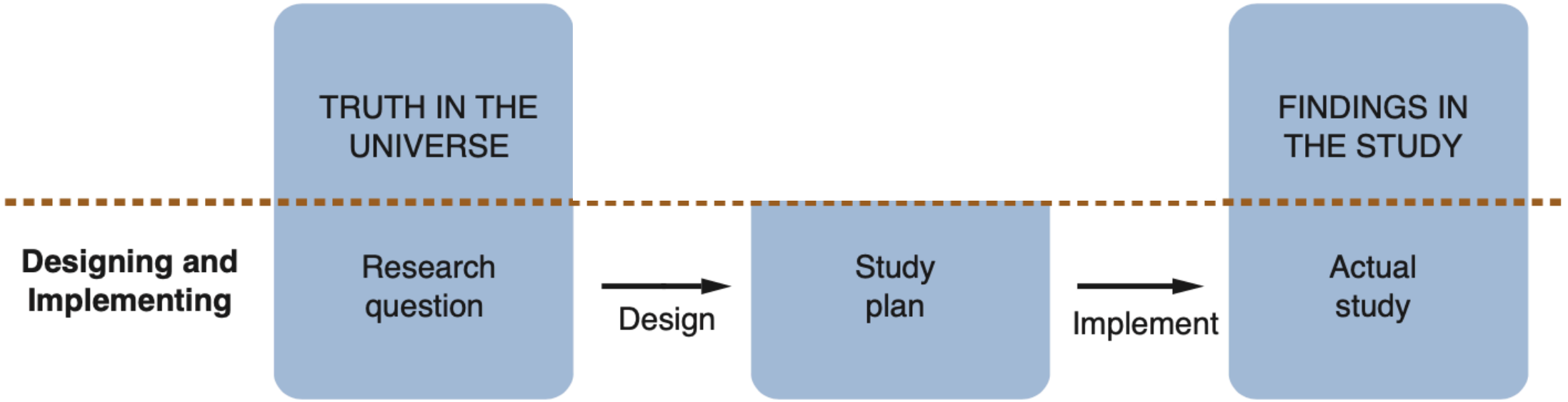
The research process



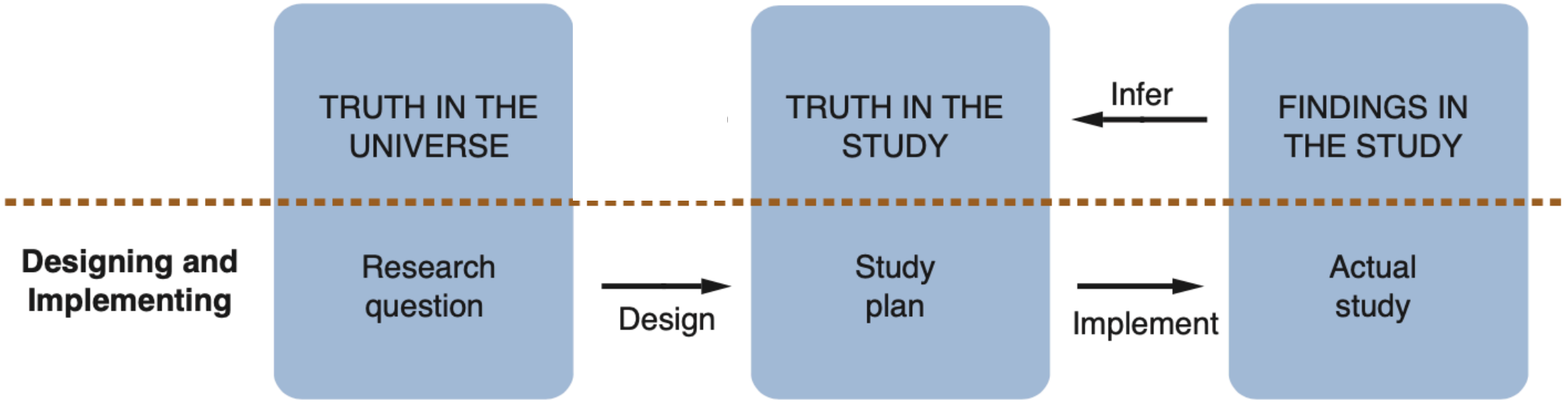
The research process



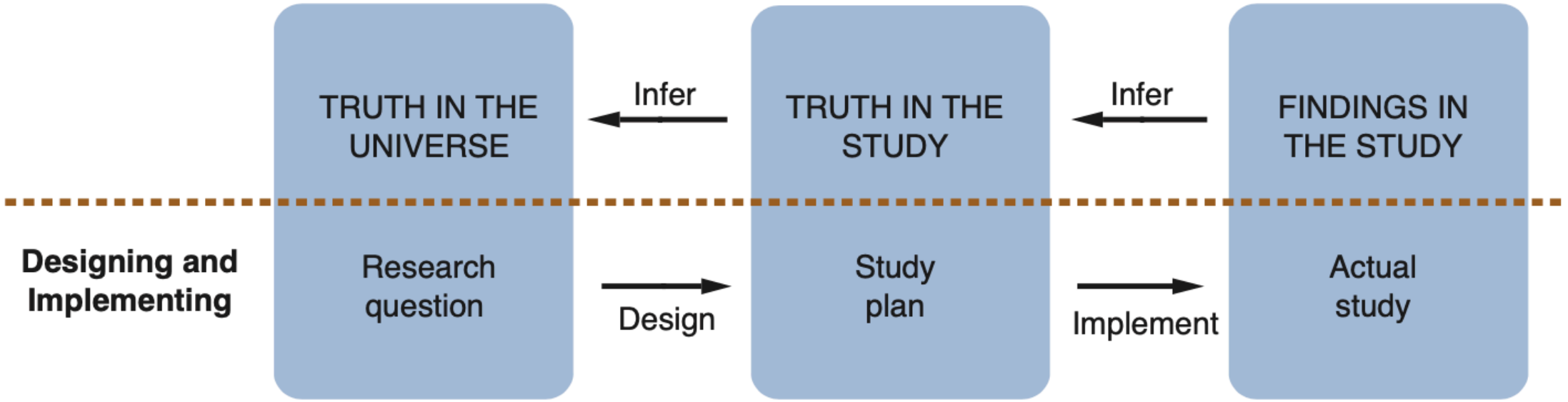
The research process



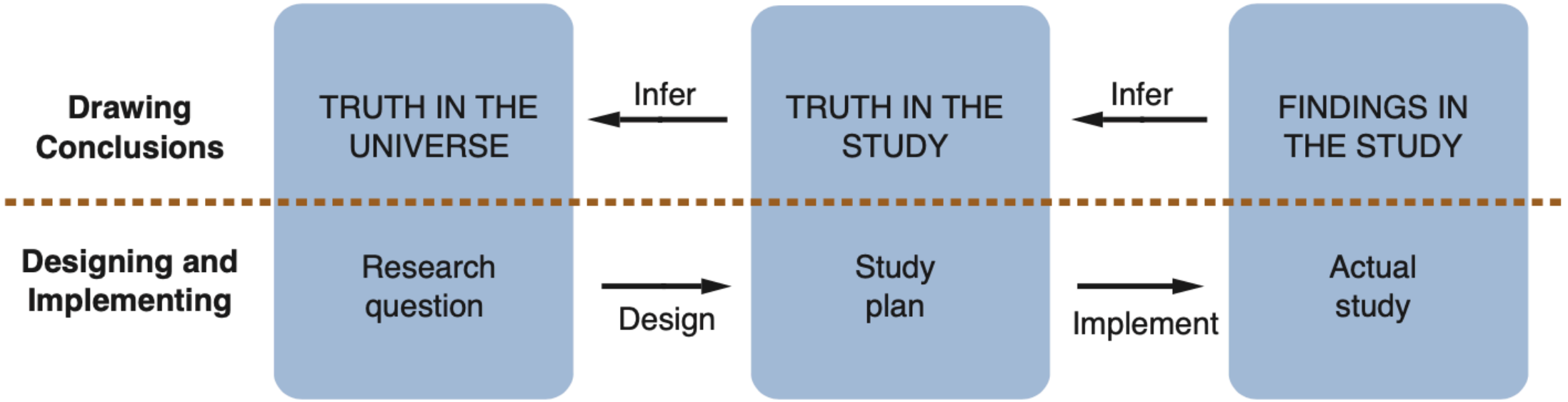
The research process



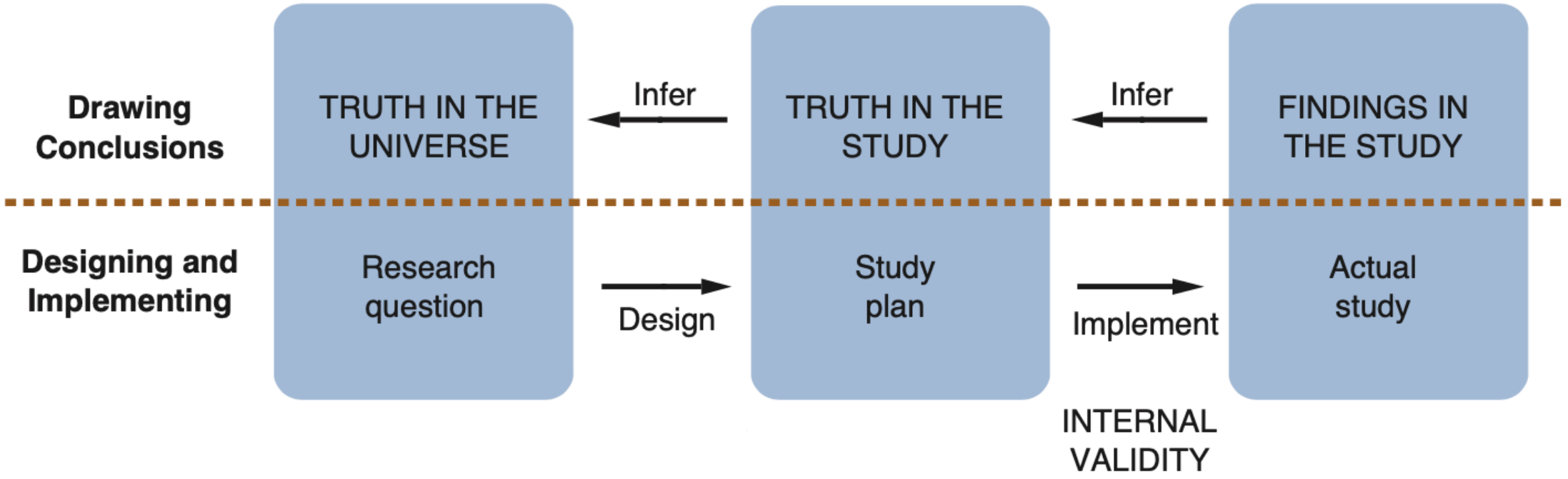
The research process



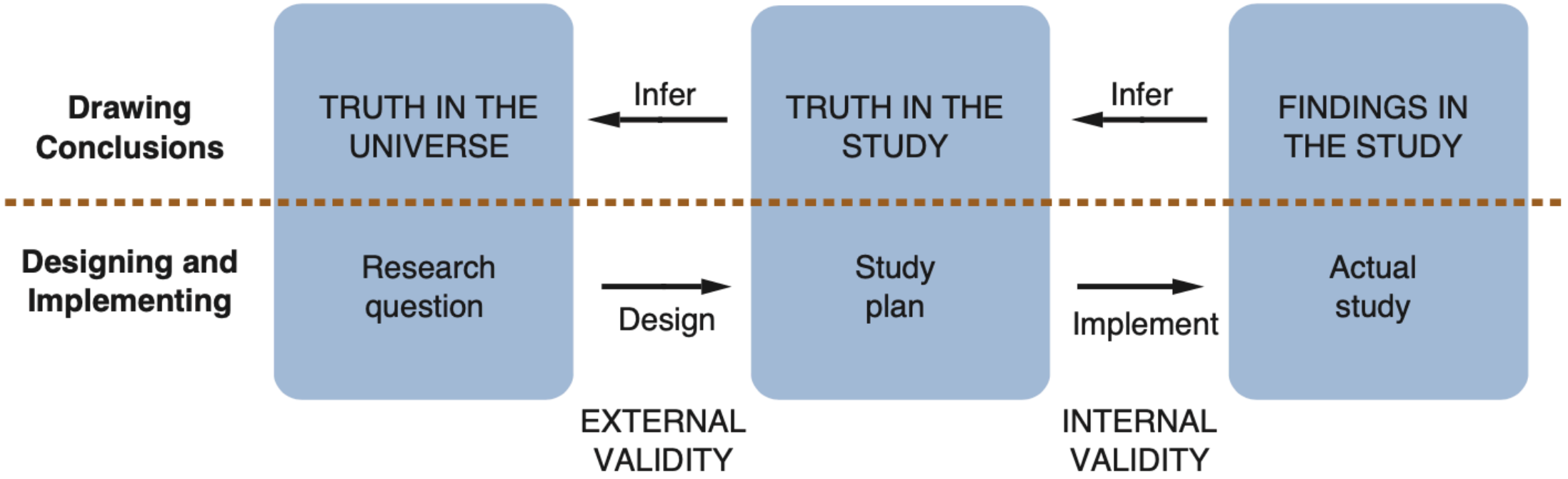
The research process



The research process

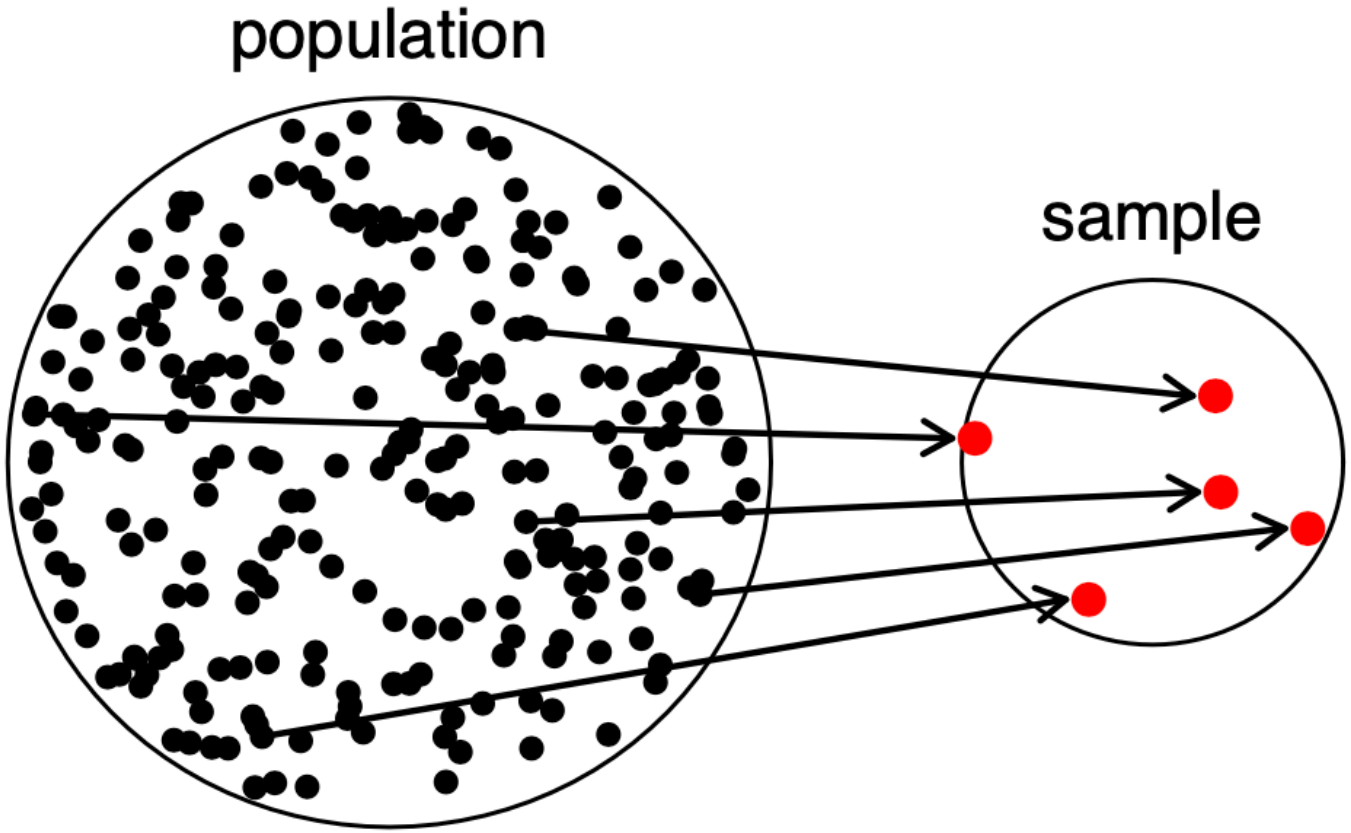


The research process



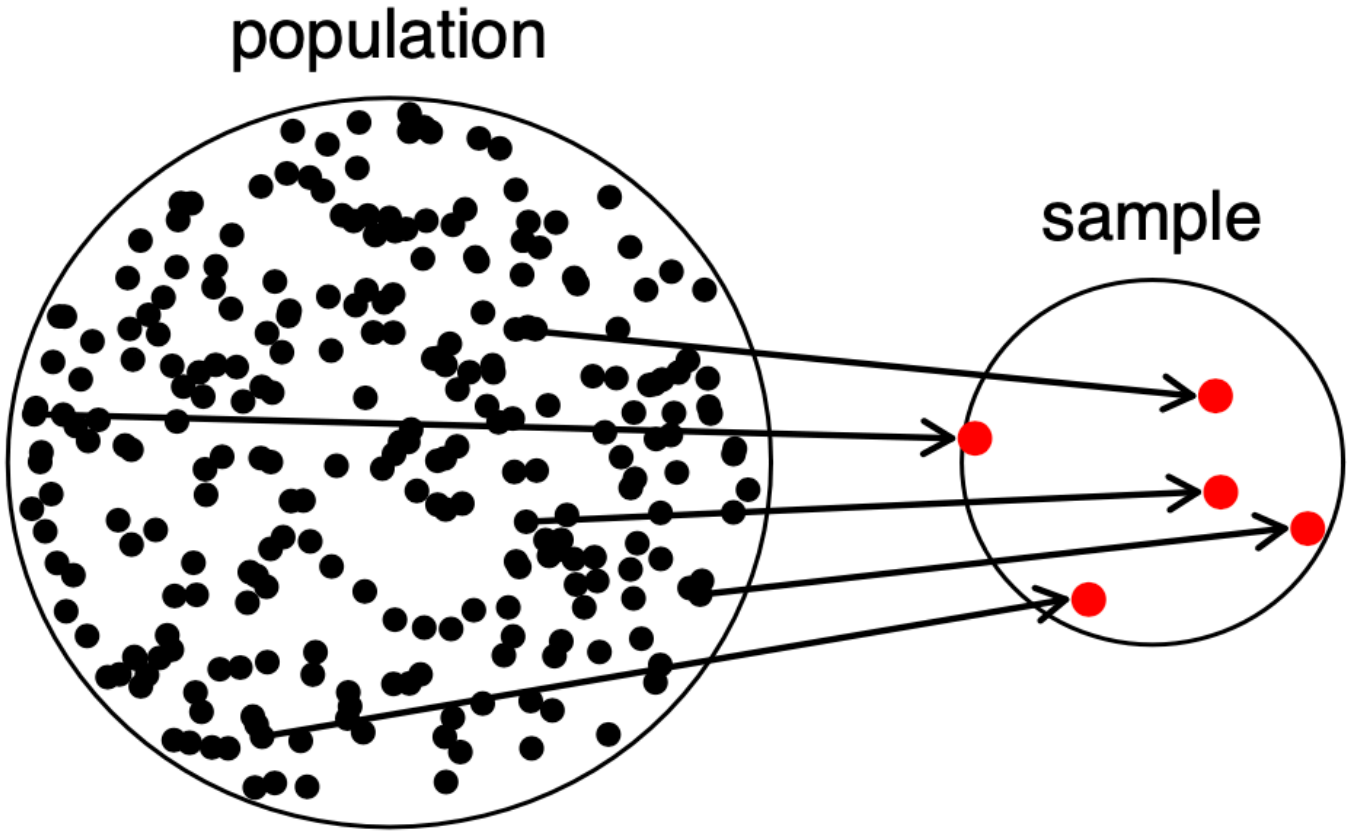
Data as sample from population

Data as sample from population



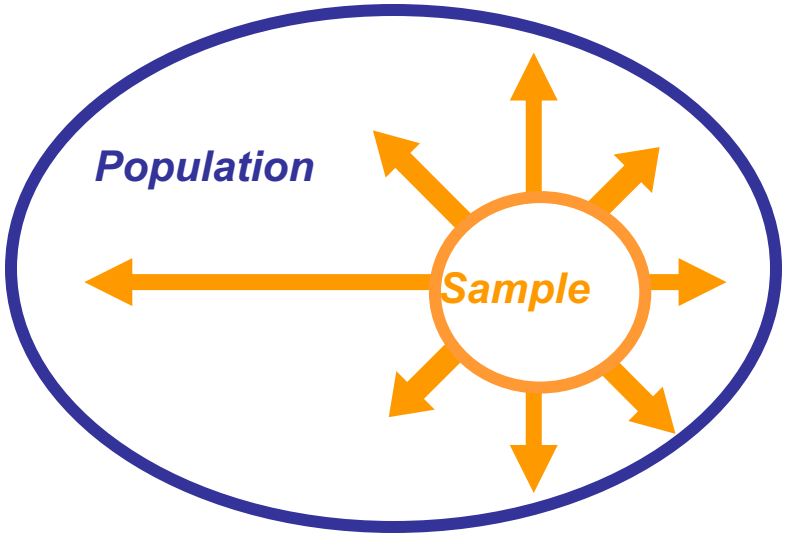
Biostatistics

INFER **DESCRIBE**

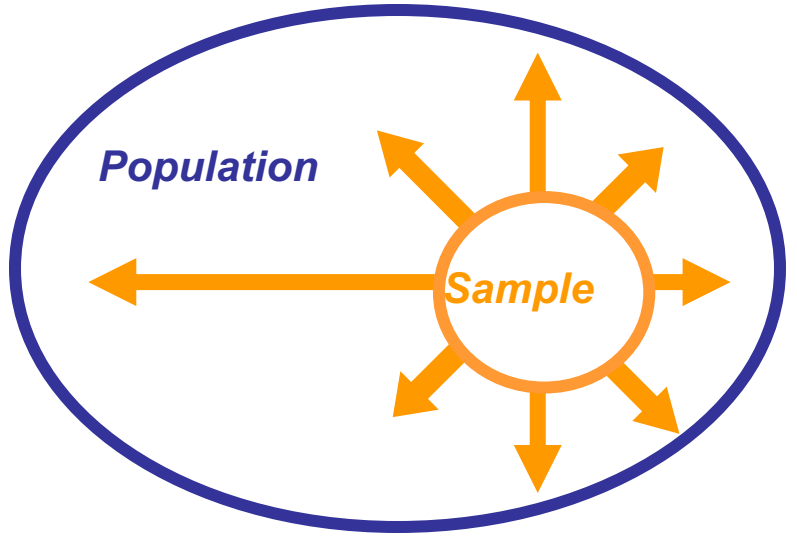


Biostatistics

- Descriptive statistics & Inferential statistics



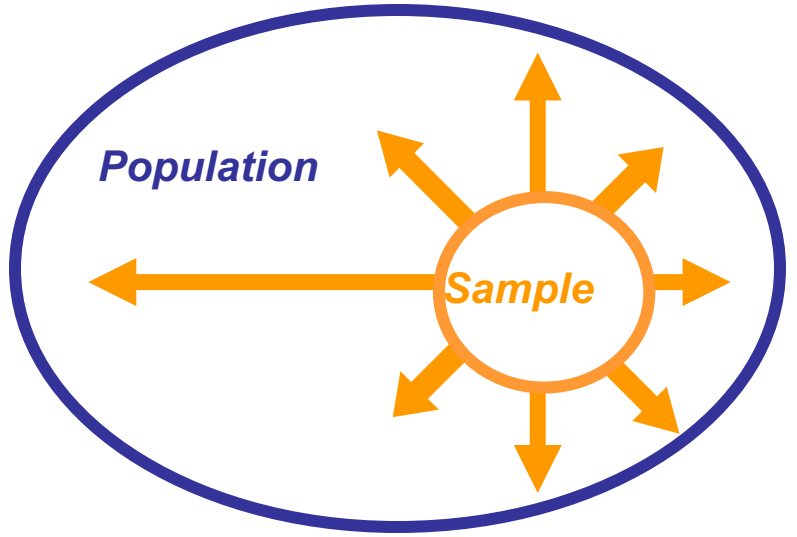
Biostatistics



- Descriptive statistics & Inferential statistics

- Descriptive statistics
  - Description/ exploration of dataset … and beyond
  - First step in data analysis

Biostatistics



- Descriptive statistics & Inferential statistics

- Descriptive statistics
  - Description/ exploration of dataset … and beyond
  - First step in data analysis

- Inferential statistics
  - Draw conclusions about a population using a sample

Descriptive statistics

- Describe data from sample

Descriptive statistics

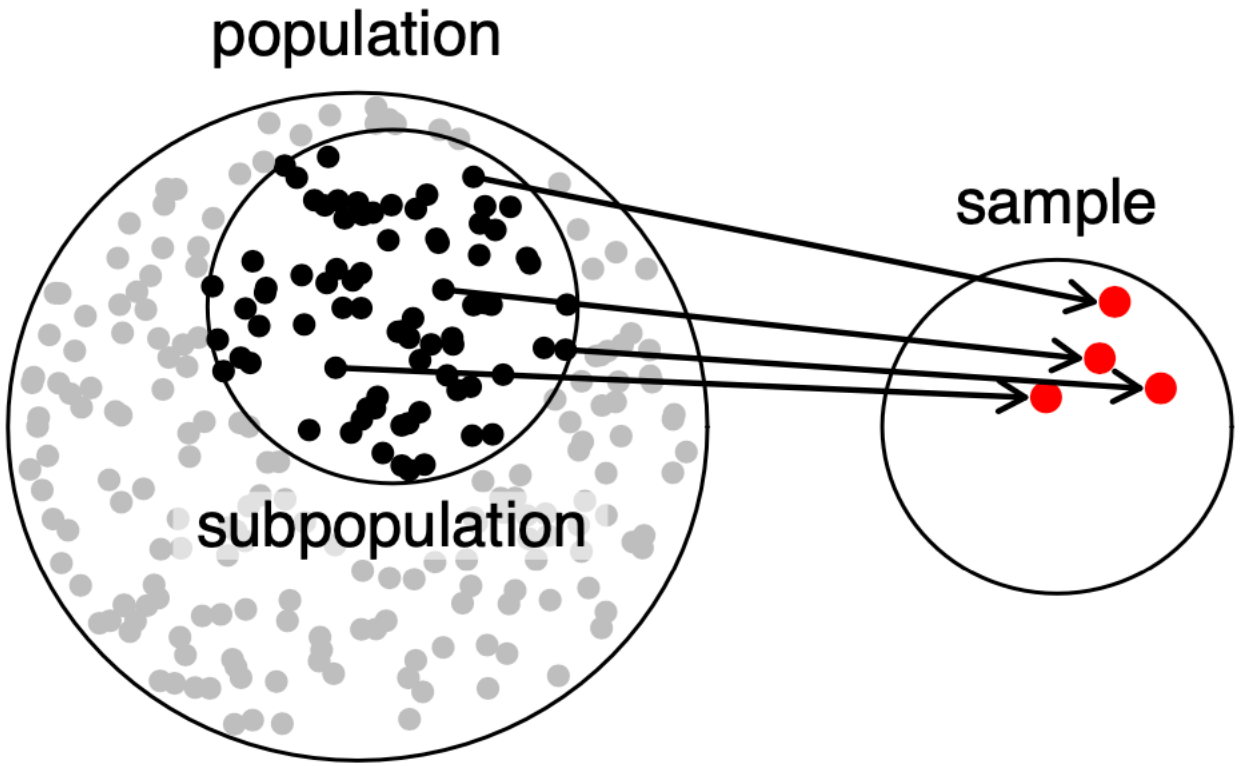
- Describe data from sample
- Use numbers and graphs

Inferential statistics

- From sample to population

Inferential statistics

- From sample to population
- But which population???



Random sampling

- Data randomly selected from some (much) larger population

Random sampling

- Data randomly selected from some (much) larger population
- Representative sample: observations (summary, relations between variables) can be transferred to population

Random sampling

- Data randomly selected from some (much) larger population
- Representative sample: observations (summary, relations between variables) can be transferred to population
  - Data summaries ≈ population summaries

Random sampling

- Data randomly selected from some (much) larger population
- Representative sample: observations (summary, relations between variables) can be transferred to population
  - Data summaries ≈ population summaries
  - Some uncertainty: only (small) sample from population
  - Result slightly different if study repeated (new sample)

Quiz: target population

- Patients with dengue shock treated at HTD during 2014-17

Quiz: target population

- Patients with dengue shock treated at HTD during 2014-17
- Which population do the data represent?

Quiz: target population

- Patients with dengue shock treated at HTD during 2014-17
- Which population do the data represent?
  - all patients with dengue shock treated at HTD, in past and (near) future
  - all patients with dengue shock in Viet Nam all patients with dengue shock

Quiz: target population

- Patients with dengue shock treated at HTD during 2014-17
- Which population do the data represent?
  - all patients with dengue shock treated at HTD, in past and (near) future
  - all patients with dengue shock in Viet Nam all patients with dengue shock

- **Always ask yourself the question which population the sample represents**

Inferential statistics

What do we want to say about the population?

Inferential statistics
What do we want to say about the population?

- Simple summary of some quantity
  - e.g. incidence of tuberculous meningitis (TBM)

Inferential statistics
What do we want to say about the population?

- Simple summary of some quantity
  - e.g. incidence of tuberculous meningitis (TBM)

- Relation between one or more variables and an outcome

Inferential statistics
What do we want to say about the population?

- Simple summary of some quantity
  - e.g. incidence of tuberculous meningitis (TBM)

- Relation between one or more variables and an outcome
  - **Prediction (clinical)**
    - Probability of death based on patient characteristics at TBM diagnosis?

Inferential statistics
What do we want to say about the population?

- Simple summary of some quantity
  - e.g. incidence of tuberculous meningitis (TBM)

- Relation between one or more variables and an outcome
  - **Prediction (clinical)**
    - Probability of death based on patient characteristics at TBM diagnosis?
  - **Etiology (scientific; "explanation")**: causal relation
    - Does dexamethasone decrease risk of dying?
    - Role of HIV coinfection in TBM disease process?

Inferential statistics
What do we want to say about the population?

- Simple summary of some quantity
  - e.g. incidence of tuberculous meningitis (TBM)

- Relation between one or more variables and an outcome
  - **Prediction (clinical)**
    - Probability of death based on patient characteristics at TBM diagnosis?
  - **Etiology (scientific; "explanation")**: causal relation
    - Does dexamethasone decrease risk of dying?
    - Role of HIV coinfection in TBM disease process?
  - **Exploration**
    - What are the risk factors for mortality in TBM patients?

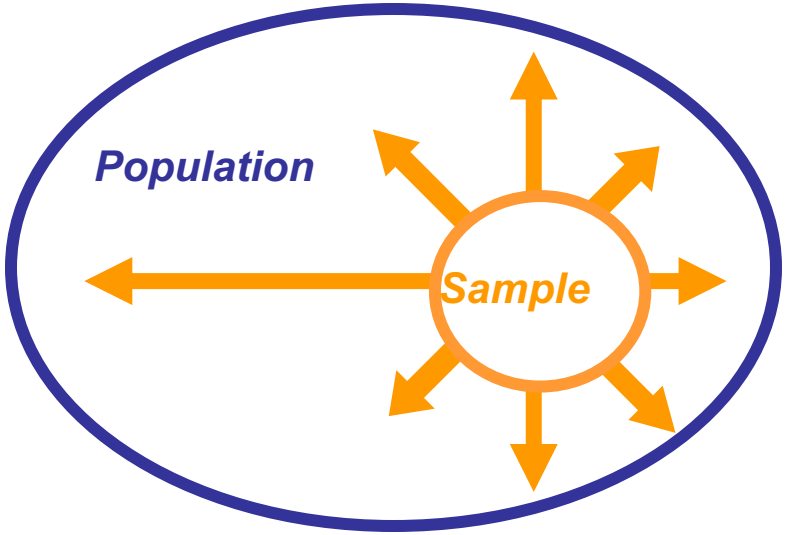
Summary

Summary

- Biostatistics
  - The development and application of statistical methods to a wide range of topics in biology and medicine

Summary

- Biostatistics
  - The development and application of statistical methods to a wide range of topics in biology and medicine
- Two main branches
  - Descriptive statistics
  - Inferential statistics



Descriptive statistics

Some definitions

Some definitions

- **Study:** sample from population

Some definitions

- **Study:** sample from population

- **Dataset:** contains observations from a study

Some definitions

- **Study:** sample from population
- **Dataset:** contains observations from a study
- **Observation:** values of variables that are measured on a unit (patients, animals, farms) at one specific time

Some definitions

- **Study:** sample from population
- **Dataset:** contains observations from a study
- **Observation:** values of variables that are measured on a unit (patients, animals, farms) at one specific time
- **Variable:** characteristic that may vary over units

Data structure

Data structure

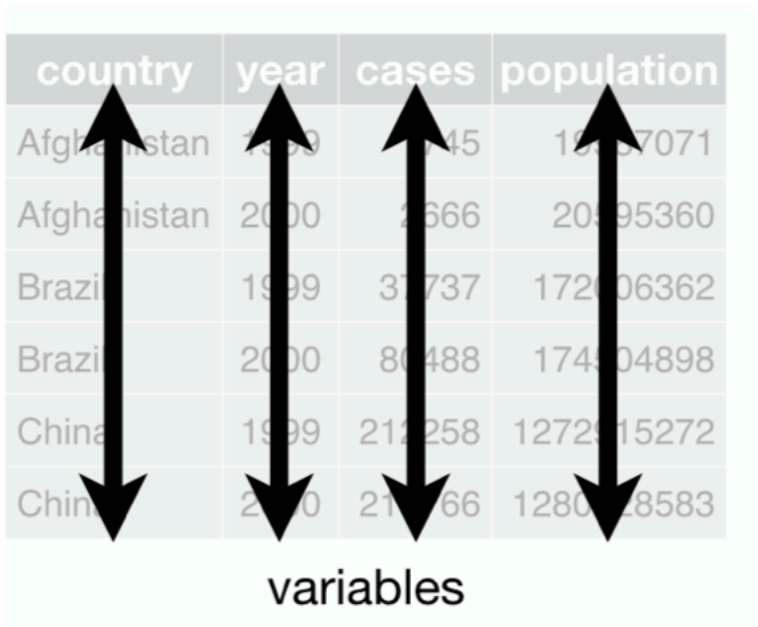
- 80% of analysis time spent on data cleaning and preparation, especially when data are **“messy”**

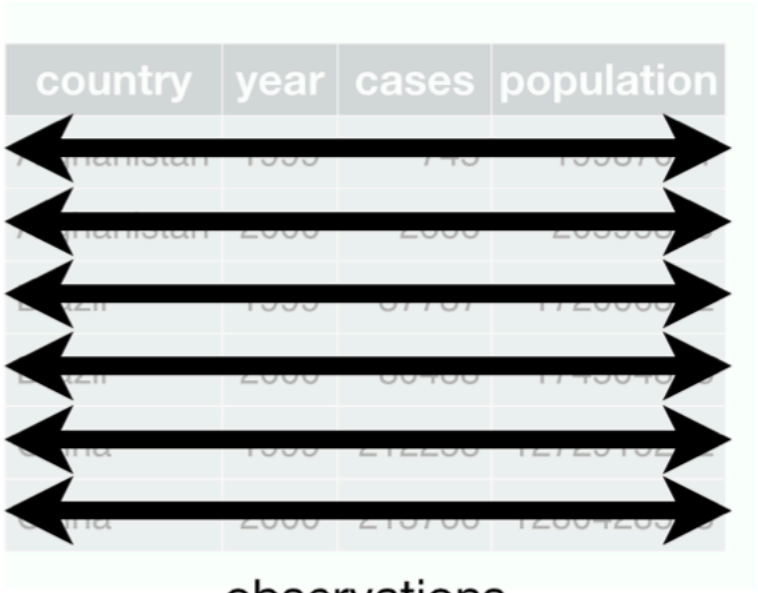
Data structure

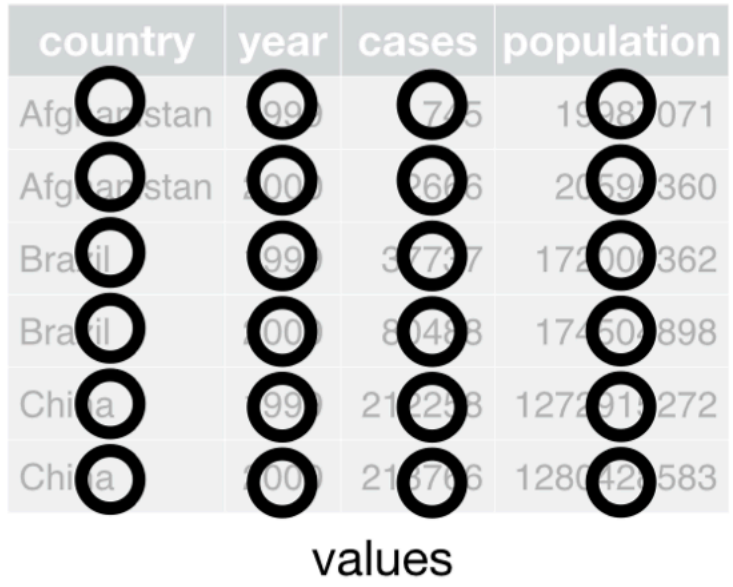
- 80% of analysis time spent on data cleaning and preparation, especially when data are **“messy”**

- **“Tidy”** data: link structure with semantics
  - Each observation forms a row
  - Each variable forms a column
  - One table for each type of observations
  - If you have multiple tables, they should include a column in the table that allows them to be linked

Tidy data


country
year
cases
population
variables


country
year
cases
population
observations


country
year
cases
population
values

Tidy data: example

studyno	Fluid	age	sex	hct1	plat1	hospdays	clinical_overload
400	Dextran	9	male	42	80000	4	no
401	Dextran	13	female	48	100000	4	no
402	Starch	10	female	50	47000	4	no
407	Starch	8	male	40	NA	5	no
410	Lactate Hartman	6	male	45	28000	6	yes
412	Dextran	10	female	52	33000	10	yes

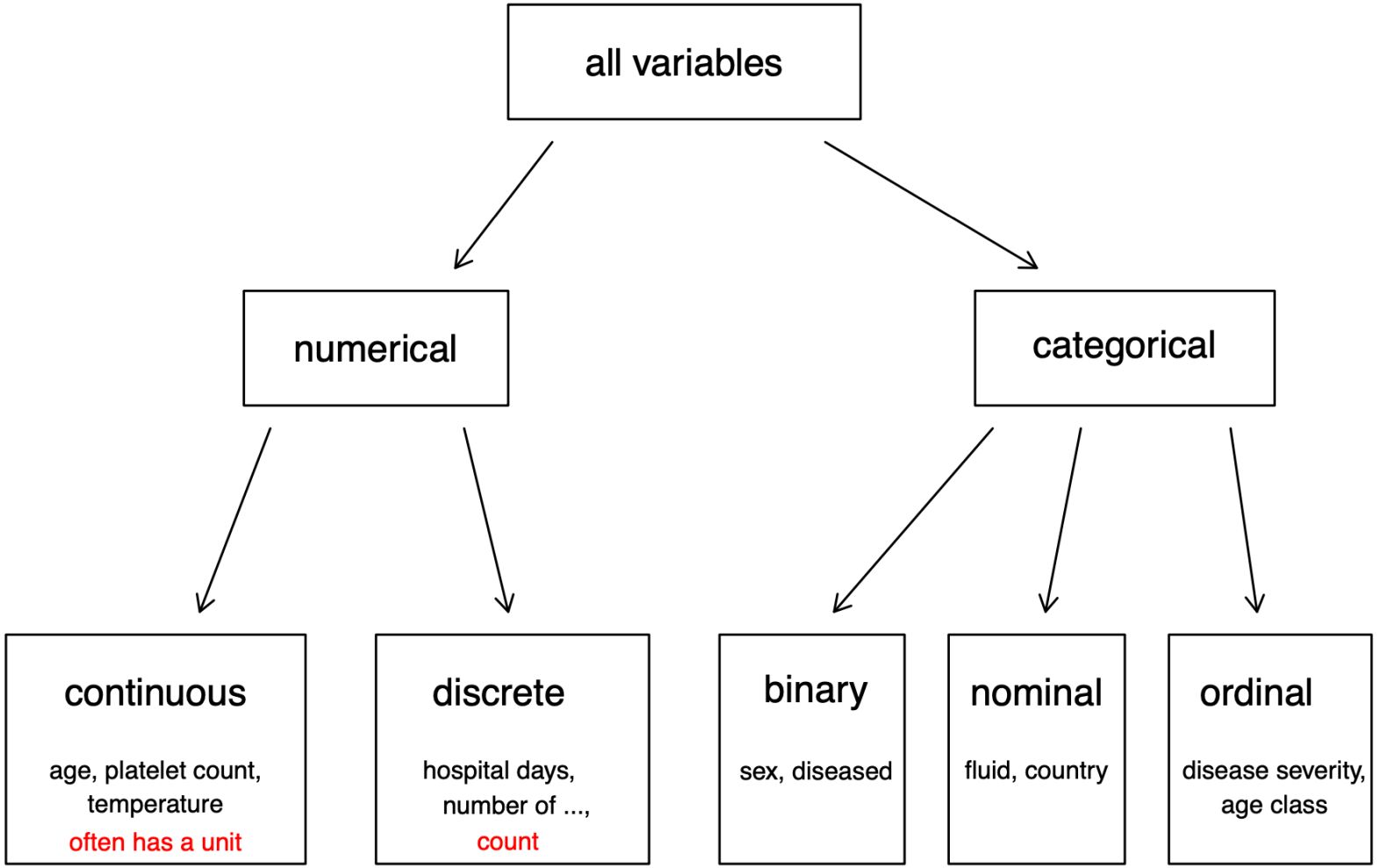
Tidy data: example

studyno	Fluid	age	sex	hct1	plat1	hospdays	clinical_overload
400	Dextran	9	male	42	80000	4	no
401	Dextran	13	female	48	100000	4	no
402	Starch	10	female	50	47000	4	no
407	Starch	8	male	40	NA	5	no
410	Lactate Hartman	6	male	45	28000	6	yes
412	Dextran	10	female	52	33000	10	yes

- Rows: units (subjects, patients at a single time point)
- Columns: variables (outcome, response; covariates, covariables, predictors)
- Cells: values

Types of variables

Types of variables



Types of variables

1) **Categorical** (binary, nominal, ordinal) and
2) **Numeric** (count, continuous):

Binary / Dichotomous - 2 levels: sex, diseased

Nominal – more than 2 levels: fluid, country

Ordinal – ordered levels: severity of disease, disability score, age class

Count – integer: number of …, calendar year, age year, hospital days

Continuous –often has a unit: temperature, platelet count, age

Quiz: Types of variables in this table?

studyno	Fluid	age	sex	hct1	plat1	hospdays	clinical_overload
400	Dextran	9	male	42	80000	4	no
401	Dextran	13	female	48	100000	4	no
402	Starch	10	female	50	47000	4	no
407	Starch	8	male	40	NA	5	no
410	Lactate Hartman	6	male	45	28000	6	yes
412	Dextran	10	female	52	33000	10	yes

Quiz: Types of variables in this table?

studyno	Fluid	age	sex	hct1	plat1	hospdays	clinical_overload
400	Dextran	9	male	42	80000	4	no
401	Dextran	13	female	48	100000	4	no
402	Starch	10	female	50	47000	4	no
407	Starch	8	male	40	NA	5	no
410	Lactate Hartman	6	male	45	28000	6	yes
412	Dextran	10	female	52	33000	10	yes

- Binary: sex, clinical_overload
- Nominal: Fluid, studyno
- Count: hospdays, age (if rounded, “life years” without unit)
- Continuous: age, hct1, plat1

Coding variable values

Coding variable values

- Categorical variables often coded numerically
  - e.g. 1=male, 2=female
  - 1=Vietnam, 2=Thailand, 3=Laos
- BUT: This does not make them numerical!

Coding variable values

- Categorical variables often coded numerically
  - e.g. 1=male, 2=female
  - 1=Vietnam, 2=Thailand, 3=Laos
- BUT: This does not make them numerical!

- Remember during the analyses (or better code them as character from the start)

Coding variable values

- Categorical variables often coded numerically
  - e.g. 1=male, 2=female
  - 1=Vietnam, 2=Thailand, 3=Laos
- BUT: This does not make them numerical!

- Remember during the analyses (or better code them as character from the start)

- Missing data
  - use special code; NA (“not available”) in R
  - always report amount of missingness per variable
  - usually excluded in analysis (but may introduce bias)

Summarize data

Summarize data

- Write on a sheet of paper: 20 numbers from 0 to 9

Summarize data

- Write on a sheet of paper: 20 numbers from 0 to 9
- Work in pair: verbally describe your numbers to partner

Summarize data

Summarize data

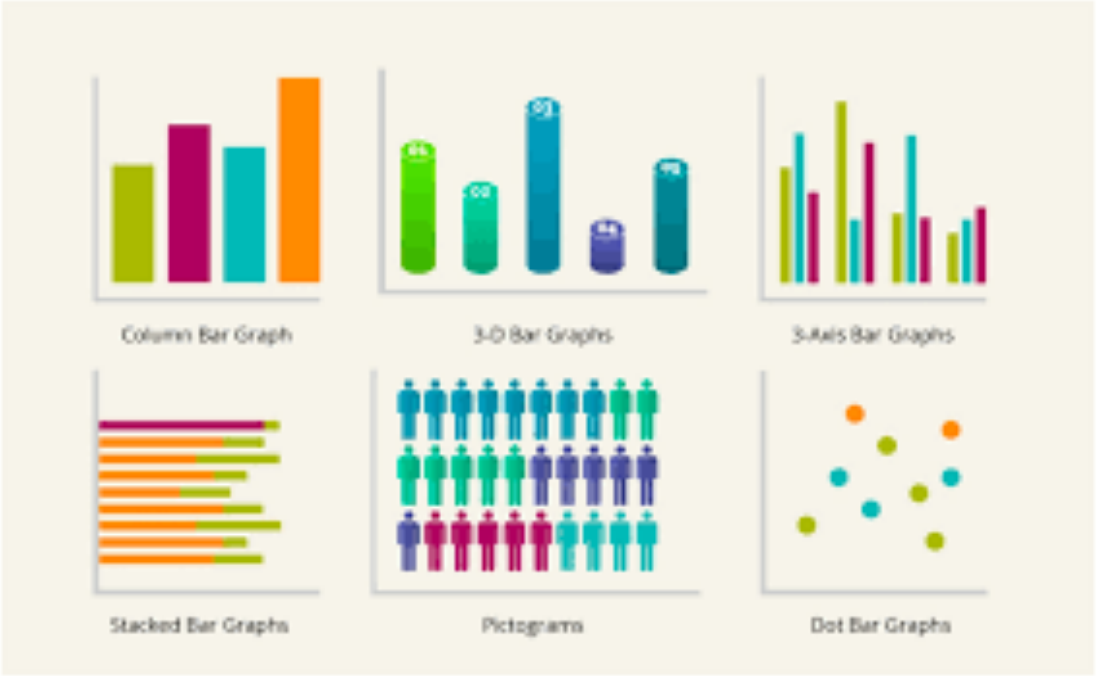
Numerical

Summarize data

Numerical

1 2 3
4 5 6
7 8 9

Graphical



Numerical summary

Numerical summary

- Categorical variables

Numerical summary

- Categorical variables
  - **Frequency:** how many subjects in each level/category

Numerical summary

- Categorical variables
  - **Frequency:** how many subjects in each level/category
  - **Relative frequency:** percentage (0%-100%) or proportion (0-1) of each category

Numerical summary

- Categorical variables
  - **Frequency:** how many subjects in each level/category
  - **Relative frequency:** percentage (0%-100%) or proportion (0-1) of each category

- Numerical variables

Numerical summary

- Categorical variables
  - **Frequency:** how many subjects in each level/category
  - **Relative frequency:** percentage (0%-100%) or proportion (0-1) of each category

- Numerical variables
  - **Location:** mean, median
  - **Dispersion:** standard deviation, quartiles, range

Location: mean

Location: mean

- Adding up all values and dividing this sum by the number of values

Location: mean

- Adding up all values and dividing this sum by the number of values
  - e.g.: 5 patients age: 20, 22, 25, 63, 75
  - mean age = 205/5 =41 yrs

- General formula $$\bar{x} := \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Location: median

Location: median

- Order data, from smallest value to largest value

Location: median

- Order data, from smallest value to largest value
- Odd sample size: middle observation
- Even sample size: average of the two „middle“ observations

Location: median

- Order data, from smallest value to largest value
- Odd sample size: middle observation
- Even sample size: average of the two „middle“ observations
- Median splits the sample in two halves: 50% are lower, 50% are higher

Location: median

- Order data, from smallest value to largest value
- Odd sample size: middle observation
- Even sample size: average of the two „middle“ observations
- Median splits the sample in two halves: 50% are lower, 50% are higher
  - e.g.: 5 patients age: 20, 22, 25, 63, 75

Location: median

- Order data, from smallest value to largest value
- Odd sample size: middle observation
- Even sample size: average of the two „middle“ observations
- Median splits the sample in two halves: 50% are lower, 50% are higher
  - e.g.: 5 patients age: 20, 22, 25, 63, 75
  - median age = 25 yrs

Location: mean vs. median

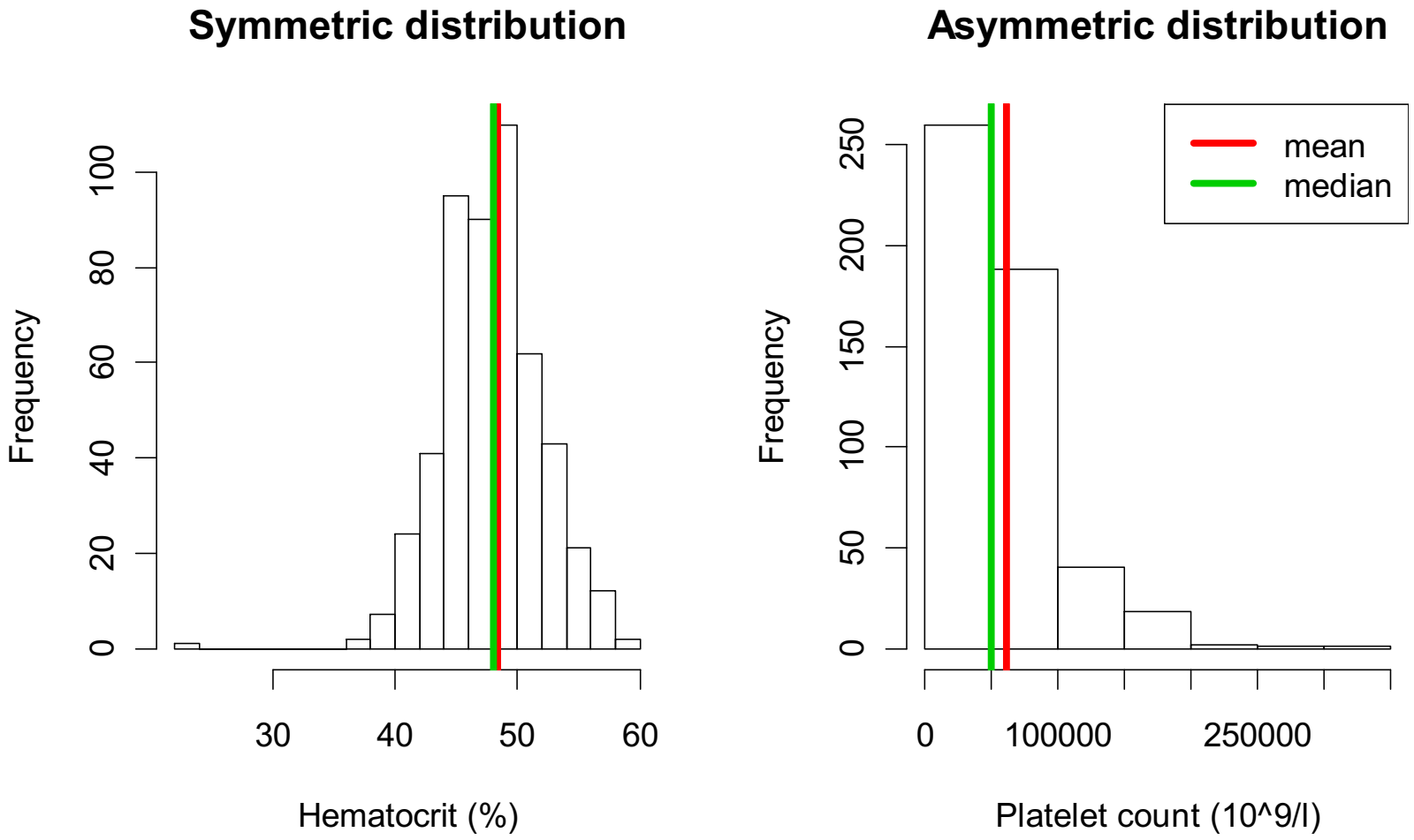
Mean

- Optimal if data distribution
  - ~ symmetric
  - No too long tails or outliers
- e.g. data from normal distribution
- Basis of most statistical models and tests

Median

- Close to sample mean if data has symmetric distribution
- Smaller than mean if distribution skewed to the right
- Still meaningful if data has extreme values

Location: mean vs. median - example



Quiz: Sample mean vs. median

Quiz: Sample mean vs. median

- Length of hospital stay after being admitted to hospital with community-acquired pneumonia

- Distribution is skewed to the right: median ~ 8 days, mean ~ 10 days

Quiz: Sample mean vs. median

- Length of hospital stay after being admitted to hospital with community-acquired pneumonia

- Distribution is skewed to the right: median ~ 8 days, mean ~ 10 days

- **Mean** or the **median** more relevant for you if
  - you are a prospective patient?
  - you are a hospital administrator interested in costs?

Dispersion

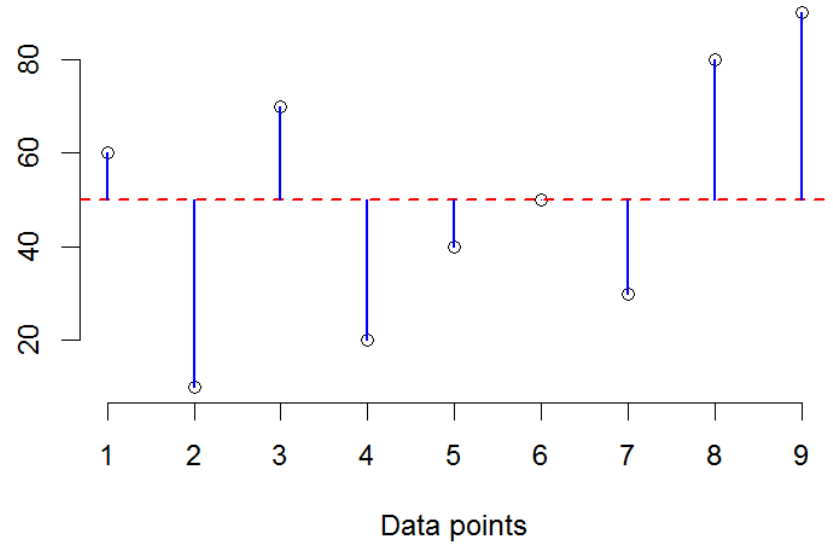
Dispersion

- Age: 60, 10, 70, 20, 40, 50, 30, 80, 90

- Age: 52, 51, 47, 49, 54, 46, 52, 46, 53

Dispersion

- Age: 60, 10, 70, 20, 40, 50, 30, 80, 90



- Age: 52, 51, 47, 49, 54, 46, 52, 46, 53

- Mean 50, but dispersion around mean is different

Dispersion I: variance and standard deviation

Dispersion I: variance and standard deviation

- Variance: square each deviation and average

$$\text{variance} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Dispersion I: variance and standard deviation

- Variance: square each deviation and average

$$\text{variance} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Standard deviation: Square root of variance

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

Dispersion I: variance and standard deviation

- Variance: square each deviation and average

$$\text{variance} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Standard deviation: Square root of variance

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

- Rule: If data has approximately normal distribution then
  - ~68% of observations lie between $\bar{x} \pm \text{sd}$
  - ~95% of observation lie between $\bar{x} \pm 2 \cdot \text{sd}$

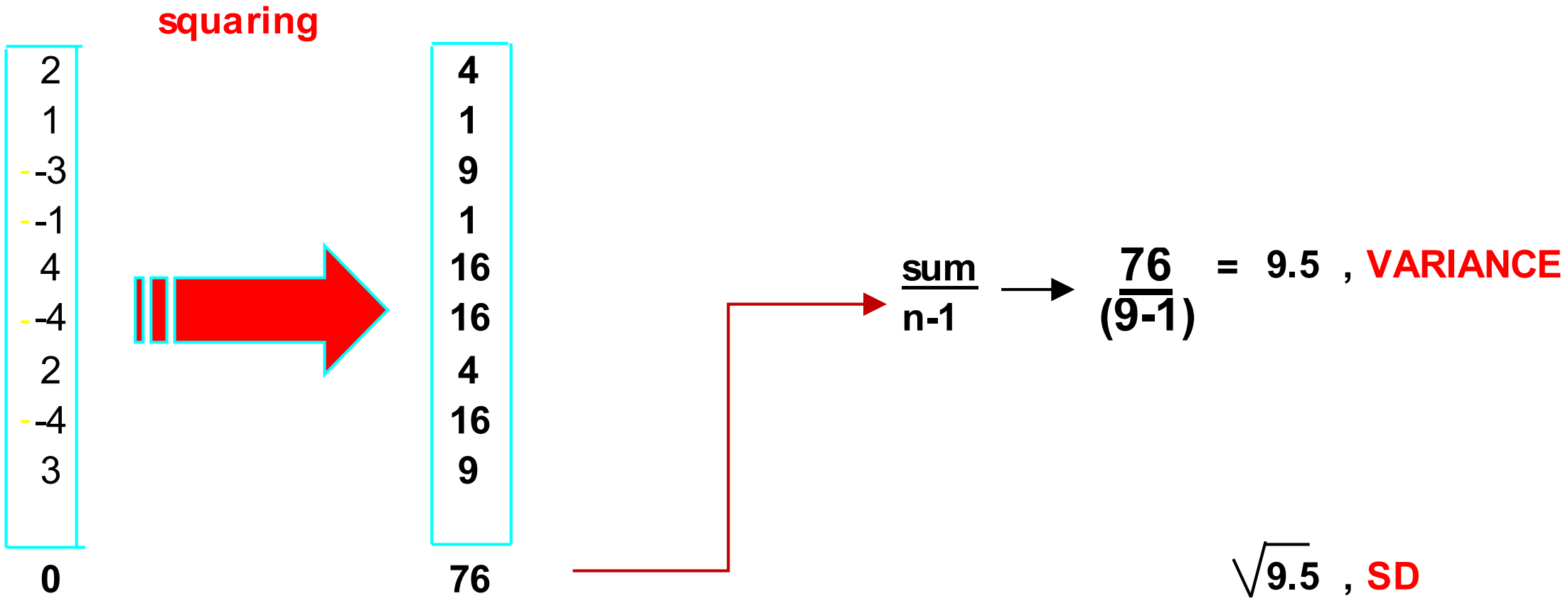
Deviation

- By definition: mean of the deviation is zero

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

The positive differences exactly cancel out the negative differences

Squared deviation



Dispersion II: Range, interquartile range (IQR)

- Range: minimum to maximum

Dispersion II: Range, interquartile range (IQR)

- Range: minimum to maximum

- Quartiles

Dispersion II: Range, interquartile range (IQR)

- Range: minimum to maximum

- Quartiles
  - First quartile q1: cuts off lowest 25% of data
  -

Dispersion II: Range, interquartile range (IQR)

- Range: minimum to maximum

- Quartiles
  - First quartile q1: cuts off lowest 25% of data
  - Second quartile q2: cuts off lowest 50% of data (median)

Dispersion II: Range, interquartile range (IQR)

- Range: minimum to maximum

- Quartiles
  - First quartile q1: cuts off lowest 25% of data
  - Second quartile q2: cuts off lowest 50% of data (median)
  - Third quartile q3: cuts off lowest 75% of data

Dispersion II: Range, interquartile range (IQR)

- Range: minimum to maximum

- Quartiles
  - First quartile q1: cuts off lowest 25% of data
  - Second quartile q2: cuts off lowest 50% of data (median)
  - Third quartile q3: cuts off lowest 75% of data
  - IQR: q3 – q1 (often reported as [q1, q3])

Quiz: medians and IQRs

Quiz: medians and IQRs

Compare distributions (1) and (2) based on their medians and IQRs

(1) 3, 5, 6, 7, 9	(2) 3, 5, 6, 7, 20
(1) 3, 5, 6, 7, 9	(2) 3, 5, 8, 7, 9
(1) 1, 2, 3, 4, 5	(2) 6, 7, 8, 9, 10
(1) 0, 10, 50, 60, 100	(2) 0, 100, 500, 600, 1000

Dispersion: comparison

Dispersion: comparison

- **Standard deviation**
  - Useful for approximately normally distributed data
  - More difficult to interpret for asymmetric data
  - Sensitive to outliers

Dispersion: comparison

- **Standard deviation**
  - Useful for approximately normally distributed data
  - More difficult to interpret for asymmetric data
  - Sensitive to outliers

- **Range**
  - Useful for very small sample size
  - Depends on sample size (increases with sample size)

Dispersion: comparison

- **Standard deviation**
  - Useful for approximately normally distributed data
  - More difficult to interpret for asymmetric data
  - Sensitive to outliers

- **Range**
  - Useful for very small sample size
  - Depends on sample size (increases with sample size)

- **IQR, quartiles**
  - Always interpretable
  - Also allows to infer the skewness of the distribution (if median is also given)

Location and dispersion – Recommendations

Location and dispersion – Recommendations

- Report median (quartiles, IQR) for descriptive statistics
  - Reason: Simple and meaningful regardless whether data is symmetric or not
  - Wording: „Median (IQR) hematocrit value was 48 (46 to 51).“
- Mean and sd may be more informative for count data with many ties
  - “median (IQR) = 3 (2-3)” not very informative
- Always give location and dispersion measure

Graphical summary

Graphical summary

Categorical variables

- Pie chart (usually not recommended)
- Bar chart
- Dotplot (often preferred over bar chart)

Graphical summary

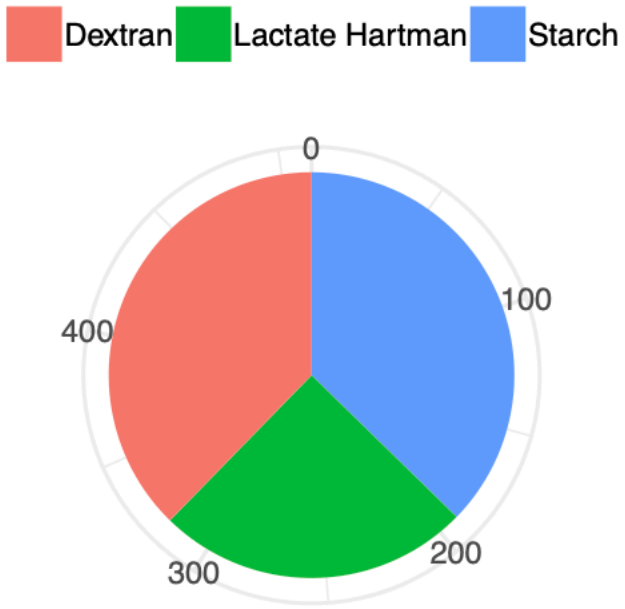
Categorical variables

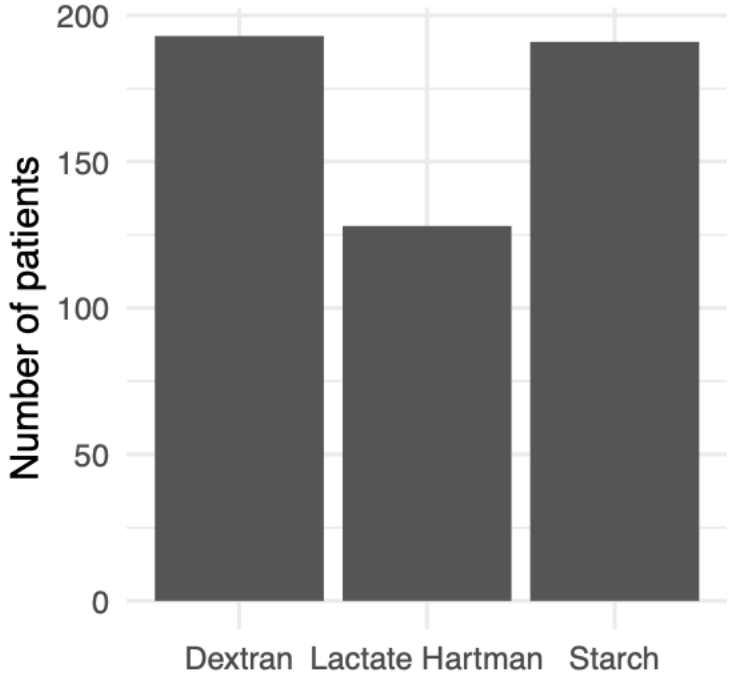
- Pie chart (usually not recommended)
- Bar chart
- Dotplot (often preferred over bar chart)

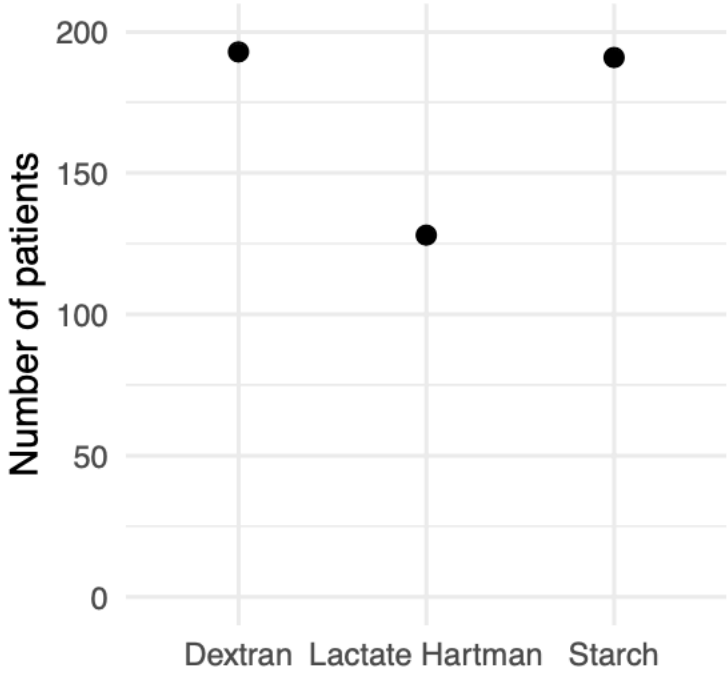
Continuous variables

- Histogram/density plot
- Boxplot

Pie, bar and dot







Histogram

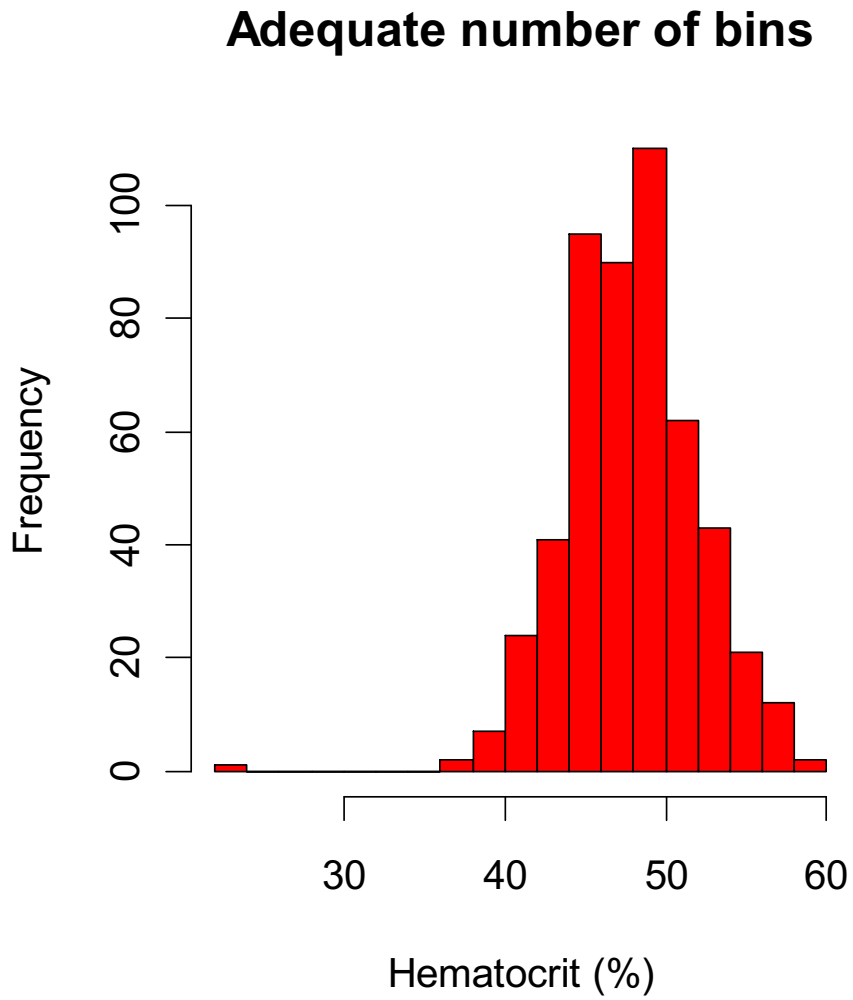
Histogram

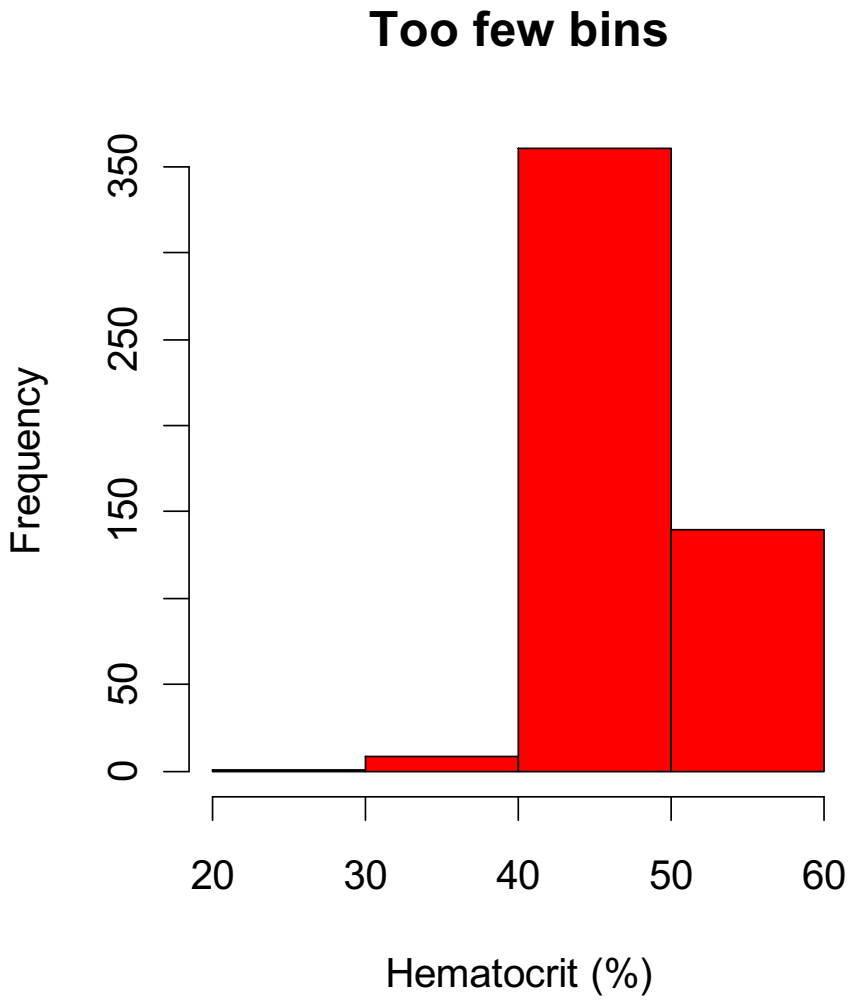
- Group values of a variable into bins of equal width; plot the number (or relative frequency) as a barchart

Histogram

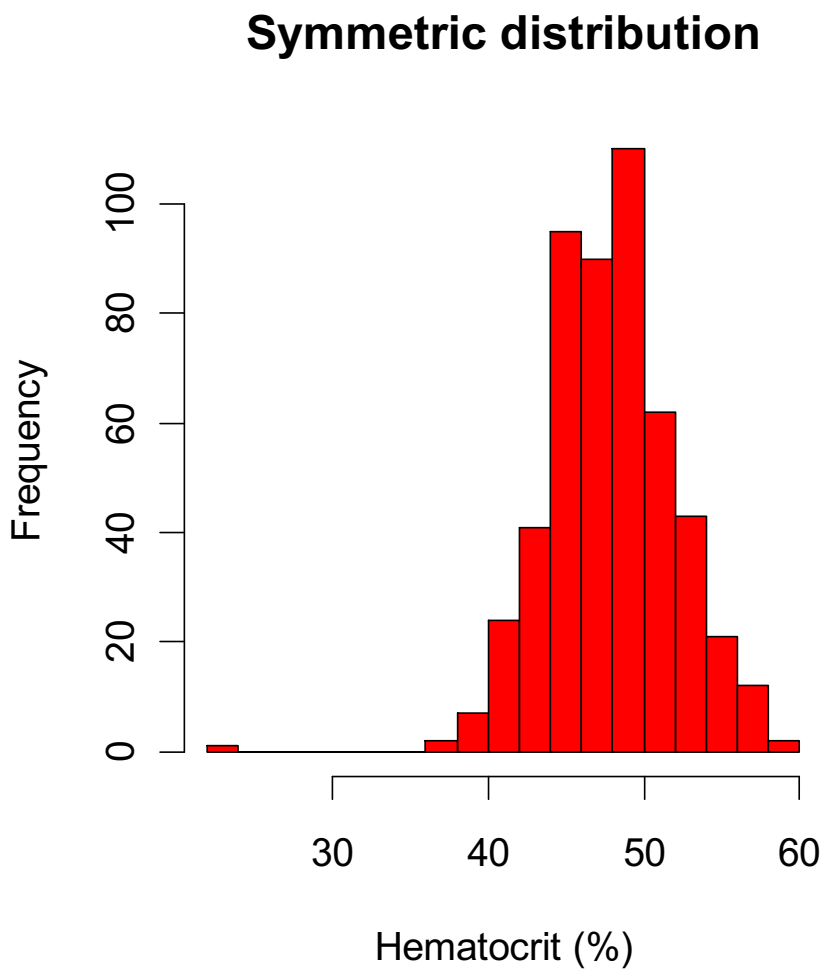
- Group values of a variable into bins of equal width; plot the number (or relative frequency) as a barchart
- Caveat: visual appearance may depend on the chosen number and location of bins
- Try several groupings of the data

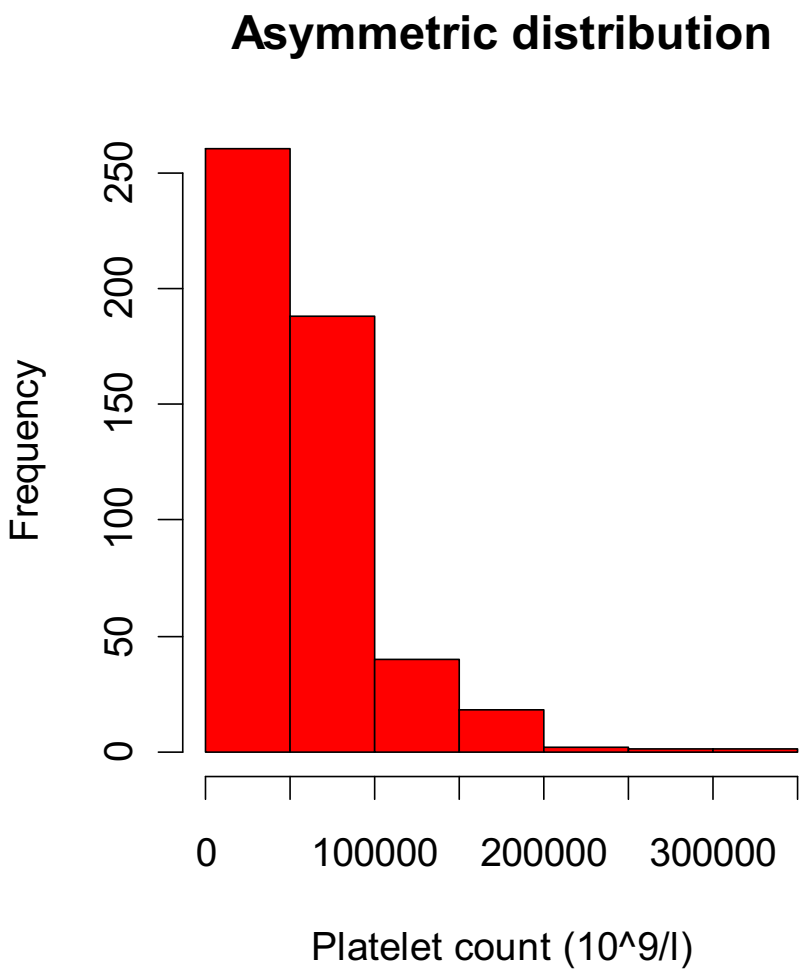
Histogram: examples



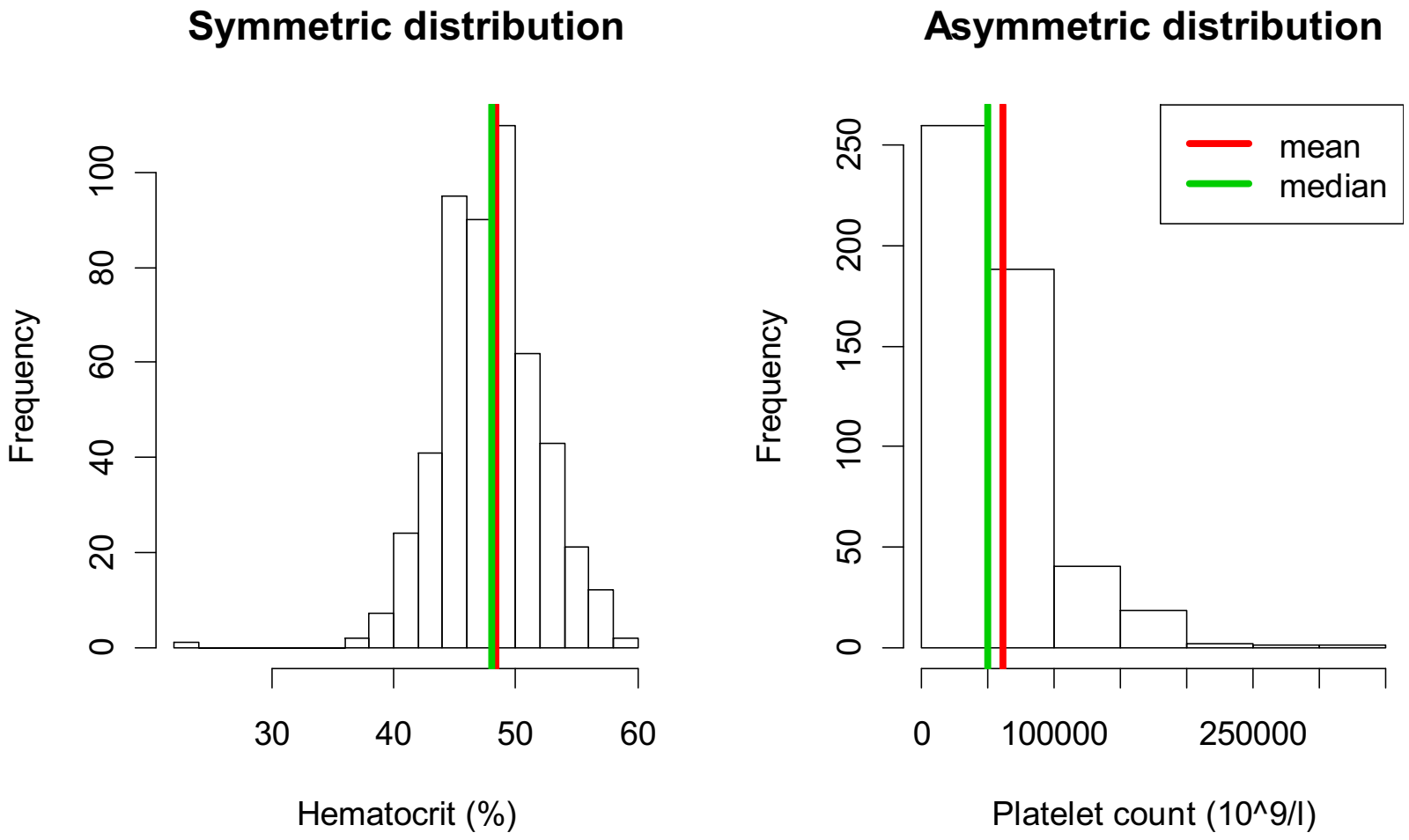


Histogram: examples





Histogram: examples

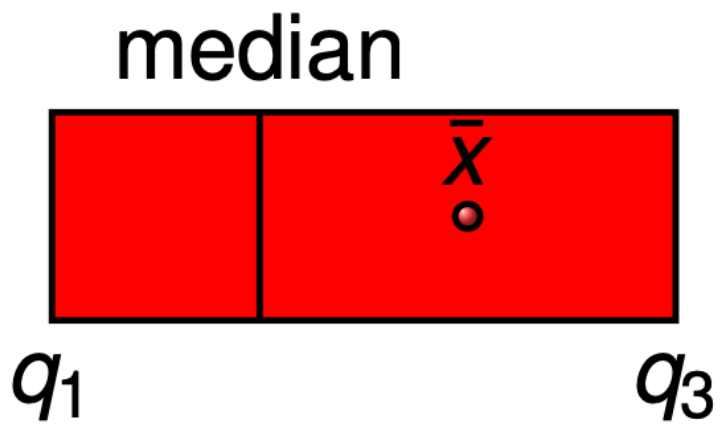


Boxplot

Boxplot

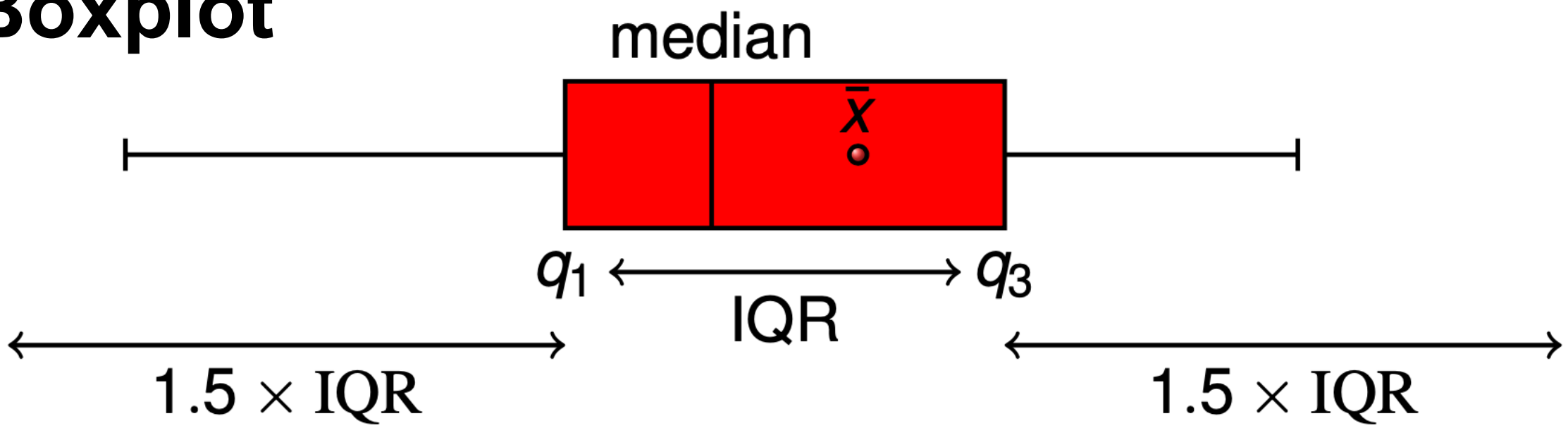
- Formal name: box-and-whisker plot

Boxplot



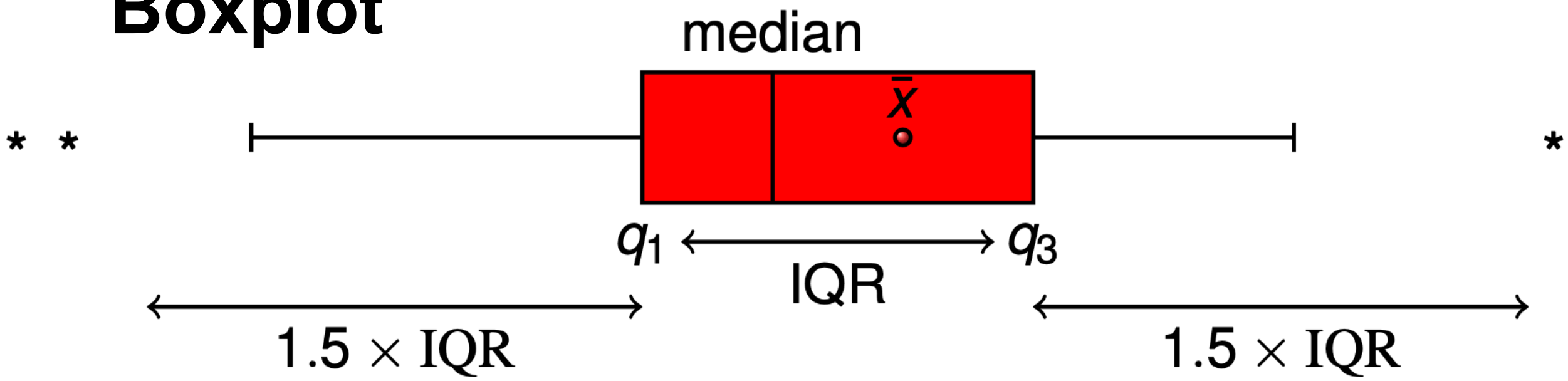
- Formal name: box-and-whisker plot
- Box: most common observations

Boxplot



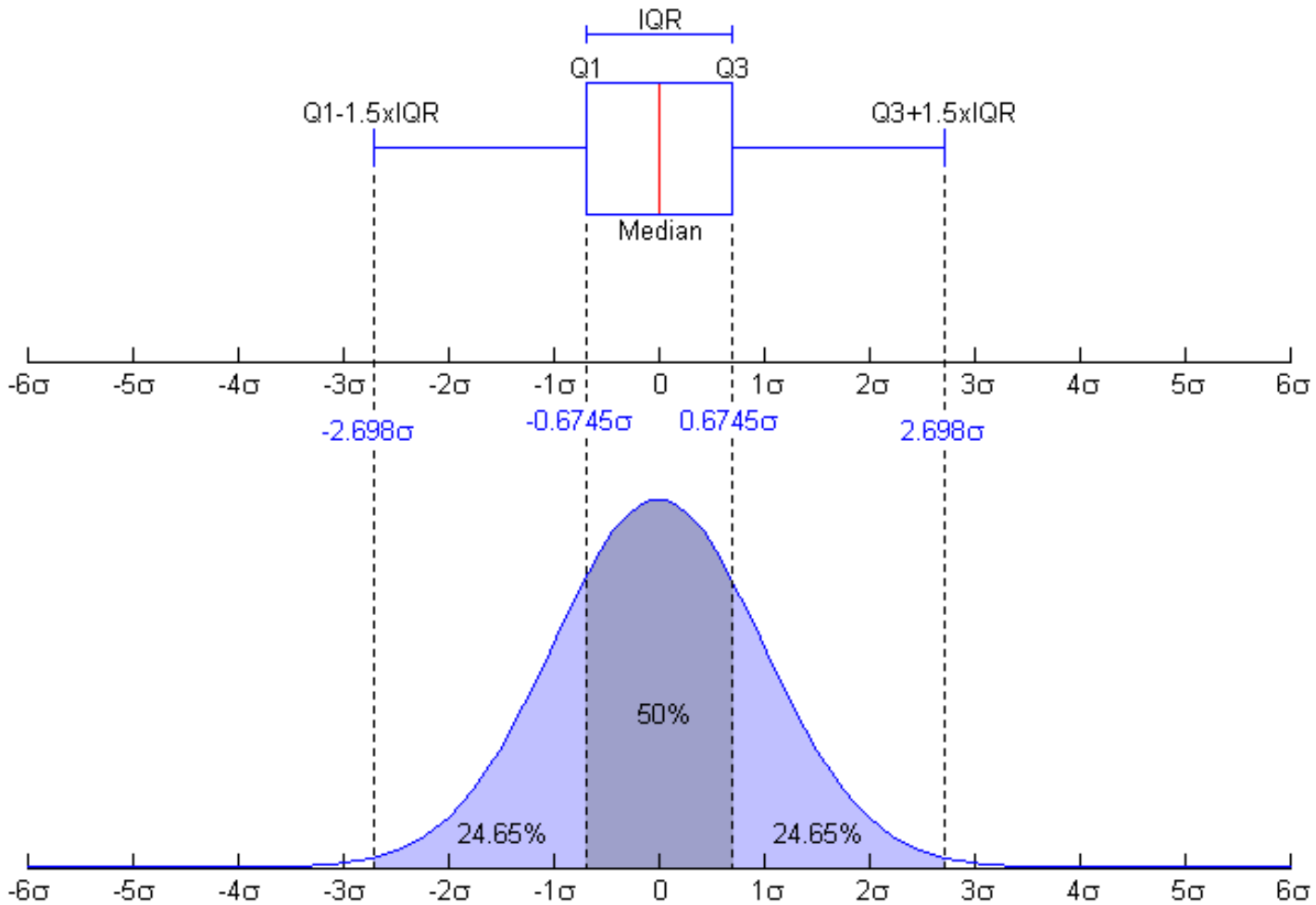
- Formal name: box-and-whisker plot
- Box: most common observations
- Whiskers: less common but still typical
  - From quartiles to the furthest away observation:
    ≥ q1 −1.5 × IQR and **≤ q3 +1.5 × IQR**

Boxplot

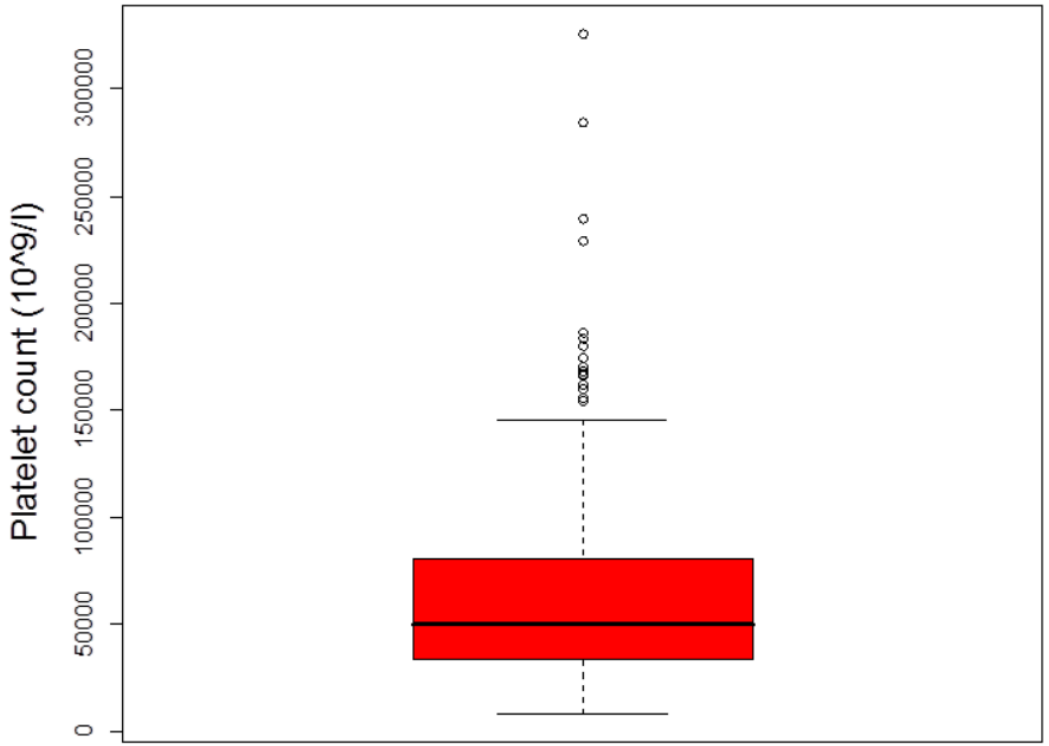


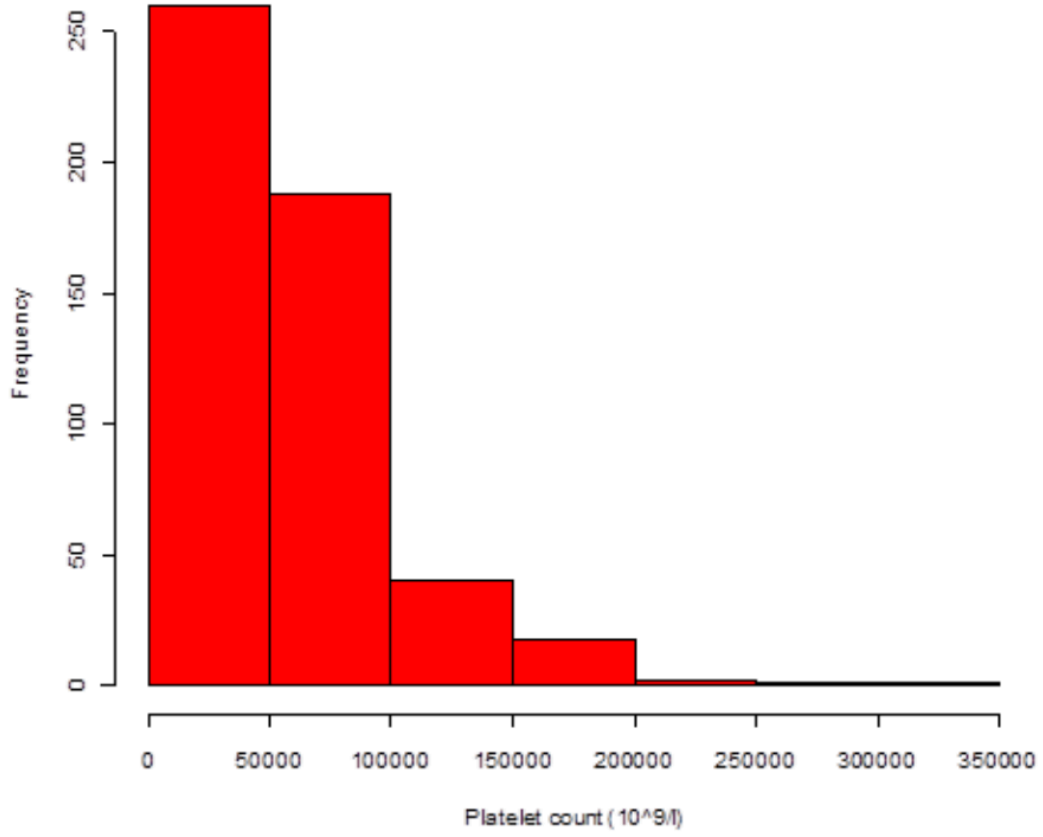
- Formal name: box-and-whisker plot
- Box: most common observations
- Whiskers: less common but still typical
  - From quartiles to the furthest away observation:
    ≥ q1 −1.5 × IQR and **≤ q3 +1.5 × IQR**
- Outliers: All points outside of the whiskers

Boxplot for normal distribution



Boxplot vs. Histogram

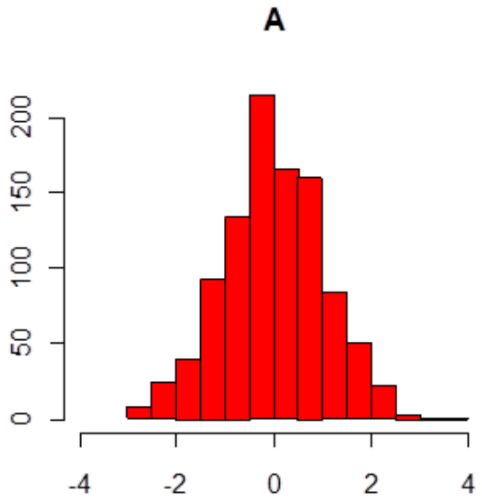


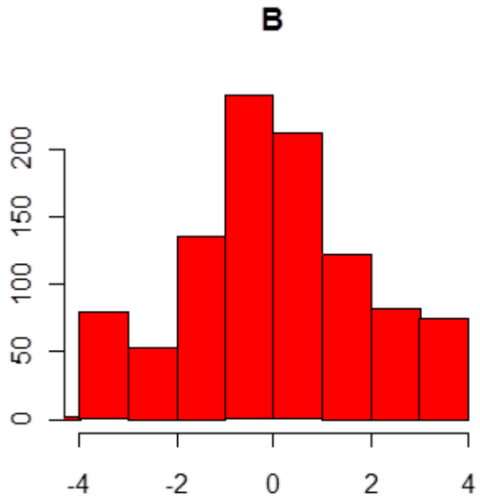


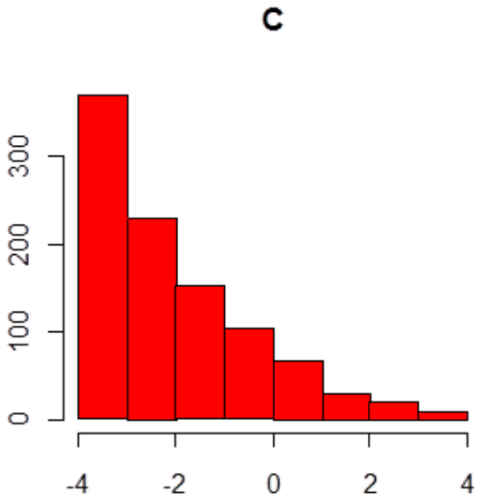
Boxplot: usage

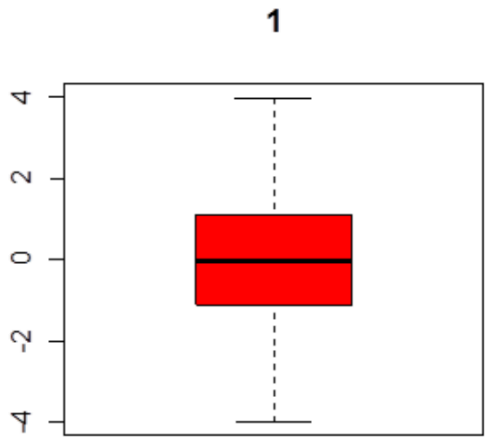
- Useful and concise summaries of the data
- Particularly useful for visual comparisons of multiple groups
- Small data sets: add individual values as dots

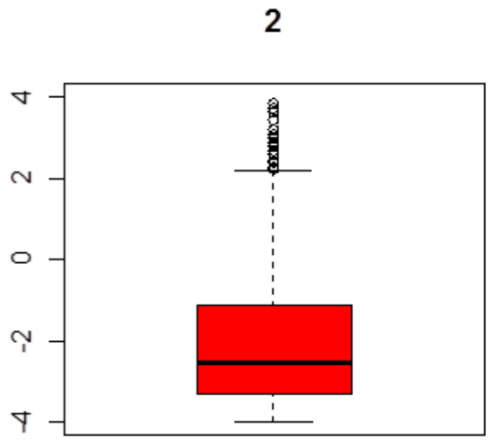
Quiz: Match histograms and boxplots

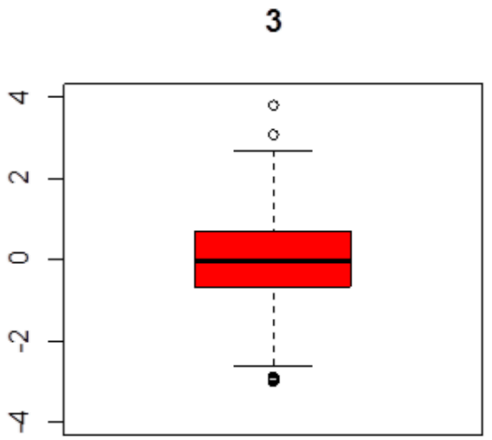












Summarizing association between continuous variables

Summarizing association between continuous variables

- Scatterplot

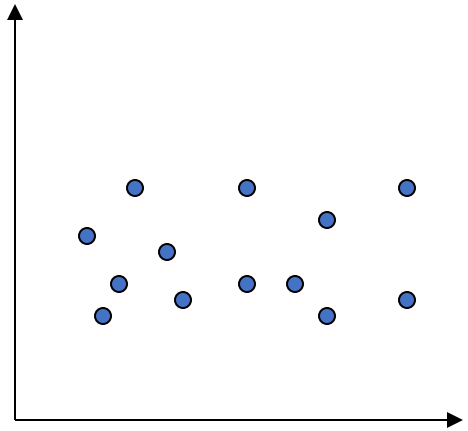
Summarizing association between continuous variables

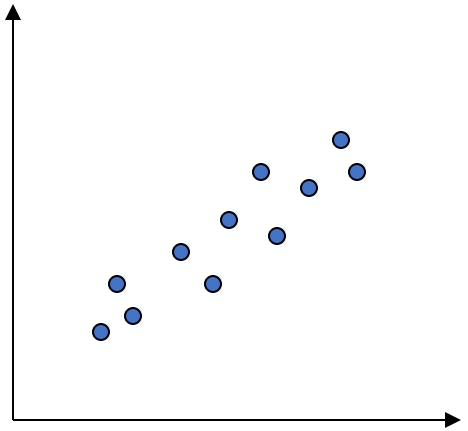
- Scatterplot
- x – independent variable (predictor, covariable)
- y – dependent variable (outcome, response)
- Sometimes not clear which variable is x or y

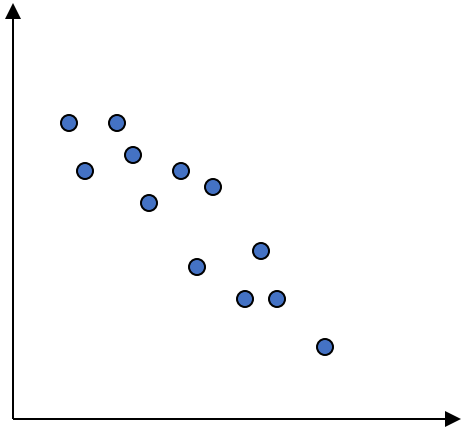
Summarizing association between continuous variables

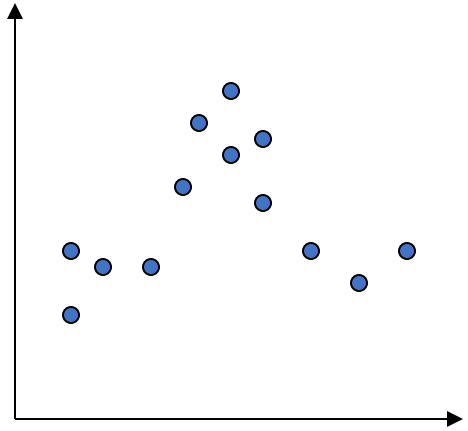
- Scatterplot
- x – independent variable (predictor, covariable)
- y – dependent variable (outcome, response)
- Sometimes not clear which variable is x or y
- Each observation is represented by one point
- Pattern of the points → relationship between variables

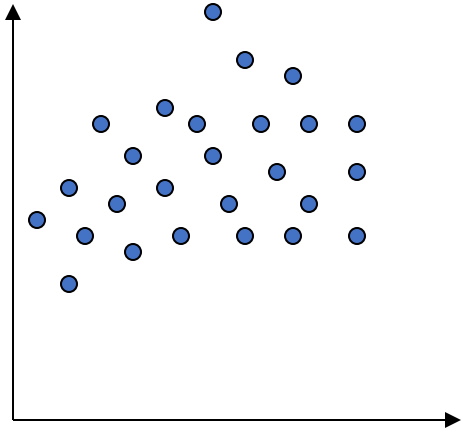
Scatterplot











No relationship

Positive correlation

Negative correlation

Linear relationship

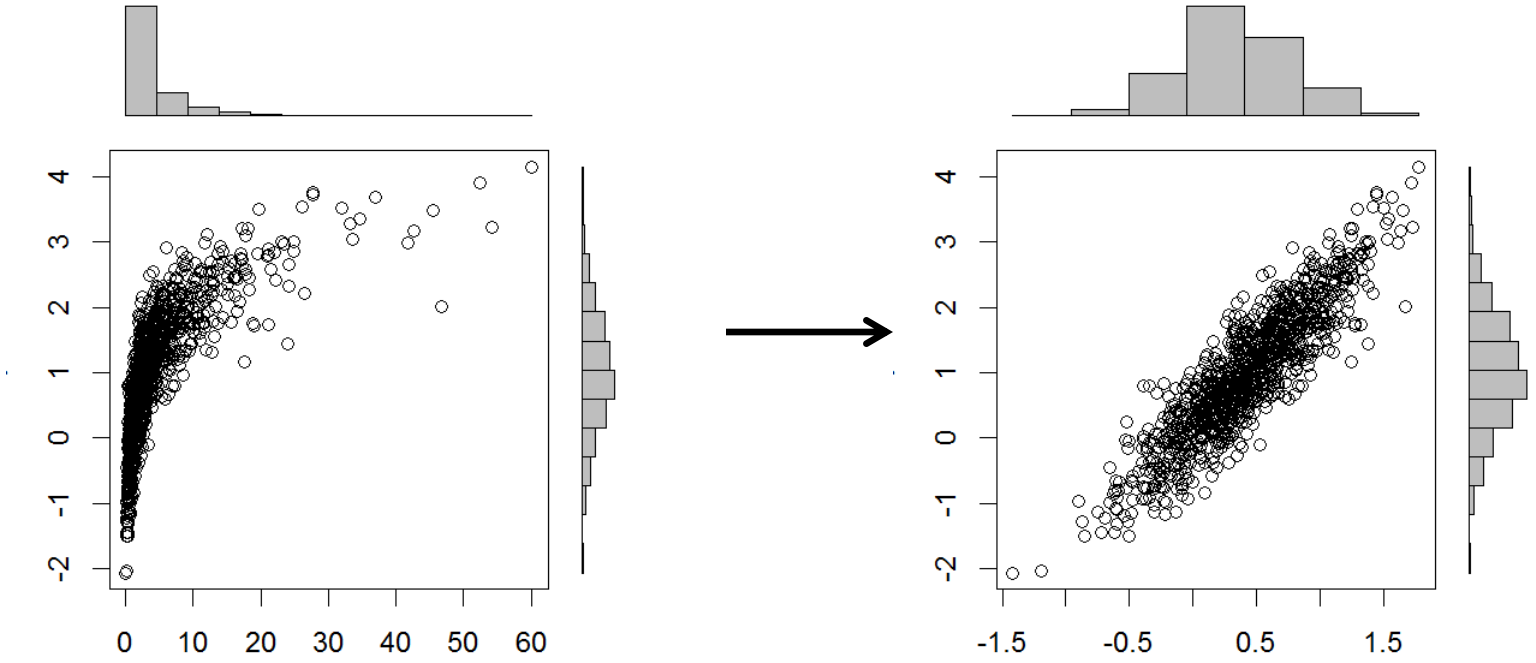
Non-linear relationship

Scatterplot

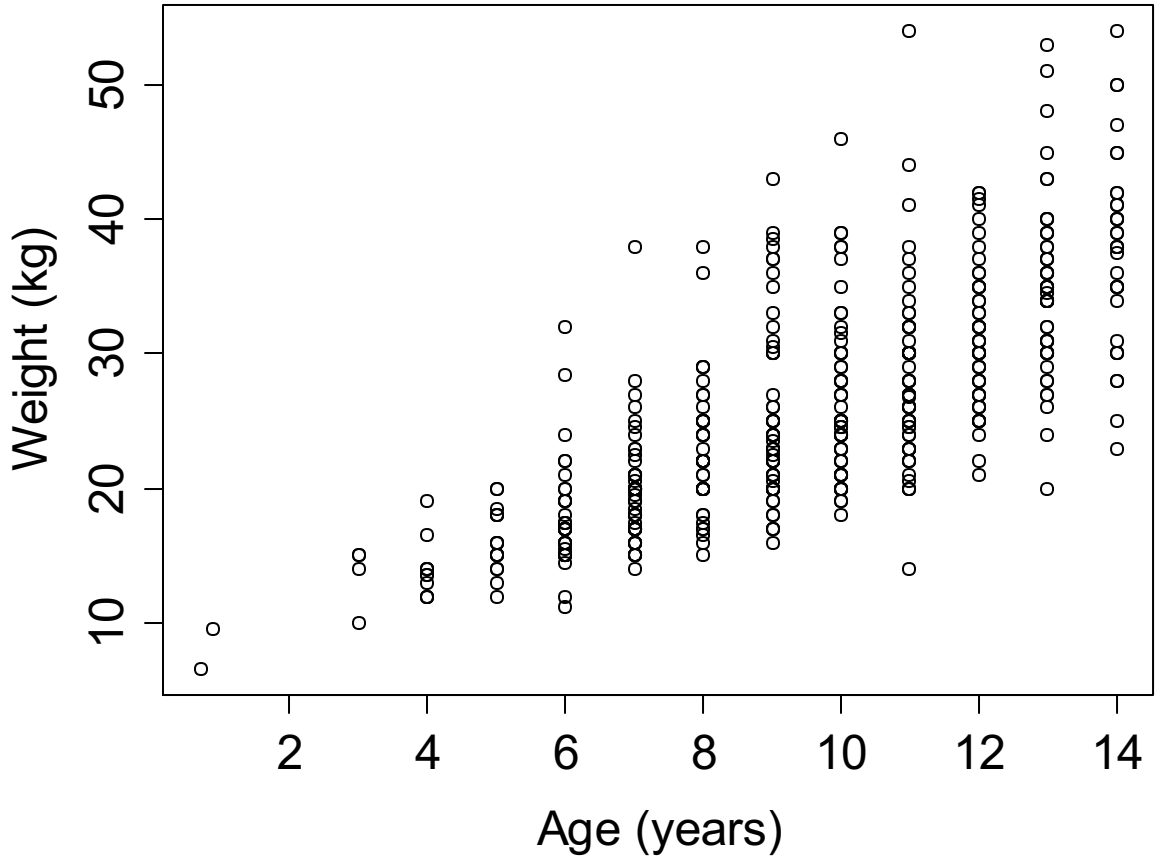
- Scatter plots are easier to interpret if variables are approximately normally distributed
- Transform data appropriately before plotting

Scatterplot

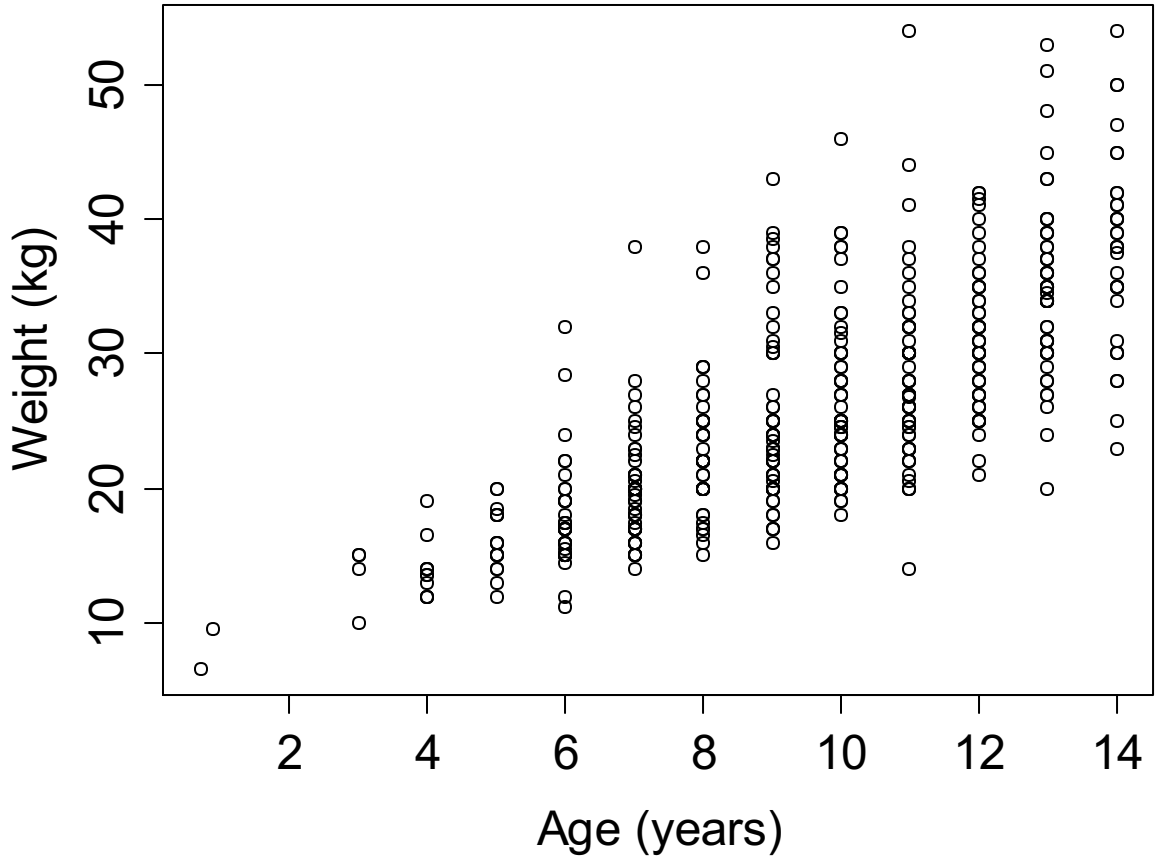
- Scatter plots are easier to interpret if variables are approximately normally distributed
- Transform data appropriately before plotting



Quiz: age vs. weight for dengue shock dataset



Quiz: age vs. weight for dengue shock dataset



- Roughly linear increase of weight with age
- A lot of variability, especially for higher age

Summary

- Data structure
- Data types: categorical/continuous
- Data summary
  - Numbers
    - Frequency, percentage, proportion
    - Location: mean, median
    - Dispersion: standard deviation, range, IQR
  - Graphs
    - Pie chart, bar chart, dotplots
    - Histogram, boxplot
    - Scatterplot

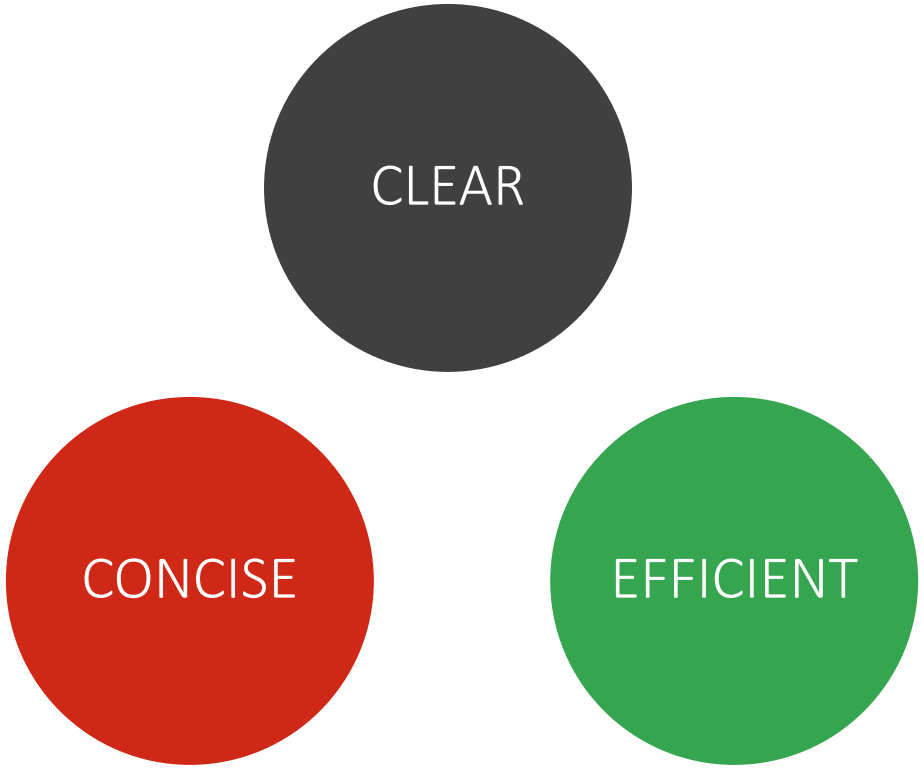
Making effective graphs and tables

Graphs and tables can be

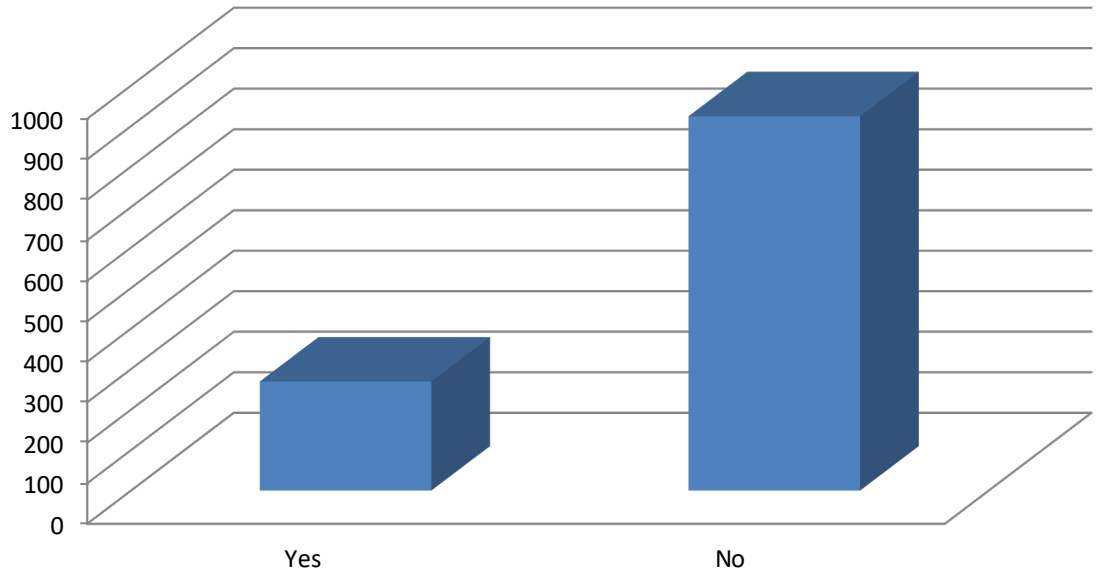
GOOD

BAD

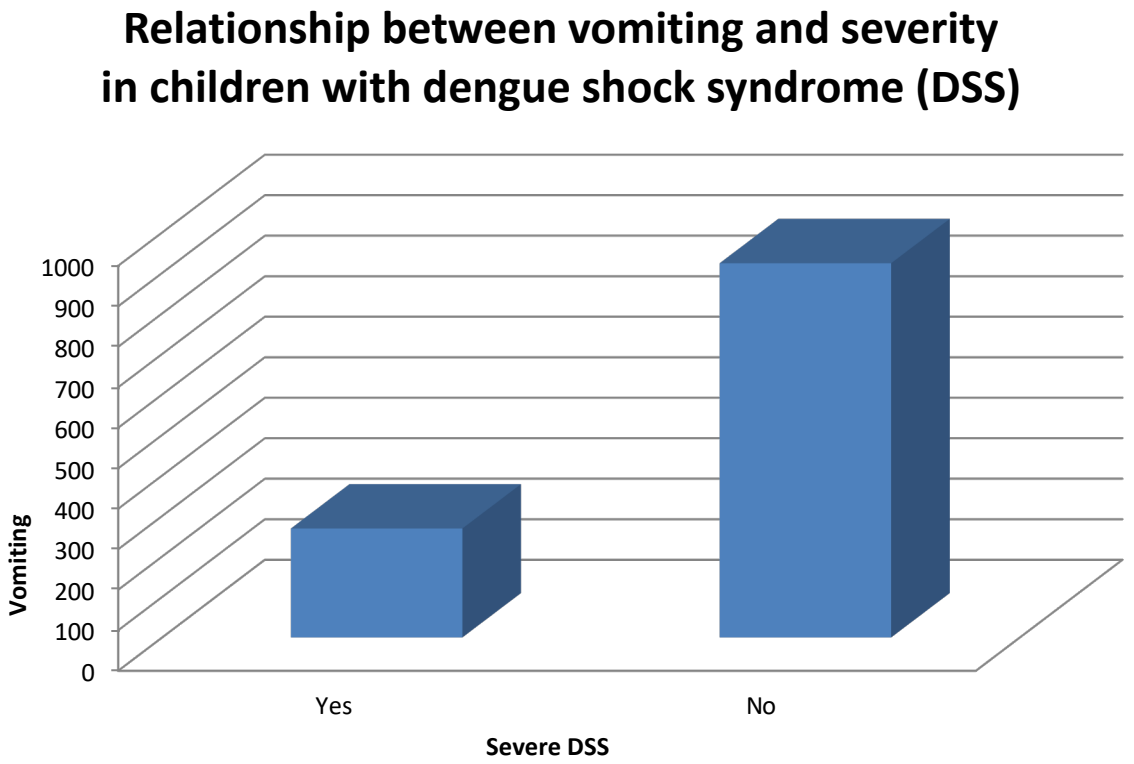
Graphs and tables should be



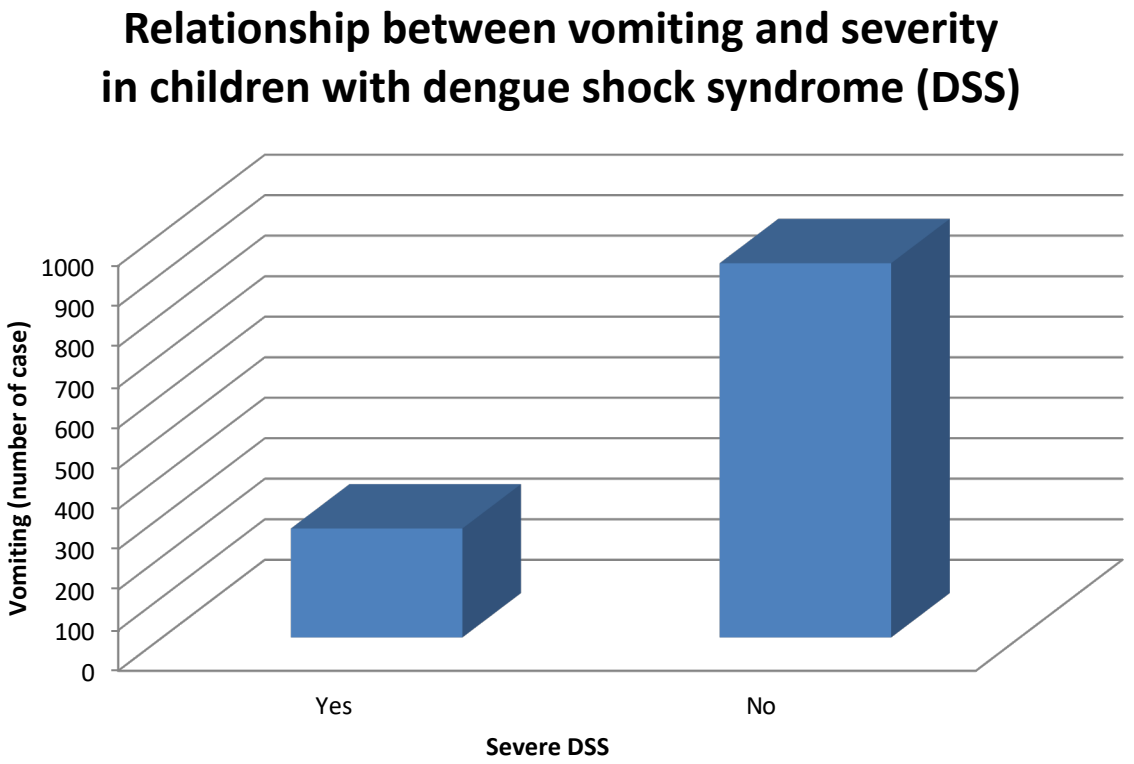
What is wrong with this graph?



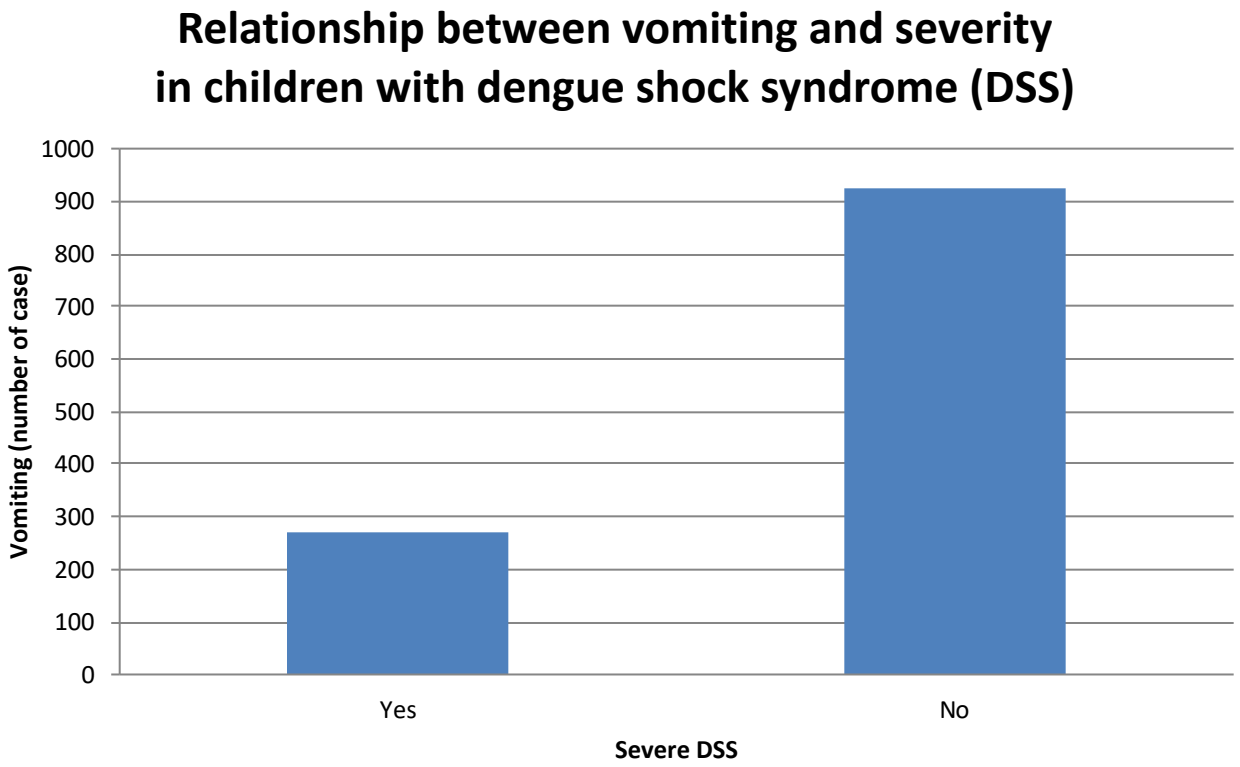
What is wrong with this graph?



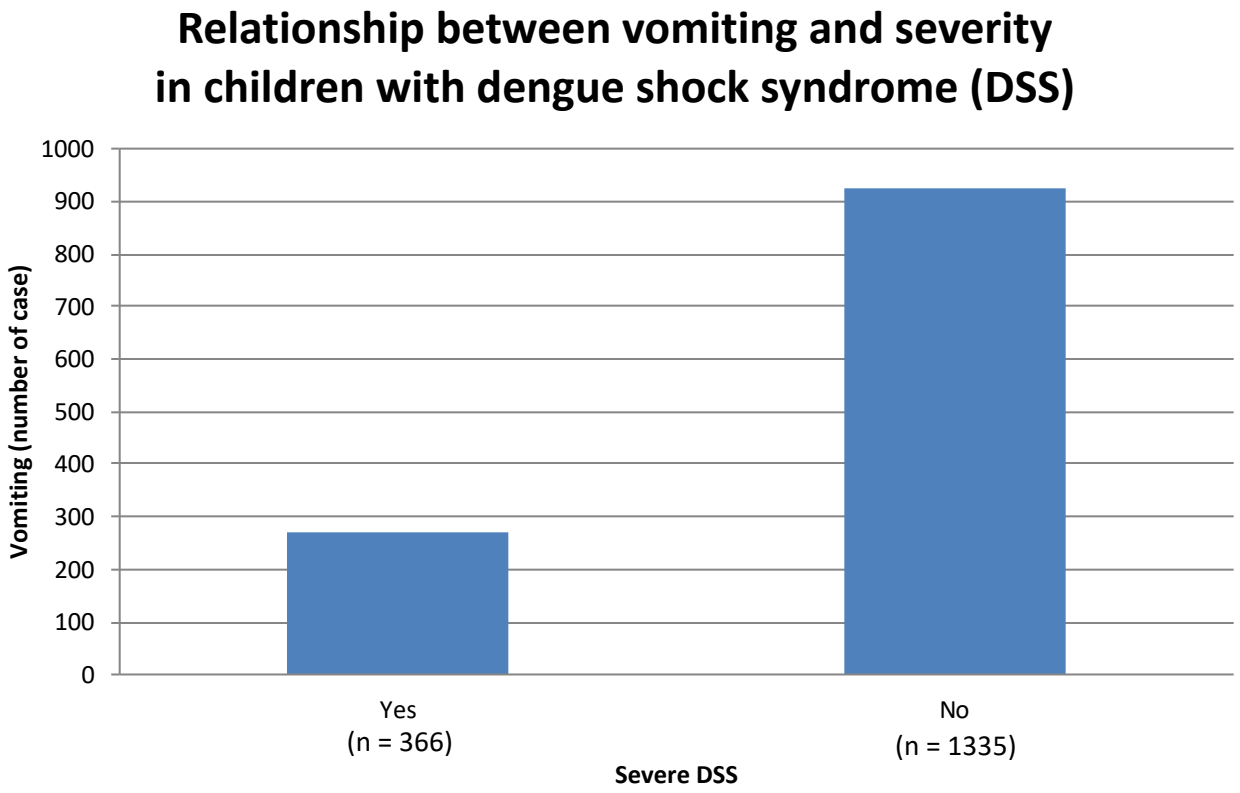
What is wrong with this graph?



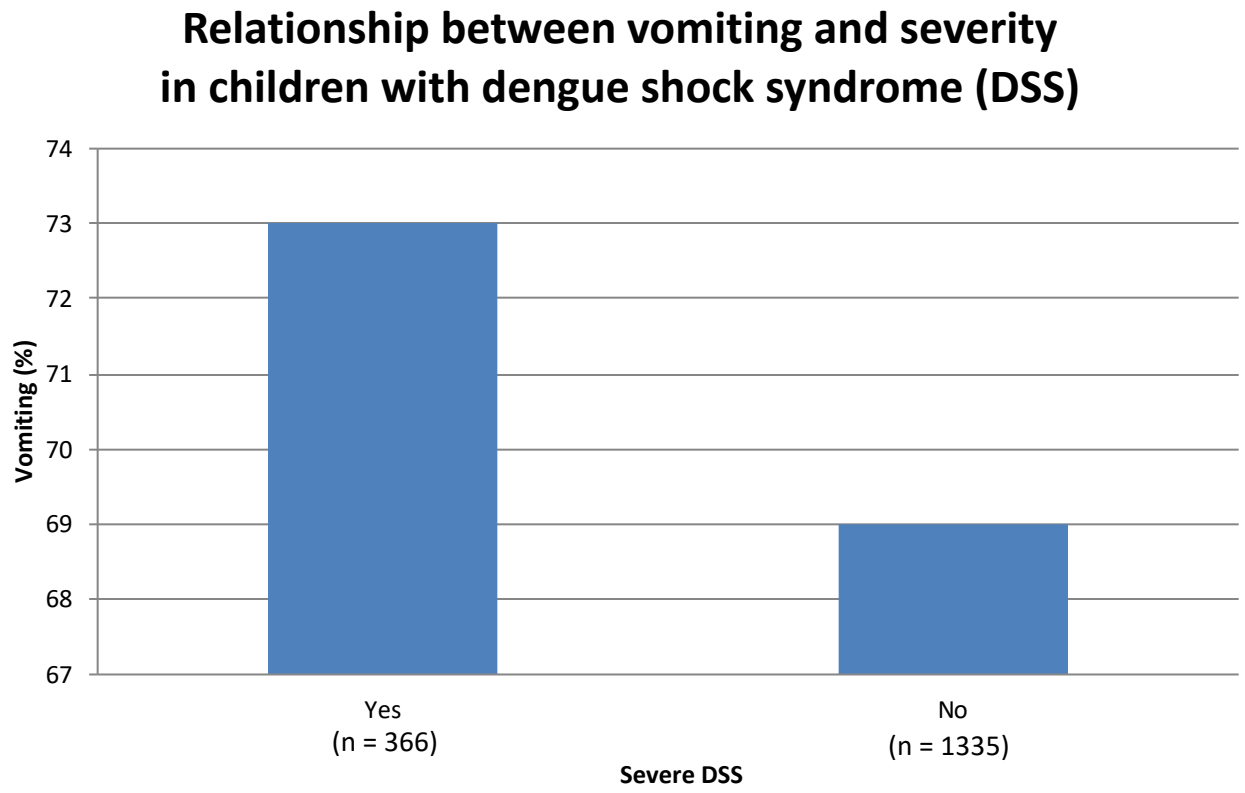
What is wrong with this graph?



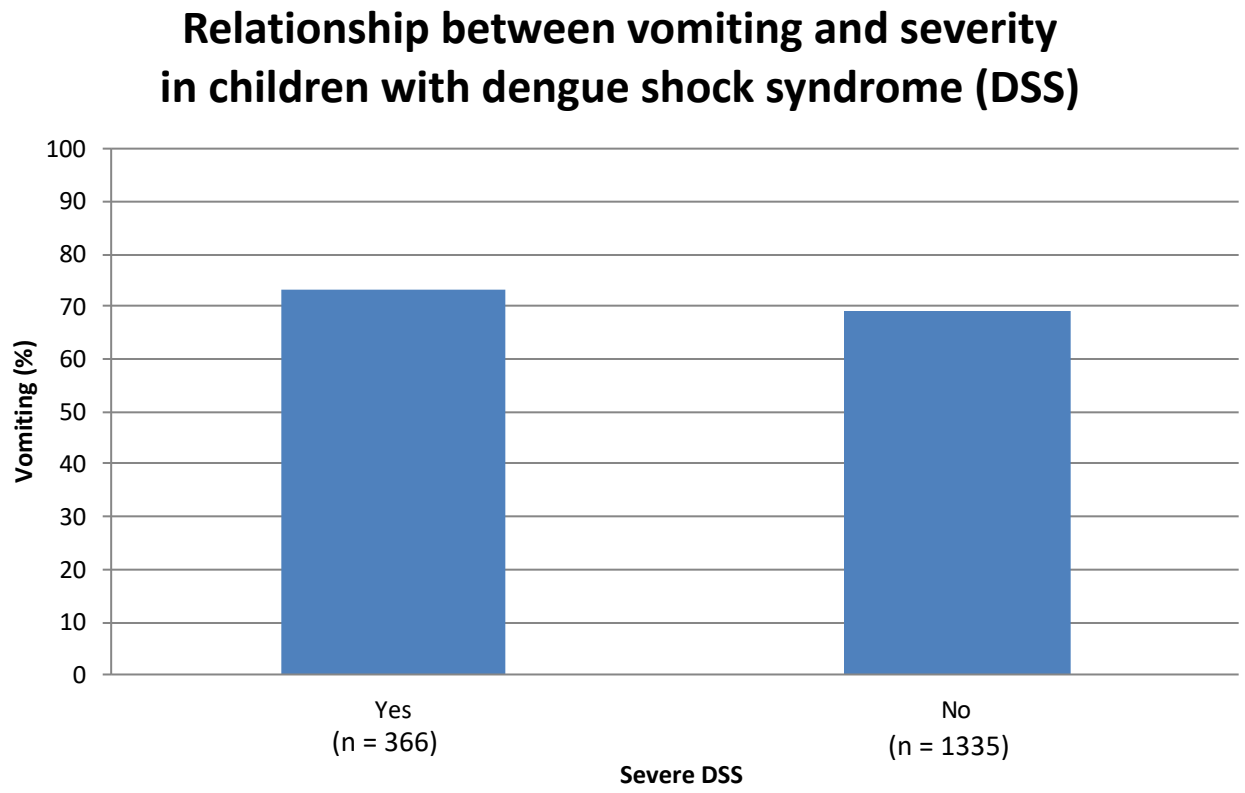
What is wrong with this graph?



What is wrong with this graph?



What is wrong with this graph?



What is wrong with this table?

Characteristics	Summary statistics	
Age	9.753	(7.012 – 12.18)
Weight	27.2	(20.14 – 35.26)
Hemorrhage		
None	493	(29.1)
Skin only	1153	(67.35)
Mucosal	73	(3.55)

What is wrong with this table?

Table 1. Baseline characteristics of the study participants

Characteristics	Summary statistics	
Age	9.753	(7.012 – 12.18)
Weight	27.2	(20.14 – 35.26)
Hemorrhage		
None	493	(29.1)
Skin only	1153	(67.35)
Mucosal	73	(3.55)

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment

Characteristics	Summary statistics	
Age	9.753	(7.012 – 12.18)
Weight	27.2	(20.14 – 35.26)
Hemorrhage		
None	493	(29.1)
Skin only	1153	(67.35)
Mucosal	73	(3.55)

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment (N = 1719)

Characteristics	Summary statistics	
Age	9.753	(7.012 – 12.18)
Weight	27.2	(20.14 – 35.26)
Hemorrhage		
None	493	(29.1)
Skin only	1153	(67.35)
Mucosal	73	(3.55)

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment (N = 1719)

Characteristics	Summary statistics	
Age	9.753	(7.012 – 12.18)
Weight	27.2	(20.14 – 35.26)
Hemorrhage		
None	493	(29.1)
Skin only	1153	(67.35)
Mucosal	73	(3.55)

Summary statistics = median (IQR) for continuous variable, frequency (%) for categorical variable

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment (N = 1719)

Characteristics	Summary statistics	
Age [year]	9.753	(7.012 – 12.18)
Weight [kg]	27.2	(20.14 – 35.26)
Hemorrhage		
None	493	(29.1)
Skin only	1153	(67.35)
Mucosal	73	(3.55)

Summary statistics = median (IQR) for continuous variable, frequency (%) for categorical variable

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment (N = 1719)

Characteristics	Summary statistics	
Age [year]	9.75	(7.01 – 12.18)
Weight [kg]	27.20	(20.14 – 35.26)
Hemorrhage		
None	493	(29.10)
Skin only	1153	(67.35)
Mucosal	73	(3.55)

Summary statistics = median (IQR) for continuous variable, frequency (%) for categorical variable

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment (N = 1719)

Characteristics	Summary statistics	
Age [year]	10	(7 – 12)
Weight [kg]	27	(20 – 35)
Hemorrhage		
None	493	(29)
Skin only	1153	(67)
Mucosal	73	(4)

Summary statistics = median (IQR) for continuous variable, frequency (%) for categorical variable

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment (N = 1719)

Characteristics	n	Summary statistics	
Age [year]	1710	10	(7 – 12)
Weight [kg]	1600	27	(20 – 35)
Hemorrhage	1719		
None		493	(29)
Skin only		1153	(67)
Mucosal		73	(4)

Summary statistics = median (IQR) for continuous variable, frequency (%) for categorical variable

What is wrong with this table?

Table 1. Baseline characteristics of the study participants at enrolment (N = 1719)

Characteristics	n	Summary statistics	
Age [year]	1710	10	(7 – 12)
Weight [kg]	1600	27	(20 – 35)
Hemorrhage	1719		
None		493	(29)
Skin only		1153	(67)
Mucosal		73	(4)

Summary statistics = median (IQR) for continuous variable, frequency (%) for categorical variable

Summary

- How to make an effective graph or table?
  - Ensure its CLARITY, PRECISION, and EFFICIENCY
  - Would you like to receive a KISS? Keep It Short and Simple

RECAP

Recap

- Biostatistics
  - Descriptive and inferential statistics
- Descriptive statistics
  - Data structure: tidy data
  - Data types: categorical/continuous
  - Data summary
    - Numbers
      - Frequency, percentage, proportion
      - Location: mean, median
      - Dispersion: standard deviation, range, IQR
    - Graphs
      - Pie chart, bar chart, dotplots
      - Histogram, boxplot
      - Scatterplot
- Making effective graphs and tables
  - Clarity, precision, efficiency