UNIVERSITY OF CALGARY

Resident Research Course:

Study Design and Analysis – Types of Data

Lecture 1

Dr. Grace Kwong, PhD, P.Stat.

grace.kwong@ucalgary.ca

These lecture notes based on others developed by Drs. Anna Funk, Alberto Nettel-Aguirre, Peter Faris, Sarah Rose, Luz Palacios-Derflingher

Lecture Objectives

- Statistics, study design, and critical thinking about bias are important

- Types of sampling methods

- Identify what types of data you are (or want to be) working with and describe it appropriately
  - Graphical presentation

- Statistical analysis plan

Lecture Objectives

- Statistics, study design, and critical thinking about bias are important
- Types of sampling methods
- Identify what types of data you are (or want to be) working with and describe it appropriately
  - Graphical presentation
- Statistical analysis plan

Lecture Objectives

- Statistics, study design, and critical thinking about bias are important

- Types of sampling methods

- Identify what types of data you are (or want to be) working with and describe it appropriately
  - Graphical presentation

- Statistical analysis plan

Lecture Objectives

- Statistics, study design, and critical thinking about bias are important
- Types of sampling methods
- Identify what types of data you are (or want to be) working with and describe it appropriately
    - Graphical presentation
- Statistical analysis plan **--> have one before you start**

Lecture Objectives

- Statistics, study design, and critical thinking about bias are important

- Types of sampling methods

- Identify what types of data you are (or want to be) working with and describe it appropriately
  - Graphical presentation

- Statistical analysis plan --> **have one before you start**

Why? Evaluate evidence from research

- **Example 1:**

- This week, you are working in a pediatric ED and you observe that all children who present with COVID-19 are discharged rapidly (mild infection). You talk about it with some other residents working this week and find that they have observed the same thing.

- You compile the data from these children and write a manuscript - you conclude that COVID-19 is mild in children.

- The news media contacts you, and you state in your interview that children across Canada should be in school, regardless of transmission trends

What type of study is this?
What do you think of the conclusions?

Why? Evaluate evidence from research

- **Example 2:**
  - You'd like to examine maternal risk factors for a rare type of congenital neurological anomaly that affects developmental outcomes.
  - You enroll all mothers of infants born with this congenital anomaly over the course of one year across all hospitals in Alberta. For your control group, you advertise for volunteer participants (mothers) with healthy births during the same time period at the same hospitals.
  - You compare the two groups and find that mothers of children with the congenital anomaly have a lower educational level and were less likely to be 100% compliant to prenatal vitamins.
  - You conclude that genetics (IQ-link) and prenatal nutritional deficiencies may interact to cause this congenital anomaly.

What kind of study is this?

What do you think of the conclusions?

Why Statistics?

- Statistics, following a robust study design, allows one to address questions such as:

Are the effects observed due to chance?

Are these conclusions valid?

How accurate are these estimates?

Considering the # of participants, and how they were enrolled, are these findings trustworthy??

Why Statistics?

- Statistics, following a robust study design, allows one to address questions such as:

…p values and results of other significance tests

Are the effects observed due to chance?

… critique statistical analysis plan and study design

Are these conclusions valid?

How accurate are these estimates?

…confidence intervals and other measures

Considering the # of participants, and how they were enrolled, are these findings trustworthy??

… sample selection, size calculation

Why statistics and study design?

- Statistical analysis (correctly done) is a <u>**vital**</u> part of good research, and is necessary to draw conclusions from any data you've collected

- Your study needs to be **well designed** and a **statistical analysis plan** should be worked into it from the beginning

A well-designed study, poorly analyzed, can be rescued by re-analysis, whereas a poorly designed study is beyond redemption, even using sophisticated statistics

Campbell MJ, Machin D. Uses and abuses of medical statistics, *Medical Statistics. A Commonsense Approach*, 1993. pg 4

Lecture objectives

- **Statistics, study design, and critical thinking about bias are important**
- Types of sampling methods
- Identify what types of data you are (or want to be) working with and describe it appropriately
  - Graphical presentation
- Statistical analysis plan **--> have one before you start**

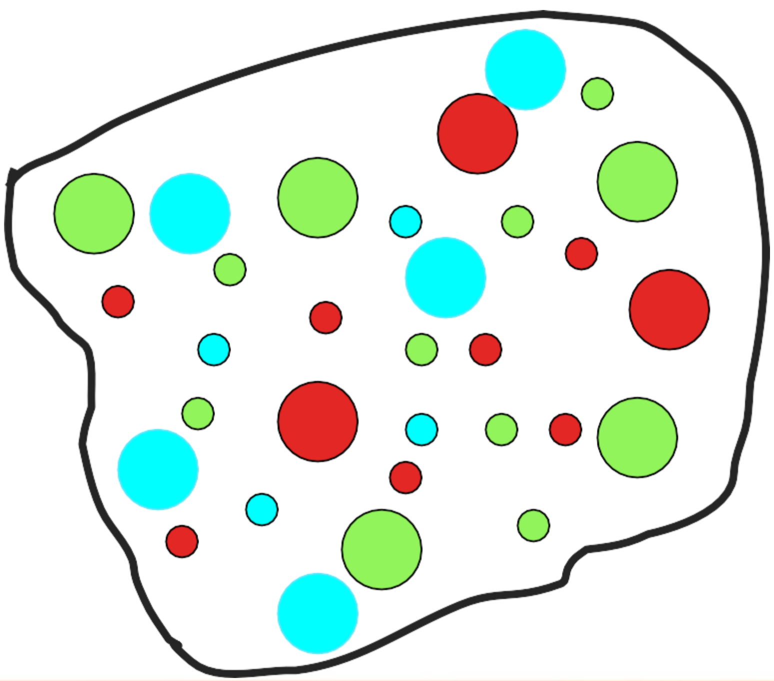
Population vs. Sample

Statistics allow us to draw inferences about a body of data (population) when only a part of the data (sample) is observed

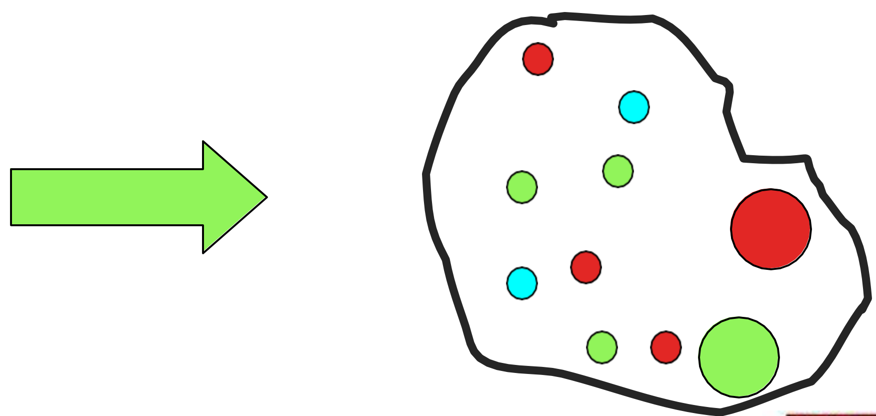
Population – ***all*** individuals or items under consideration pertaining to a study question

Sample – a selected part of the population from which information is obtained.

Everyone in the world with liver cancer



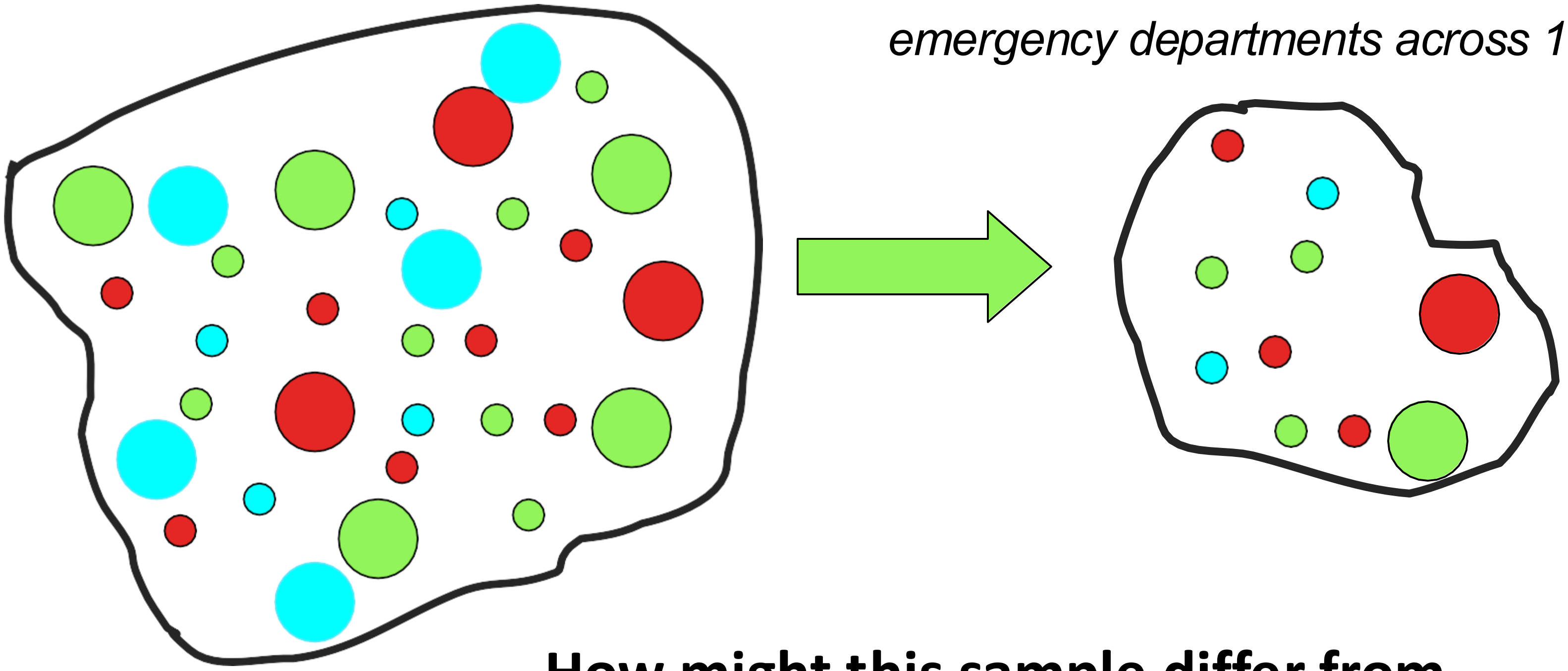
Persons with liver cancer presenting for treatment at one specialized centre in Cairo



Population vs. Sample

All children with COVID-19

Children with COVID-19 presenting to 50 emergency departments across 14 countries



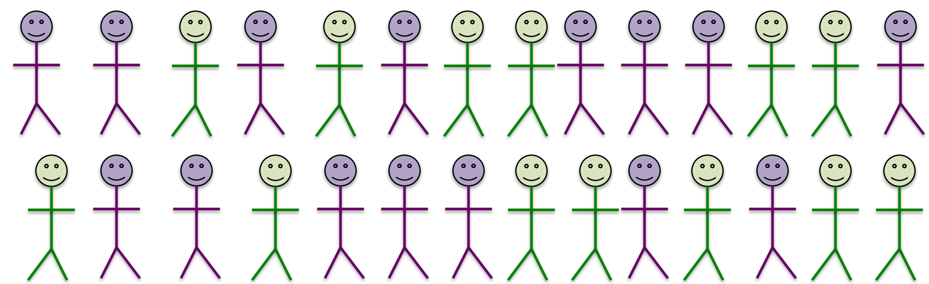
How might this sample differ from the population?

How *representative* is this sample?

Sampling Designs

- **Probabilistic (random) designs**
  - Simple random sampling
  - Systematic sampling
  - Cluster sampling
  - Stratified sampling

All children with a rare genetic disorder in Alberta



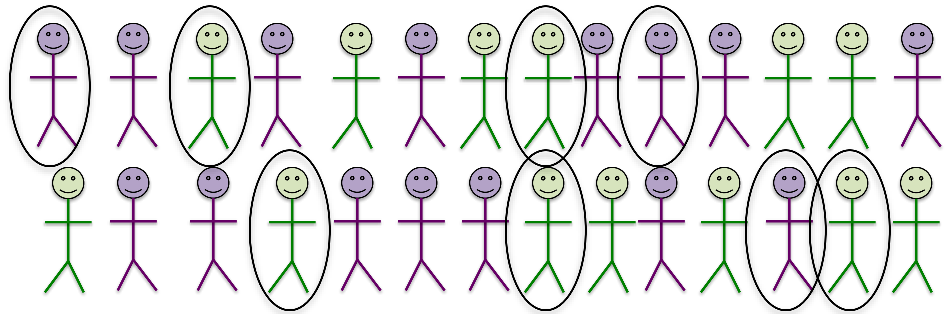
Demographic characteristic 1
(e.g. male or low socioeconomic status)

Demographic characteristic 2
(e.g. female or high socioeconomic status)

Sampling Designs

- **Probabilistic (random) designs**
  - **Simple random sampling**
  - Systematic sampling
  - Cluster sampling
  - Stratified sampling

All children with a rare genetic disorder in Alberta



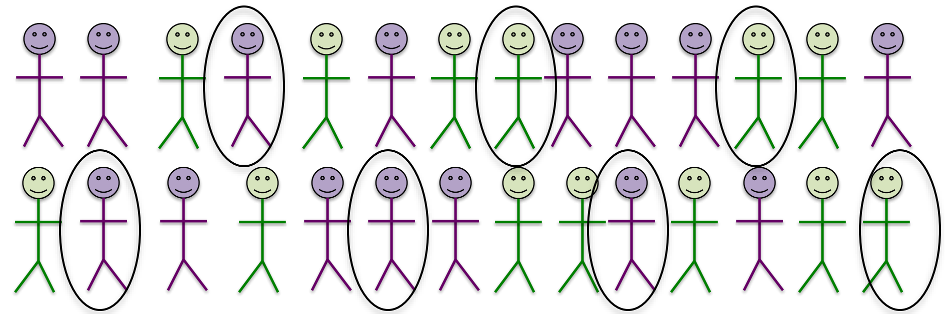
Demographic characteristic 1
(e.g. male or low socioeconomic status)

Demographic characteristic 2
(e.g. female or high socioeconomic status)

Sampling Designs

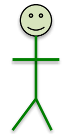
- **Probabilistic (random) designs**
  - Simple random sampling
  - **Systematic sampling**
  - Cluster sampling
  - Stratified sampling

All children with a rare genetic disorder in Alberta



(Every fourth child registered)

Demographic characteristic 1
(e.g. male or low socioeconomic status)

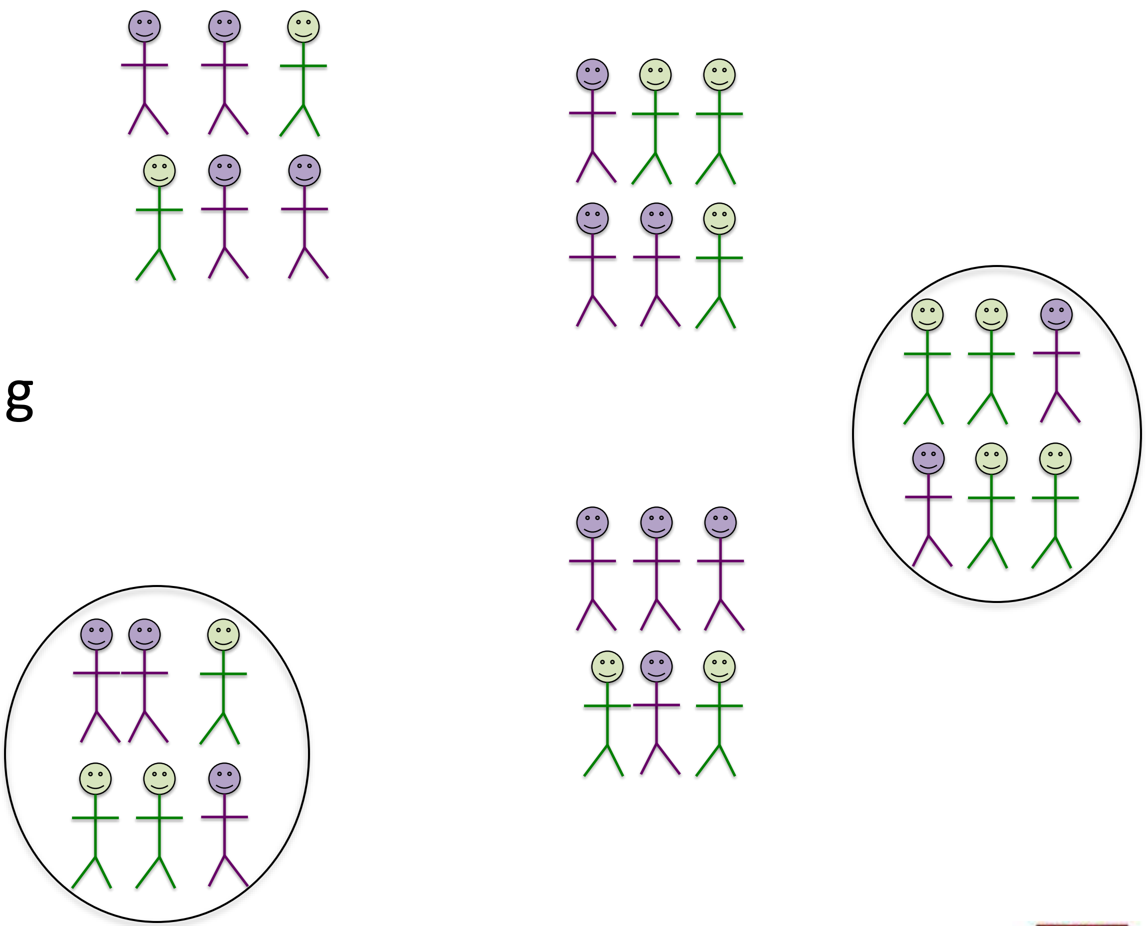


Demographic characteristic 2
(e.g. female or high socioeconomic status)

Sampling Designs

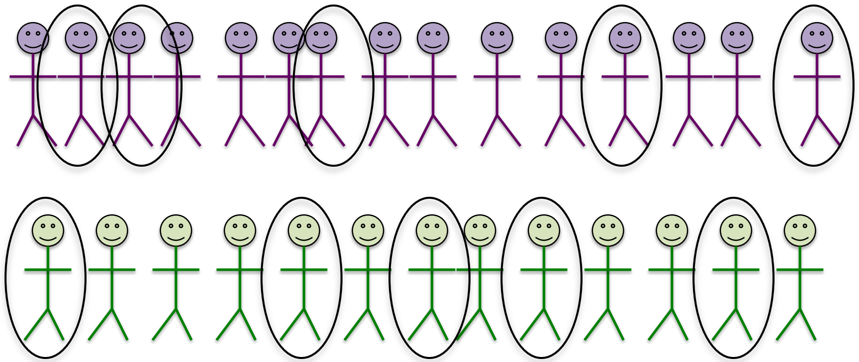
- **Probabilistic (random) designs**
  - Simple random sampling
  - Systematic sampling
  - **Cluster sampling**
  - Stratified sampling

All children with a rare genetic disorder in Alberta



Sampling Designs

All children with a rare genetic disorder in Alberta



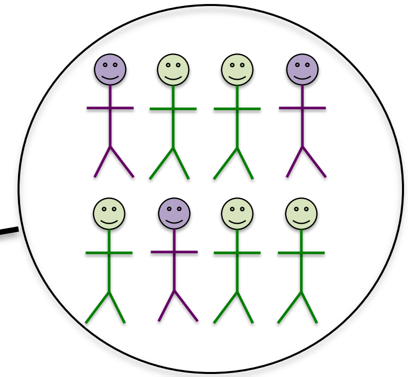
- **Probabilistic (random) designs**
  - Simple random sampling
  - Systematic sampling
  - Cluster sampling
  - **Stratified sampling**

Demographic characteristic 1
(e.g. male or low socioeconomic status)

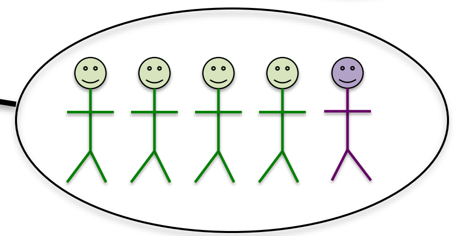
Demographic characteristic 2
(e.g. female or high socioeconomic status)

Sampling Designs

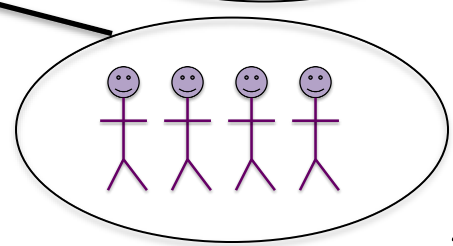
- **Non-probabilistic (deterministic) designs**
  - Convenience sample
  - Volunteer
  - Purposive
  - Snowball sampling



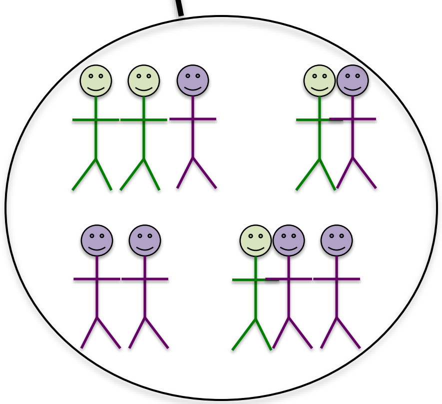
Kids who attended clinic A in 2020



Kids who's parent saw the online study notice and wanted their child to take part



Kids of lower socio-economic status families



Participants and their friends / acquaintances

Sampling - takeaway

- **Why should you know sampling strategies?**

  - For you (medical researcher)
    - Choose appropriate one
    - How *representative* is your study population?
    - What kind of statistics / analysis will you need?
    - How strong can you state your conclusions

  - For knowledge users and stakeholders
    - How seriously should we take these findings?

Lecture objectives

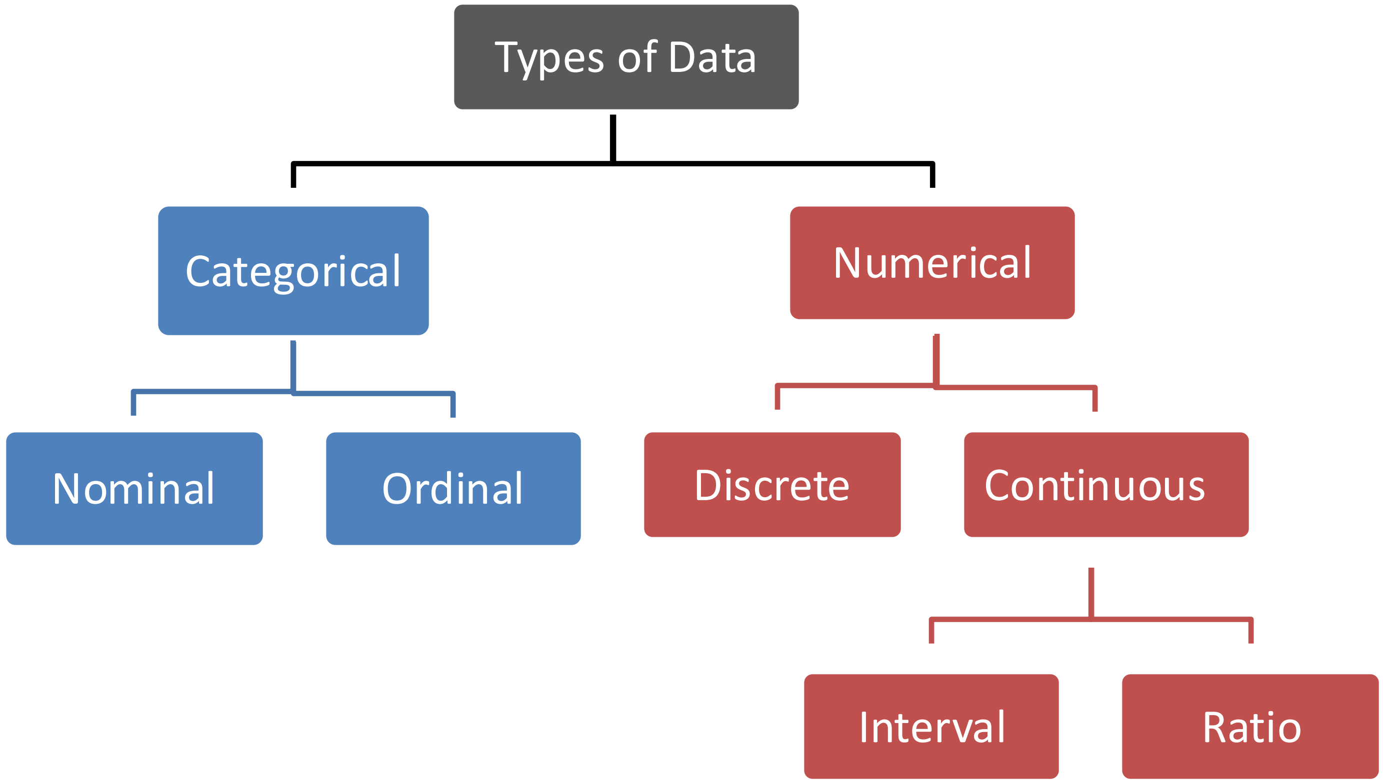
- **Statistics, study design, and critical thinking about bias are important**

- Types of sampling methods

- Identify what types of data you are (or want to be) working with and describe it appropriately
  - Graphical presentation

- Statistical analysis plan **--> have one before you start**

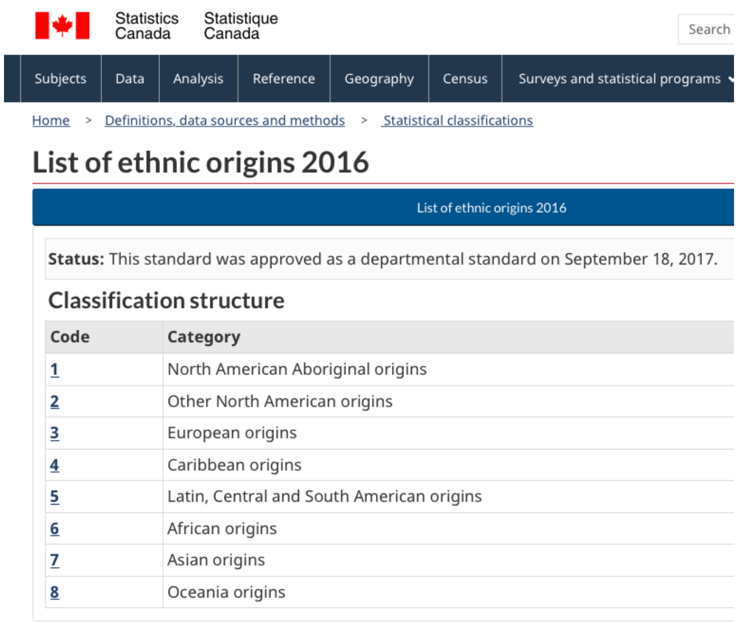
Types of Data Overview



Categorical Data

- **Nominal** (names) **variables**:
  - Labels based on an attribute of each element
  - Two or more categories, no ordering

Biological sex



Statistics Canada Statistique Canada

Search

Subjects | Data | Analysis | Reference | Geography | Census | Surveys and statistical programs

Home > Definitions, data sources and methods > Statistical classifications

List of ethnic origins 2016

List of ethnic origins 2016

Status: This standard was approved as a departmental standard on September 18, 2017.

Classification structure

Code	Category
1	North American Aboriginal origins
2	Other North American origins
3	European origins
4	Caribbean origins
5	Latin, Central and South American origins
6	African origins
7	Asian origins
8	Oceania origins

Disease classification (when <u>not</u> related to severity):

E.g. Types of respiratory illness

- Pneumonia
- Asthma
- Emphysema

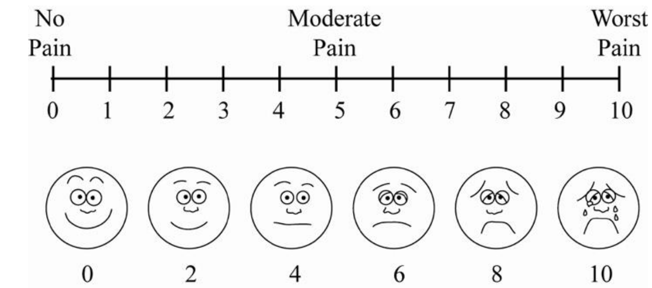
E.g. location (segment) of liver tumour

E.g. first symptom of COVID-19 experienced
- fever, chills, headache, cough

Categorical Data

- **Ordinal** (ordered) **variables**:
  - Have properties of nominal data **and** can be used to rank or order the observations
  - Two or more categories, ordered in magnitude
  - Difference between successive not necessarily the same

Visual analog scale



Disease classification (related to severity):

E.g. Acute respiratory distress syndrome (ARDS)
- Mild
- Moderate
- Severe

Government of Canada | Gouvernement du Canada

Federal tax rates for 2020

- 15% **on the first** $48,535 of taxable income, **plus**
- 20.5% **on the next** $48,534 of taxable income (on the portion of taxable income over 48,535 up to $97,069), **plus**
- 26% **on the next** $53,404 of taxable income (on the portion of taxable income over $97,069 up to $150,473), **plus**
- 29% **on the next** $63,895 of taxable income (on the portion of taxable income over 150,473 up to $214,368), **plus**
- 33% of taxable income **over** $214,368

Categorical Data

- **Special case – dichotomous (binary) variable**
  - Only two possible options
  - May indicate an order but not necessarily

Biological sex
(?)

old - young

Presence of fever?

☐ Yes

☒ No

Note: Yes / No often turned into a ‘dummy’ variable (0/1) in analysis

Displaying categorical data

- **Tables / Summaries**
  - Frequencies (number of observations that fit in that category) with percentages (proportions)
    - **DO NOT CALL IT A RATE**

Example: Cardiac arrest

In a study, there are **747 males** and **434 females**

- **276 males** and **195 females** had a cardiac arrest

Sex	Frequency	%
Male	747	63.25
Female	434	36.75
Total	1181	100.00

Cardiac arrest	Frequency	%
Yes	471	39.88
No	710	60.12
Total	1181	100.00

Tabulation

Example: Cardiac arrest

N or n Row % Col %	Cardiac arrest	No cardiac arrest	Total
Male	276 36.95 58.60	471 63.05 66.34	747 100.00 63.25
Female	195 44.93 41.40	239 55.07 33.66	434 100.00 36.75
Total	471 39.88 100.00	710 60.12 100.00	1181 100.00 100.00

Cross Tabulation

Displaying categorical data

- **95% Confidence Intervals (95% CI)**

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

$\hat{\mathbf{p}}$ = proportion
z = value from a standard normal distribution to represent the confidence level you desire (95% = 1.96)
n = sample size

E.g. In the cardiac arrest example, 276 (36.95%) of the 747 male had a cardiac arrest.

What are the lower and upper 95% CIs?

Altogether what does this confidence interval ‘range’ mean?

Displaying categorical data

- **95% Confidence Intervals (95% CI)**

$$\hat{p} \pm z^{*} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

p̂ = proportion

z = value from a standard normal distribution to represent the confidence level you desire (95% = 1.96)

n = sample size

$$\mathbf{95\%\ CIs} = \mathbf{0.3695 \pm 1.96 \times \sqrt{\frac{0.3695 \times (1-0.3695)}{747}}}$$
$$= \mathbf{(0.3349, 0.4041)}$$
$$= \mathbf{(33.5\%, 40.4\%)}$$

E.g. In the cardiac arrest example, 276 (36.95%) of the 747 male had a cardiac arrest.

What are the lower and upper 95% CIs?

Altogether what does this confidence interval 'range' mean?

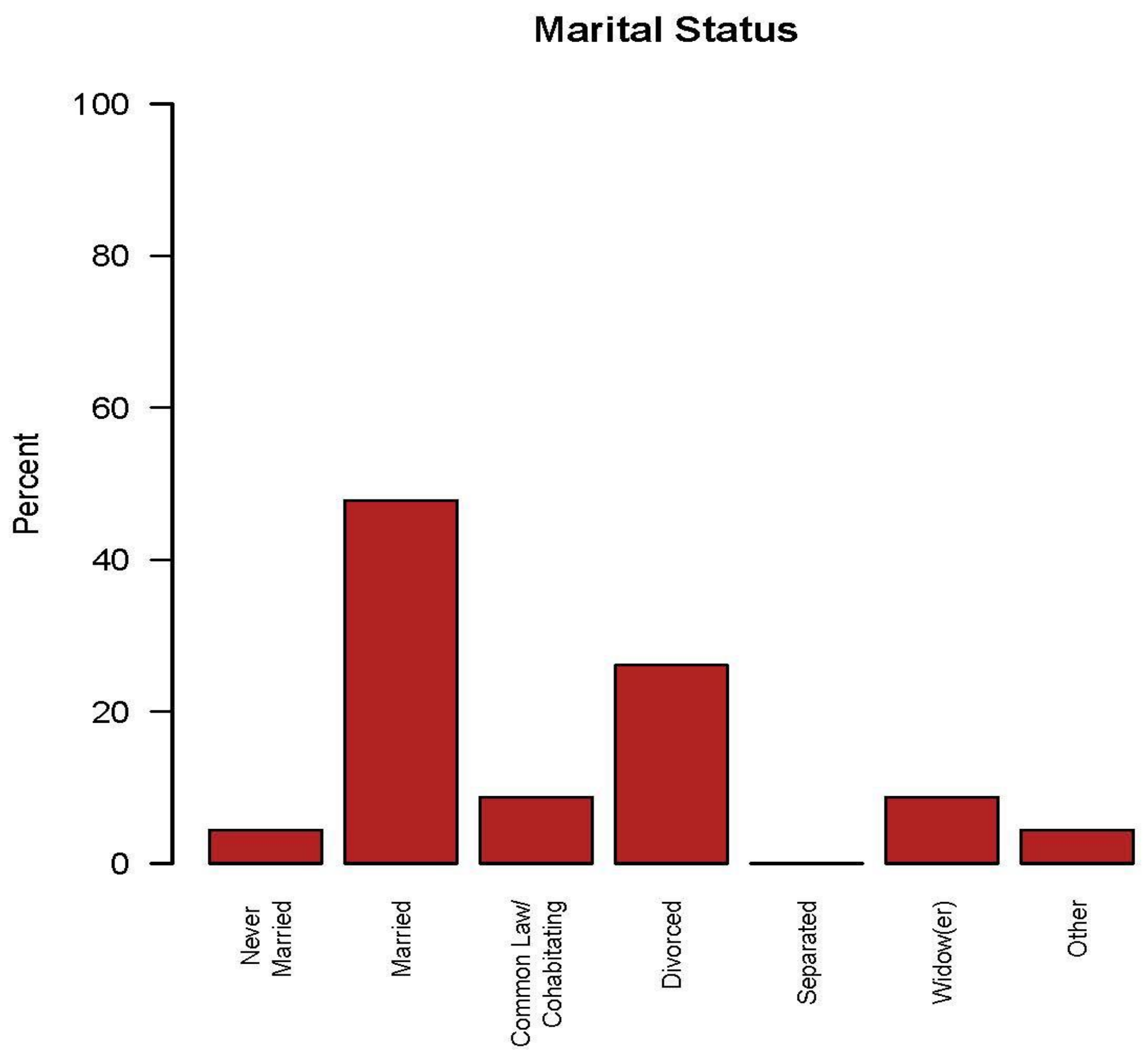
UNIVERSITY OF CALGARY

Example: Cardiac arrest

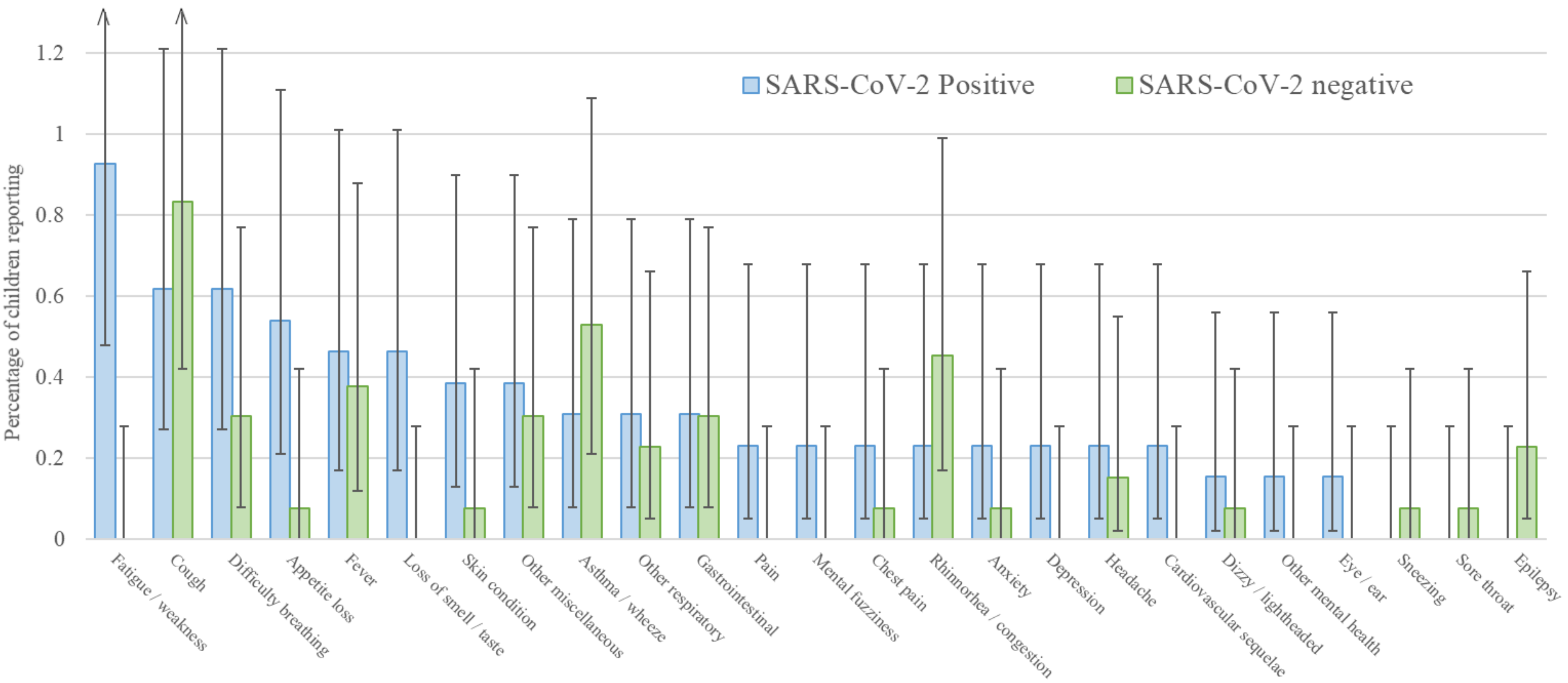
N or n Row % 95 % CI	Cardiac arrest	No cardiac arrest	Total
Male	276 36.95 33.5 – 40.4	471 63.05 59.4 – 66.5	747
Female	195 44.93 40.2 – 49.7	239 55.07 50.3 – 59.8	434
Male or Female (prev Total)	471 39.88 37.1 – 42.7	710 60.12 57.3 – 62.9	1181

95% Confidence Intervals

Displaying categorical data: Bar charts



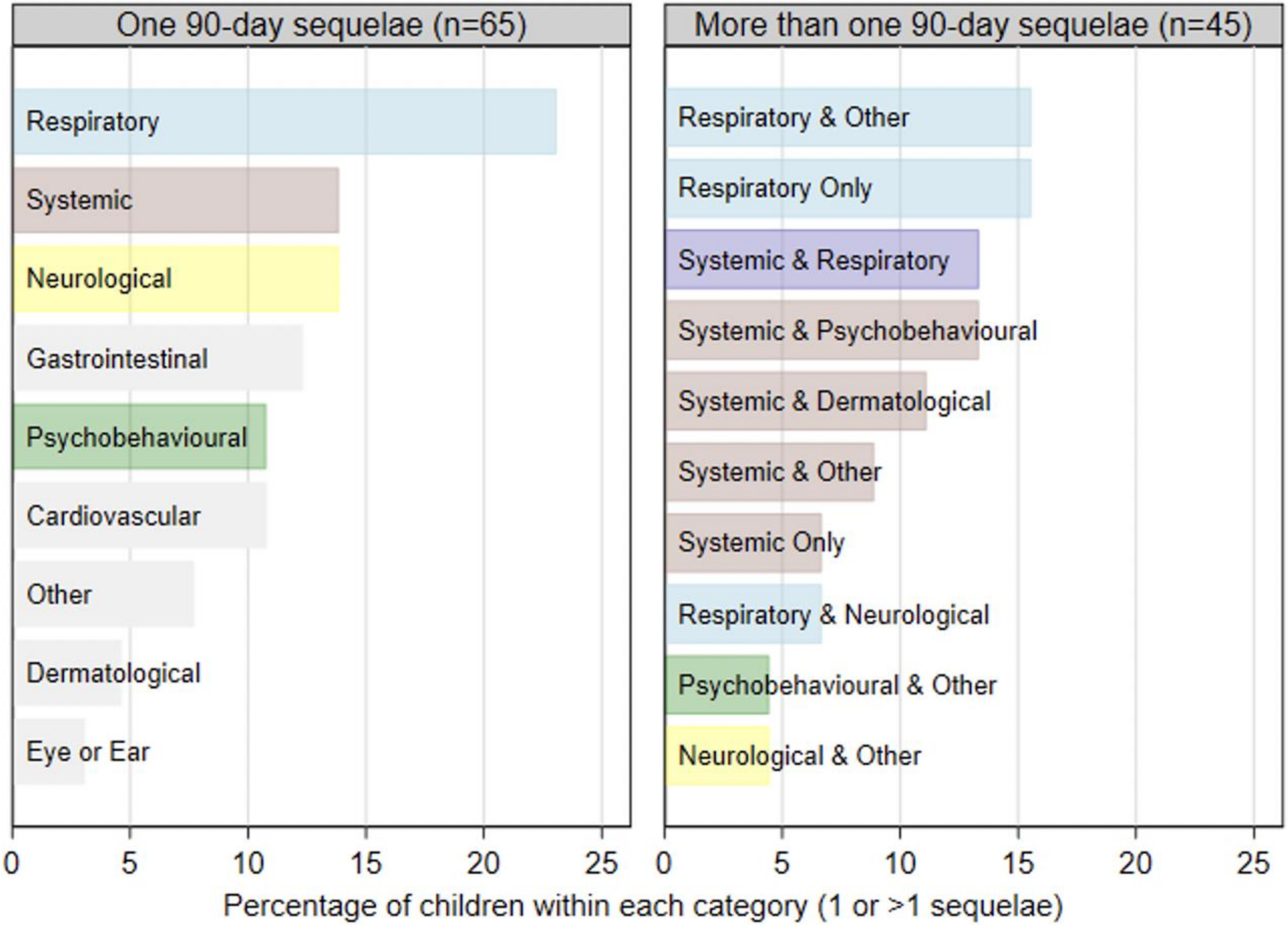
Displaying categorical data: Bar charts



Persistent symptoms in SARS-CoV-2 positive (N=1295) and SARS-CoV-2 negative (N=1321) children **not** hospitalized during the acute phase of SARS-CoV-2

Funk AL, Kuppermann N, Florin TA, Tancredi DJ, Xie J, Kim K, Finkelstein Y, Neuman MI, Salvadori MI, Yock-Corrales A, Breslin KA. Post–COVID-19 conditions among children 90 days after SARS-CoV-2 infection. JAMA network open. 2022 Jul 1;5(7):e2223253-.

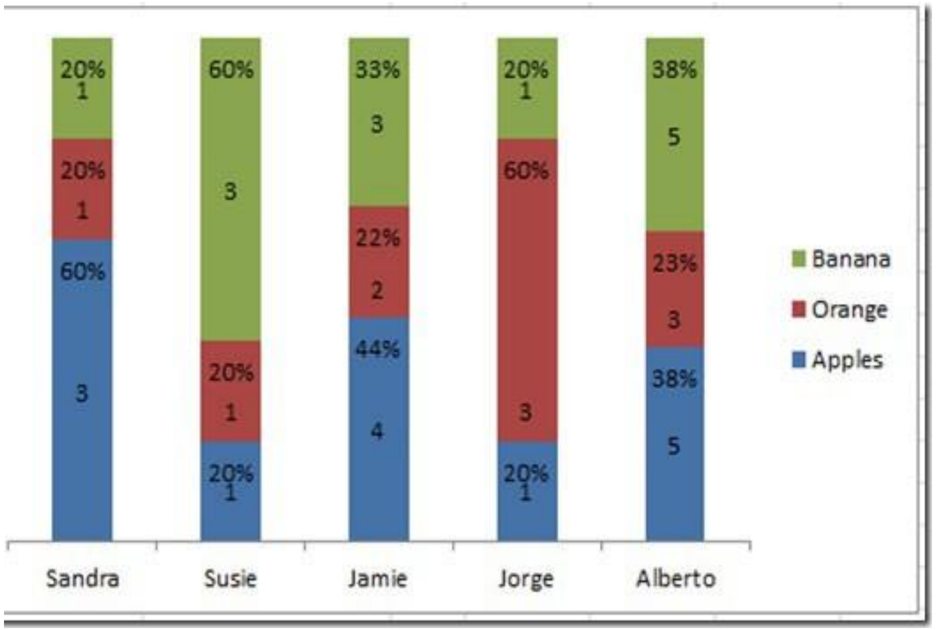
Displaying categorical data: Bar charts

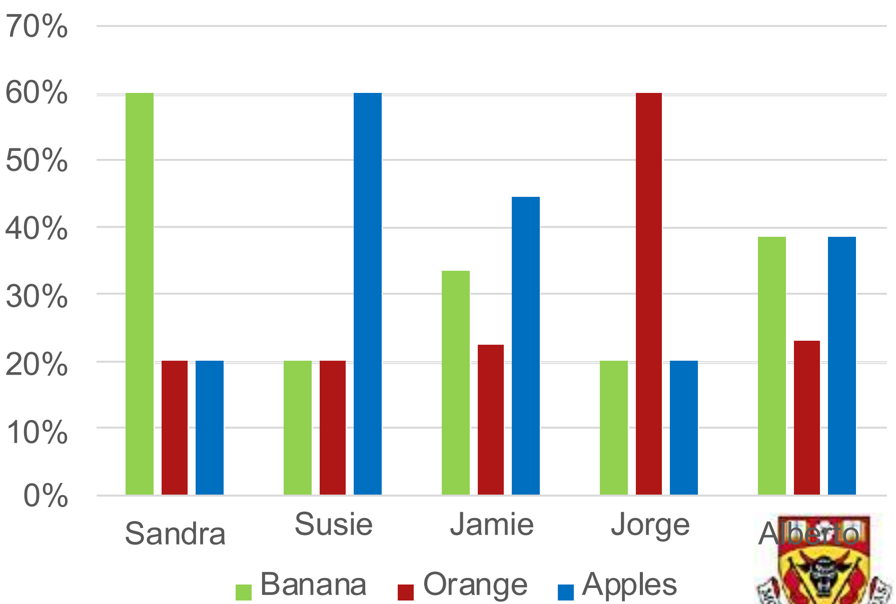


Continued… persistent symptoms in SARS-CoV-2 positive children

What not to do!

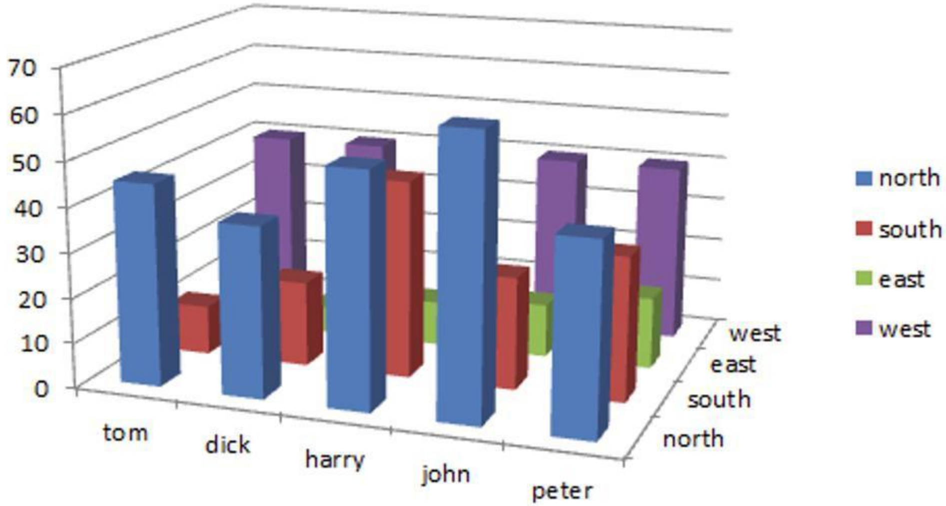
- **Don't overuse stacked bar graphs**
  - Use a graph to make comparison easy → not repeat all numbers (that's what tables are for!)
  - In stacked bars you can not easily tell different heights of different bars

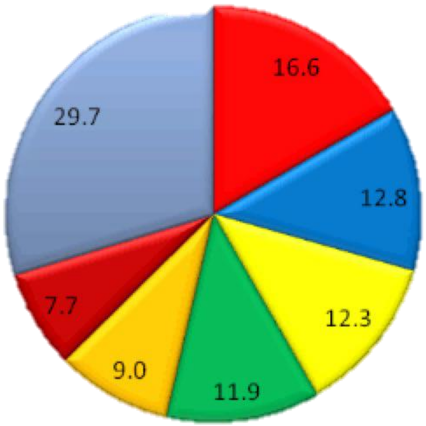




What not to do!

- Don't add "dimensions" that are not real dimensions and which distort the graph → keep it simple & clear
- Use pie charts: slices are not easily comparable and numbers are often added… Hence it is just a colored disarrayed table!





http://www.essaybuilder.net/PieCharts.html

Numerical Data

Discrete

- Can only take on certain numerical values

- E.g. Counts of events
    - Number of children hospitalized with COVID-19
    - Number of concussions in an adolescent football player
    - Number of children per household in Southern Alberta

- Discrete variables that would not be a ‘count’?

Numerical Data

Continuous

- **Interval data**
  - Ordered, difference between any two values is meaningful
  - Can theoretically be ***any value*** within a range
  - No true ‘zero’ measurement that indicates the lack of characteristic
  - **Examples:**
    - Date (with/without time) = 19/10/2020 09:04:56.864
    - Temperature in Celsius/Fahrenheit = -14.8 Celsius
    - pH level = 7.4

Numerical Data

Continuous

- **Ratio data**
  - Interval data **BUT** a ‘true zero’ does exist
    - Because of this, a *ratio* of two data points is meaningful
      - 10 kg is twice as heavy as 5 kg (ratio)
      - 10 ^{o}C is **not** twice as hot as 5 ^{o}C (interval)
  - **Examples:**
    - Viral load in copies/ml
    - Height in cm
    - Temperature in Kelvin (with its absolute zero that represents no temperature)
    - Time to an event (e.g. survival after transplant)

Numerical → Categorical

Interval and ratio data can be converted into categorical data

- Examples:
  - Precise age → < 1 year, 1-<3 years, 3-<5 years
  - Precise height and weight → specific BMI value → BMI category
- Lose some precision / detail, *however, is a 1 unit increase in your numerical data meaningful?*

Describing numerical data - overview

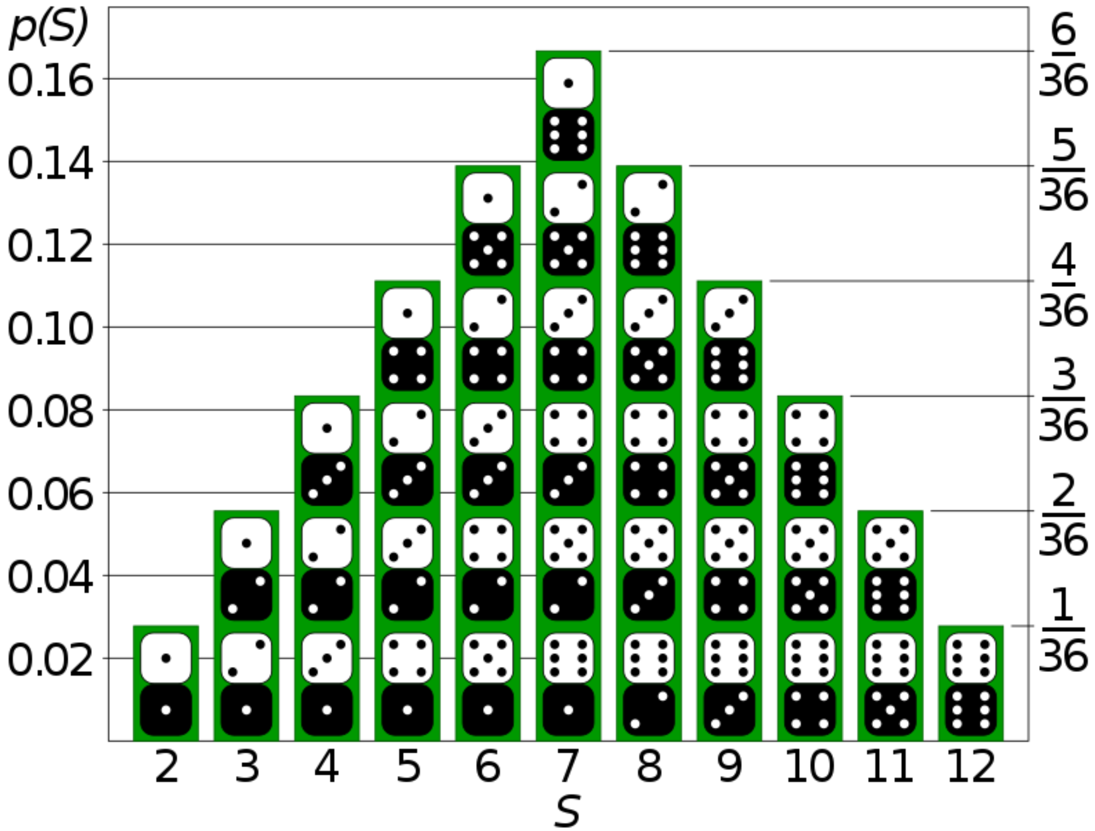
- **Data distributions and their features**
  - Probability distributions
    - Measures of central tendency - mean, median, mode
    - Measures of dispersion
    - Skewness
- **Graphical representation of numerical data**

Probability Distributions

- In ‘**probability distributions**’, each possible event gets/has a certain amount of probability (chance) of happening.

Probability mass function →

Retrieved at:
https://commons.wikimedia.org/wiki/File:Dice_Distribution_(bar).svg

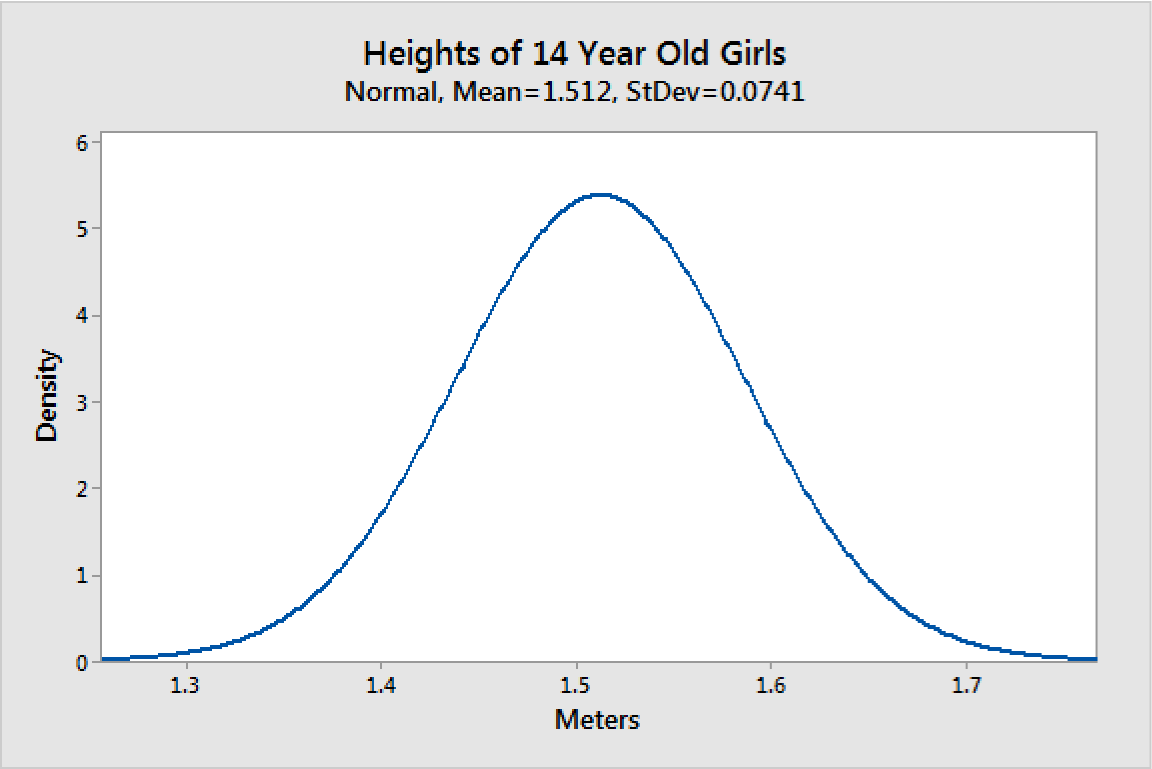


Probability Distributions

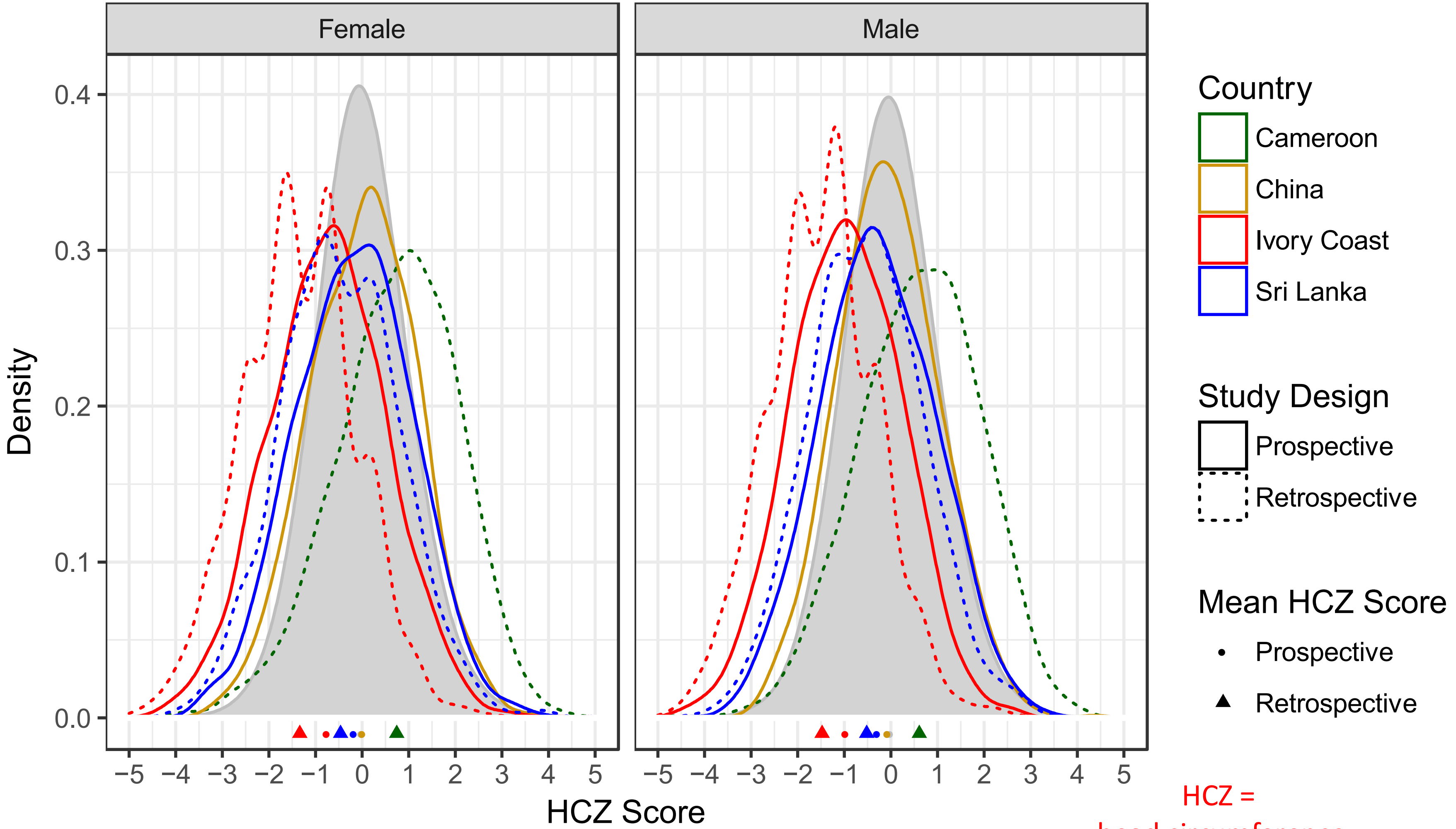
- In ‘**probability distributions**’, each possible event gets/has a certain amount of probability (chance) of happening.

Normal (Gaussian) distribution →

Retrieved at:
https://statisticsbyjim.com/basics/normal-distribution/



Probability Distributions



HCZ = head circumference Z-score

Zika-virus related microcephaly surveillance across 4 countries

Features of distributions - 1

- **Central tendency or location parameters**
  - The value that observations are "centered" around
  - **Mean** – average value
    - Sum of values divided by the number in the sample
  - **Median** – value at which 50% of the observations have accumulated
    - *To calculate:* rank data, and find the value for which 50% of observations fall above and 50% below, *or* use rank formula
  - **Mode** – most frequently *observed* value

Example

Weight (in kg) of 10 five year old boys

- **Data:** 14, 14, 15, 17, 17, 18, 18, 18, 19, 20

Example

Weight (in kg) of 10 five year old boys

- **Data:** 14, 14, 15, 17, 17, 18, 18, 18, 19, 20

- **Mean:** (14+14+15+17+17+18+18+18+19+20) /10 = 17
- **Median** (via rank formula):
  - Index **i** = (pth percentile*# observations)/100
  - 50th percentile*10 observations/100 = 5
    - If **i** is not an integer (e.g. 3.5, 46.5) the next value is pth percentile
    - If **i** is an integer (e.g. 4, 40), average **i**th and ith +1
  - Our **i** is 5 (average 5th and 6th values), median = 17+18 / 2 = 17.5
- **Mode:** most frequent value
  - ‘18’ appears the most (3 times), mode = 18

Example continued…

Weight (in kg) of 10 five year old boys

- **Data:** 14, 14, 15, 17, 17, 18, 18, 18, 32, 40
- **Mean:** (14+14+15+17+17+18+18+18+32+40) /10 = $\underbrace{20.3}_{\text{Changed!}}$
- **Median** (via rank formula):
  - median = 17+18 / 2 = 17.5
- **Mode:** most frequent value
  - '18' appears the most (3 times), mode = 18

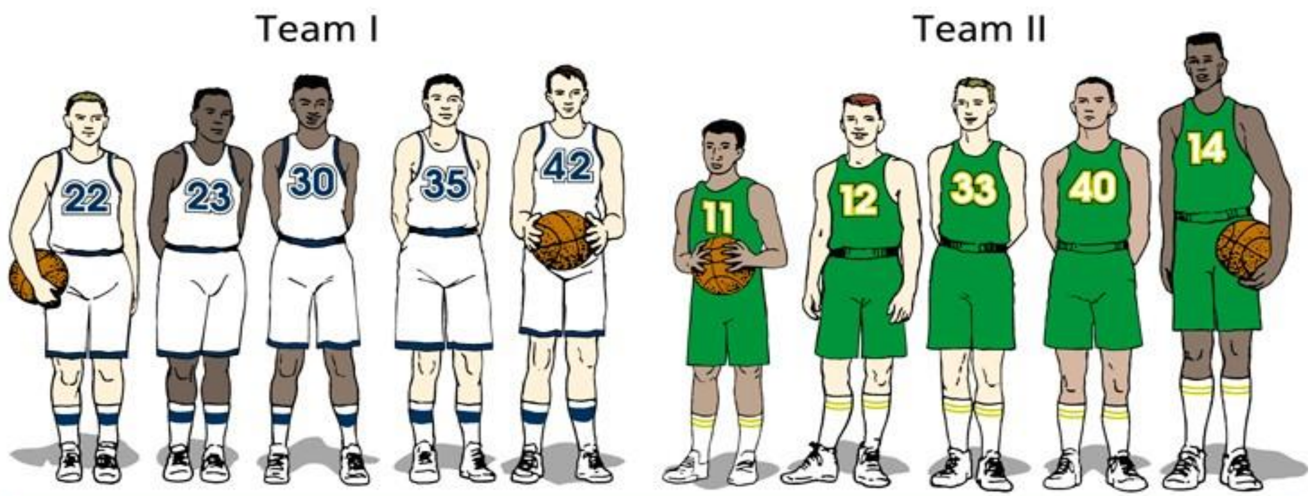
Stayed the same!

Using Measures of Central Tendency

- **Mean**
  - When data **aren't skewed**
    - If symmetric distribution, mean ~ median
  - Lack of outliers / extreme values
    - Highly influenced by extreme values
- **Median**
  - When data are skewed or when there are outliers
    - Robust to extreme values
- **Mode**
  - Use with nominal data (e.g. most common ethnicity)
  - Use to check for bimodal distribution

Features of distributions - 2

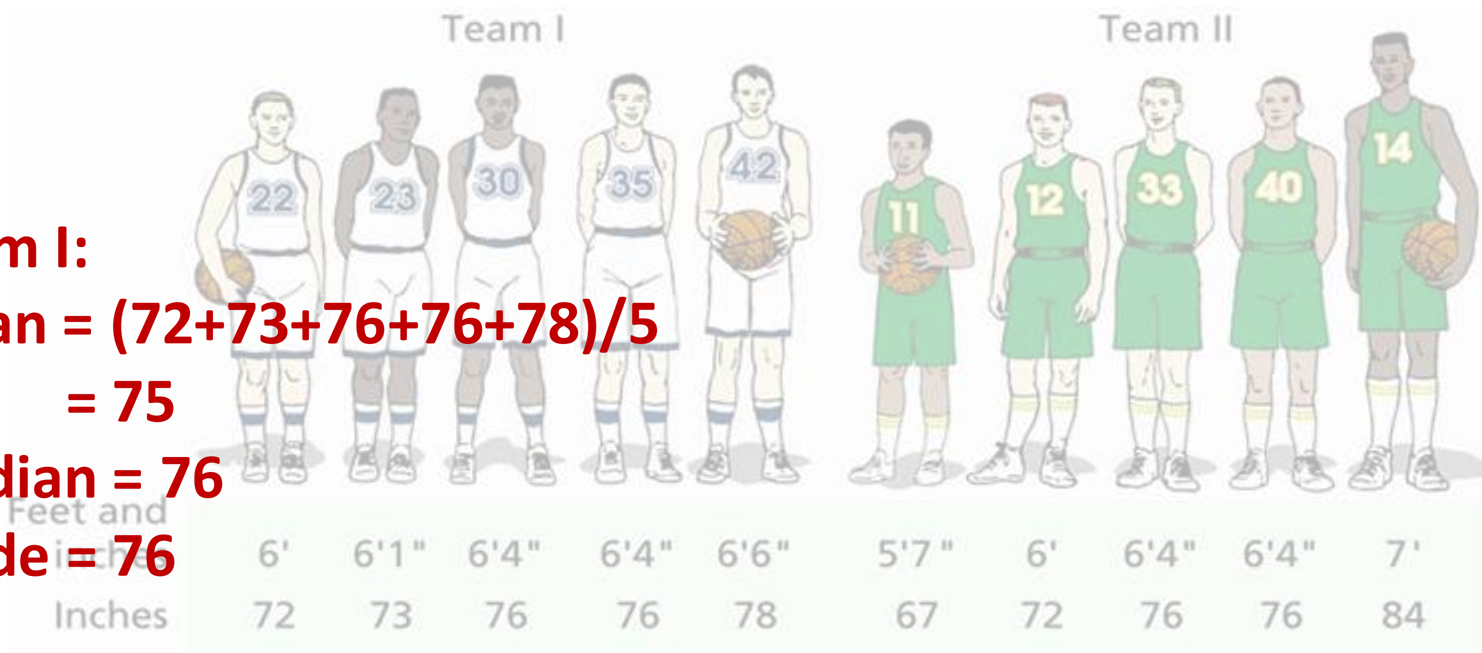
- **Dispersion**
  - How **spread out** the observations are in the distribution (sometimes around the "centre").



	Team I					Team II				
Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

Features of distributions - 2

- **Dispersion**
  - How **spread out** the observations are in the distribution (sometimes around the “centre”).



Team I:

mean = (72+73+76+76+78)/5

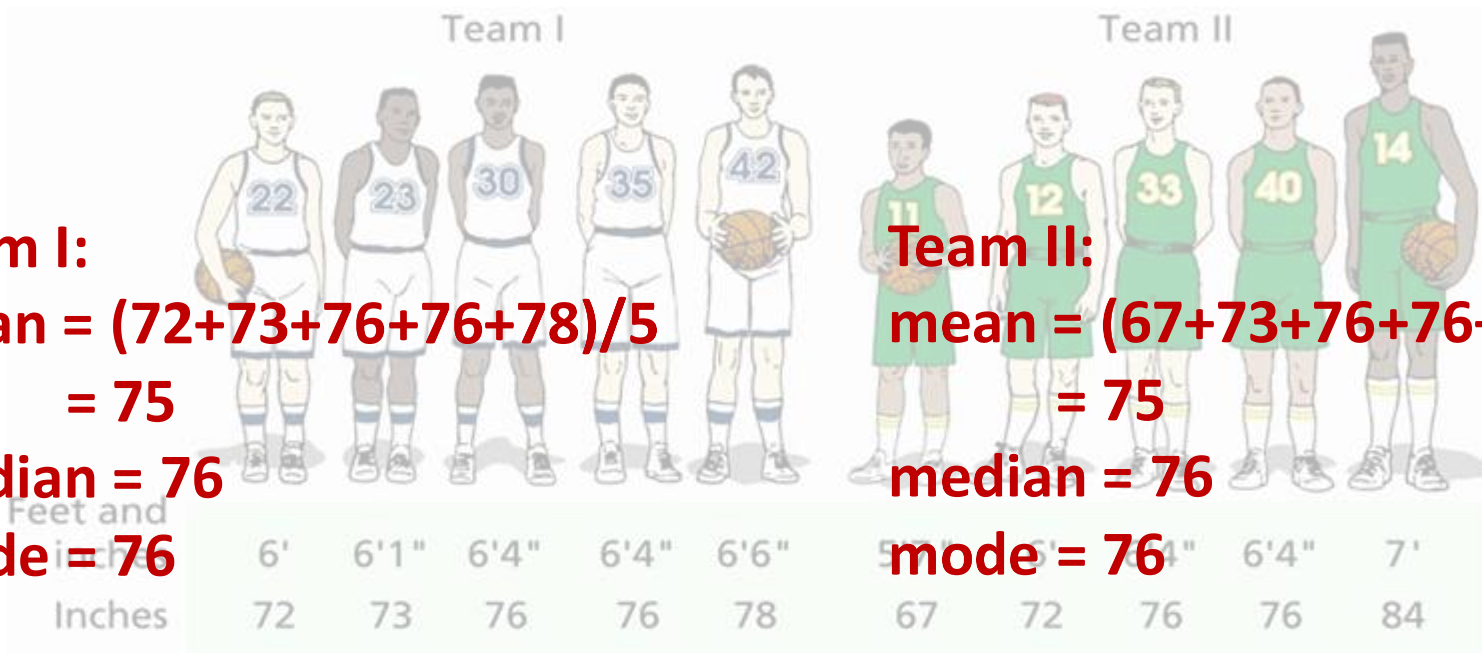
= 75

median = 76

mode = 76

Features of distributions - 2

- **Dispersion**
  - How **spread out** the observations are in the distribution (sometimes around the "centre").

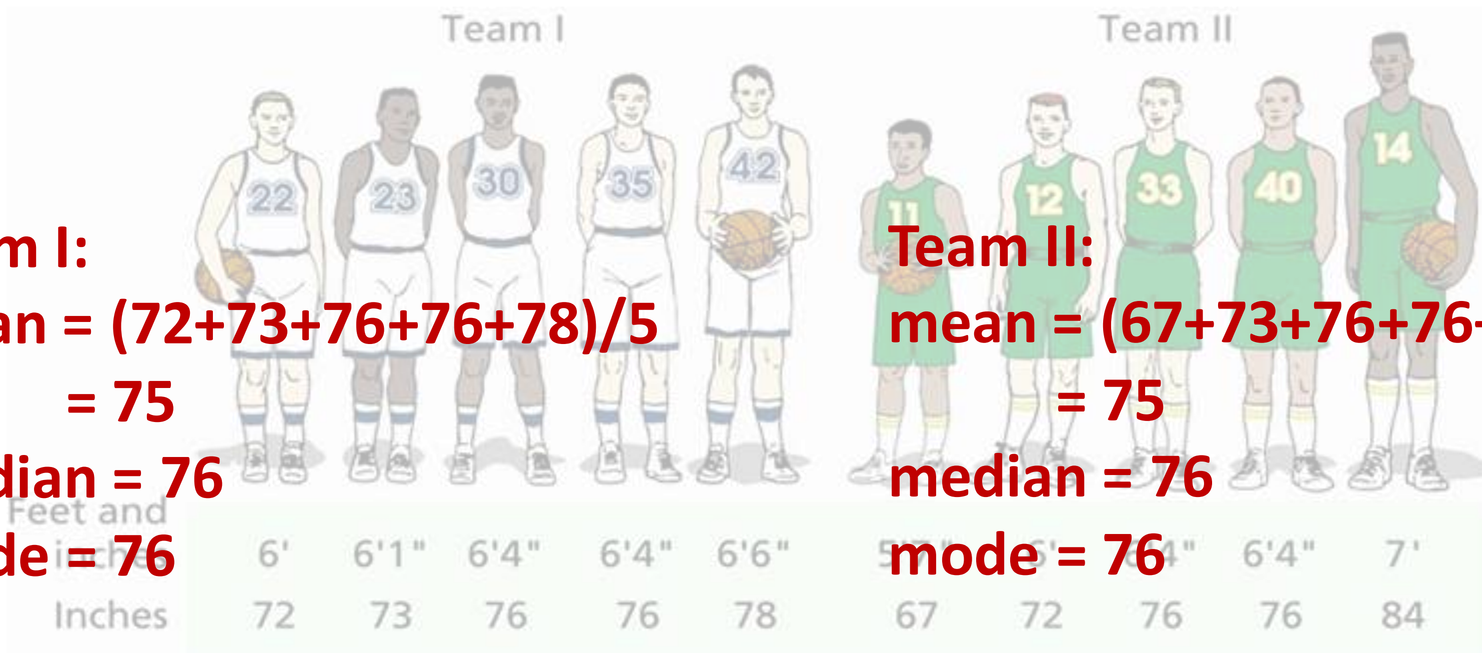


Team I:
mean = (72+73+76+76+78)/5
= 75
median = 76
mode = 76

Team II:
mean = (67+73+76+76+84)/5
= 75
median = 76
mode = 76

Features of distributions - 2

- **Dispersion**
  - How **spread out** the observations are in the distribution (sometimes around the “centre”).



Team I:
mean = (72+73+76+76+78)/5
= 75
median = 76
mode = 76

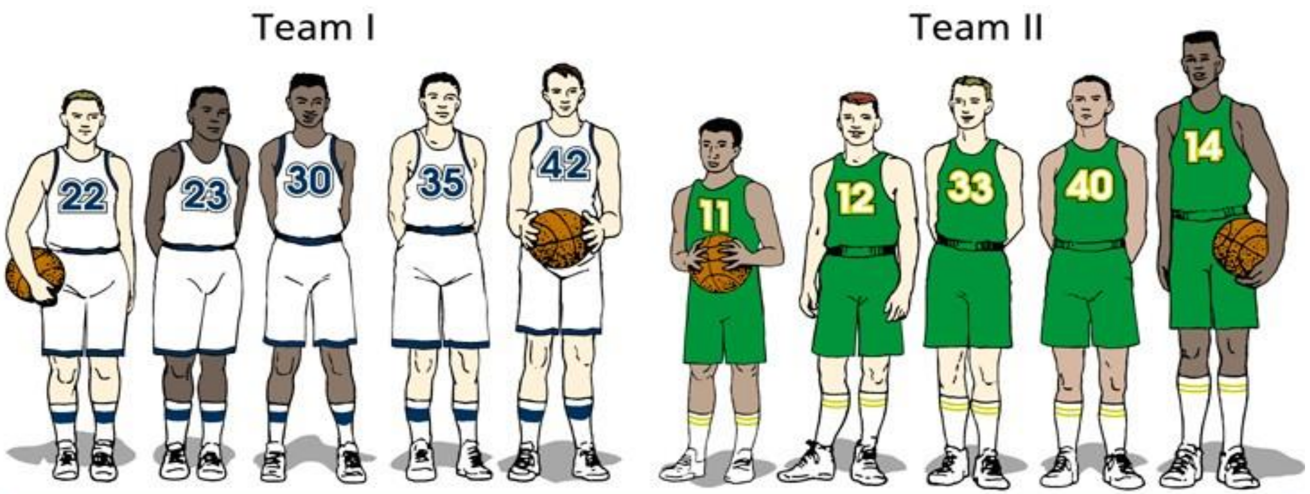
Team II:
mean = (67+73+76+76+84)/5
= 75
median = 76
mode = 76

The “data sets” have the same Mean, Median, and Mode yet clearly differ!

Measures of dispersion

- **Range**
  - Difference between largest and smallest values
- **Interquartile range → IQR**
  - Difference between the third quartile (75th percentile) and first quartile (25th percentile)
  - Central 50% of the data
- **Standard deviation → S or SD**
  - Square root of the sample variance (S^2)
    - (Sum of squared deviations from mean) / n-1
  - Most commonly used, measures spread around the mean … & highly affected by mean

Features of distributions - 2



Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

Range: 72-78
IQR: 73 – 76

Range: 67-84
IQR: 72 – 76

Features of distributions - 2

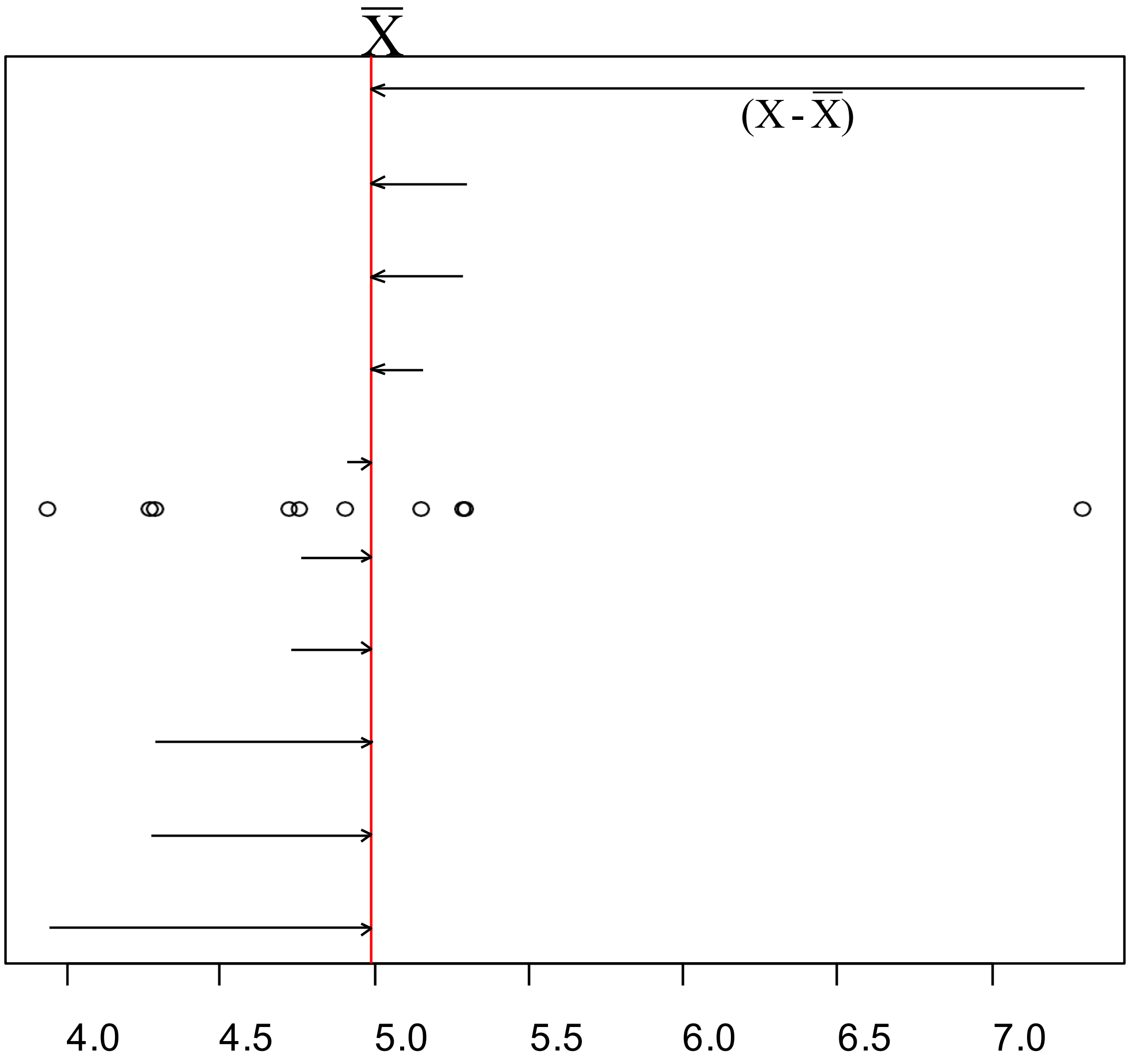
Largers vs smaller standard deviations

Retrieved at: https://www.scribbr.com/statistics/standard-deviation/

Calculating the variance and standard deviation
(X = total cholesterol)

Subject (i)	X
1	3.94
2	4.73
3	5.30
4	5.16
5	4.91
6	5.29
7	4.29
8	7.30
9	4.76
10	4.27

Calculating the variance and standard deviation
(X = total cholesterol)



Calculating the variance and standard deviation
(X = total cholesterol)

i	**X**	$\bar{X}$	$(X - \bar{X})$	$(X - \bar{X})^2$
1	3.94	4.99	-1.05	1.11
2	4.73	4.99	-0.26	0.07
3	5.30	4.99	0.31	0.09
4	5.16	4.99	0.17	0.03
5	4.91	4.99	-0.08	0.01
6	5.29	4.99	0.30	0.09
7	4.29	4.99	-0.70	0.50
8	7.30	4.99	2.31	5.31
9	4.76	4.99	-0.23	0.06
10	4.27	4.99	-0.72	0.53
$\sum$	49.95			7.79

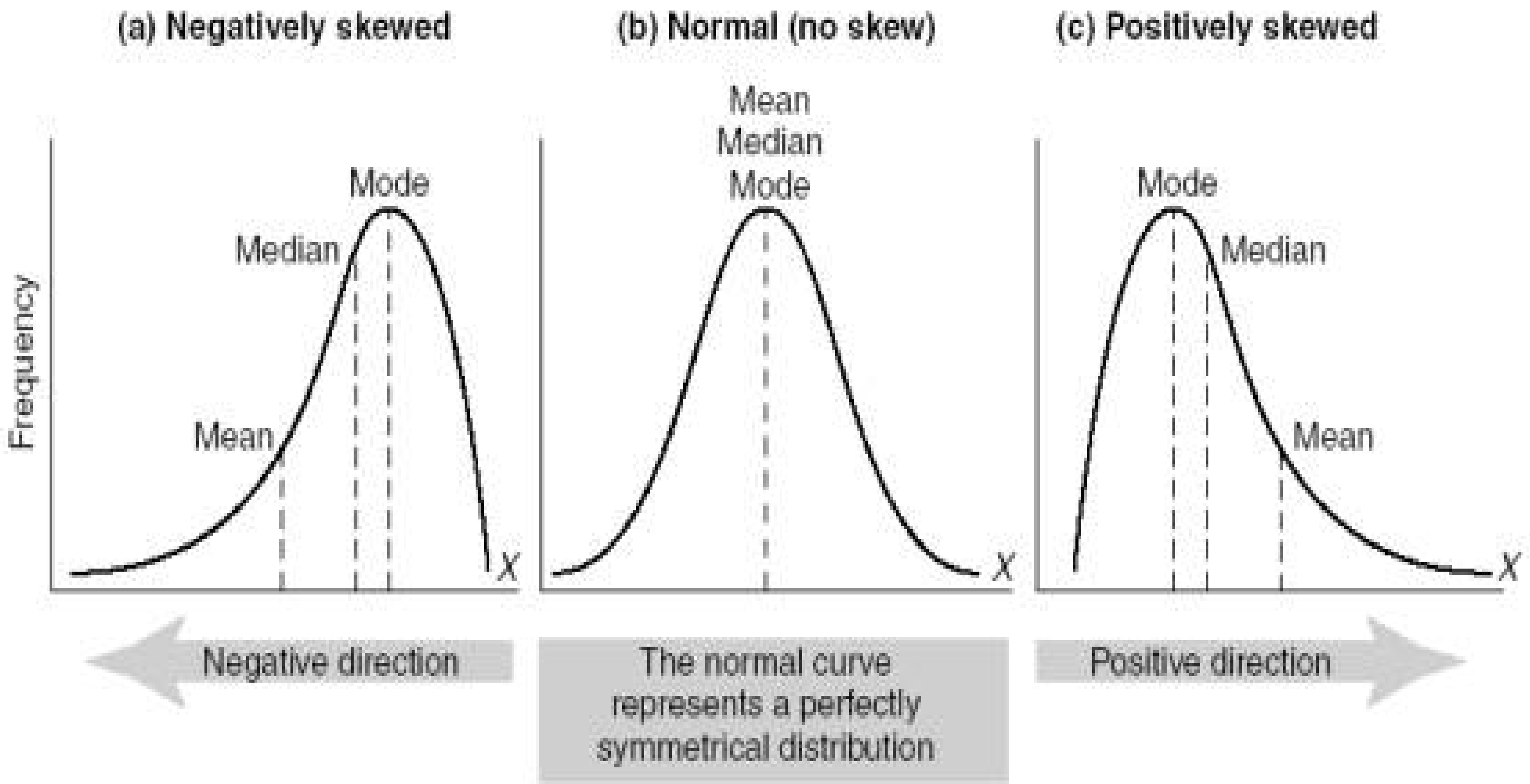
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 4.99$$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = 0.87$$

$$S = \sqrt{0.87} = 0.93$$

Features of distributions - 3

- **Skewness** *(measured by the skewness statistic)*
  - Are observations symmetrically distributed? Do they tend to extend farther to the left or to the right ?



Displaying numerical data

- **Histograms**
  - Bars representing frequency (or % of data observed) for specific values or groups of values

- **Boxplots**
  - Three number summary with outlier identification
    - 25% (lower quartile), 50% (median), 75% (upper quartile)
  - Good for visually comparing two or more groups

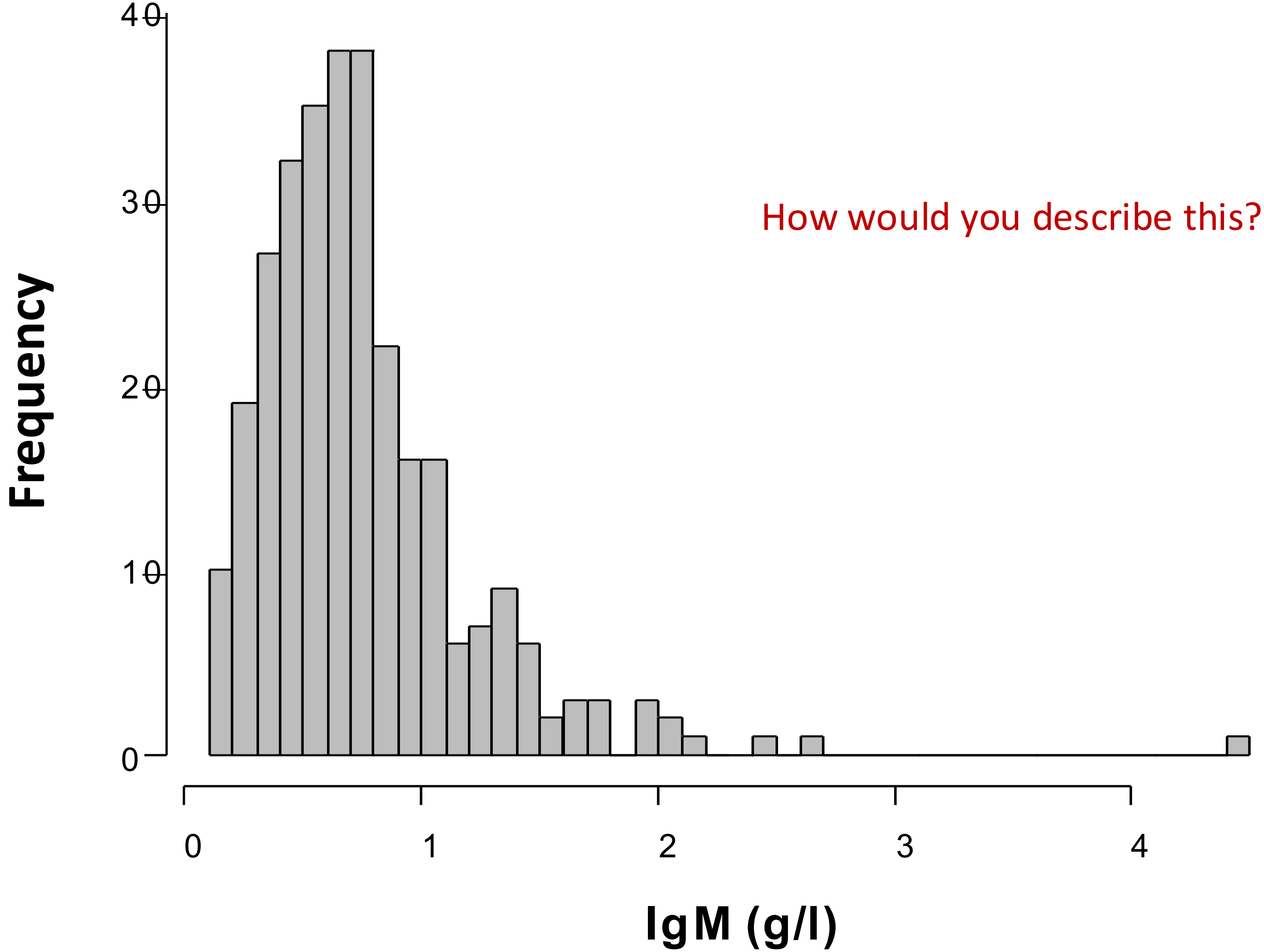
- **Other – scatter plots** (for correlation, not today)

IgM data for 298 healthy children (age 2-6 yrs)

Concentration (g/l)	Frequency	Concentration (g/l)	Frequency
0.1 to < 0.2	3	1.6 to < 1.7	2
0.2 to < 0.3	7	1.7 to < 1.8	3
0.3 to <0.4	19	1.8 to < 1.9	3
0.4 to < 0.5	27	1.9 to < 2	0
0.5 to < 0.6	32	2 to < 2.1	3
0.6 to < 0.7	35	2.1 to < 2.2	2
0.7 to <0.8	38	2.2 to < 2.3	1
0.8 to < 0.9	38	2.3 to < 2.4	0
0.9 to < 1	22	2.4 to < 2.5	0
1 to < 1.1	16	2.5 to < 2.6	1
1.1 to < 1.2	16	2.6 to < 2.7	0
1.2 to < 1.3	6	2.7 to <2.8	1
1.3 to < 1.4	7	...	
1.4 to < 1.5	9	...	
1.5 to < 1.6	6	4.5 to < 4.6	1

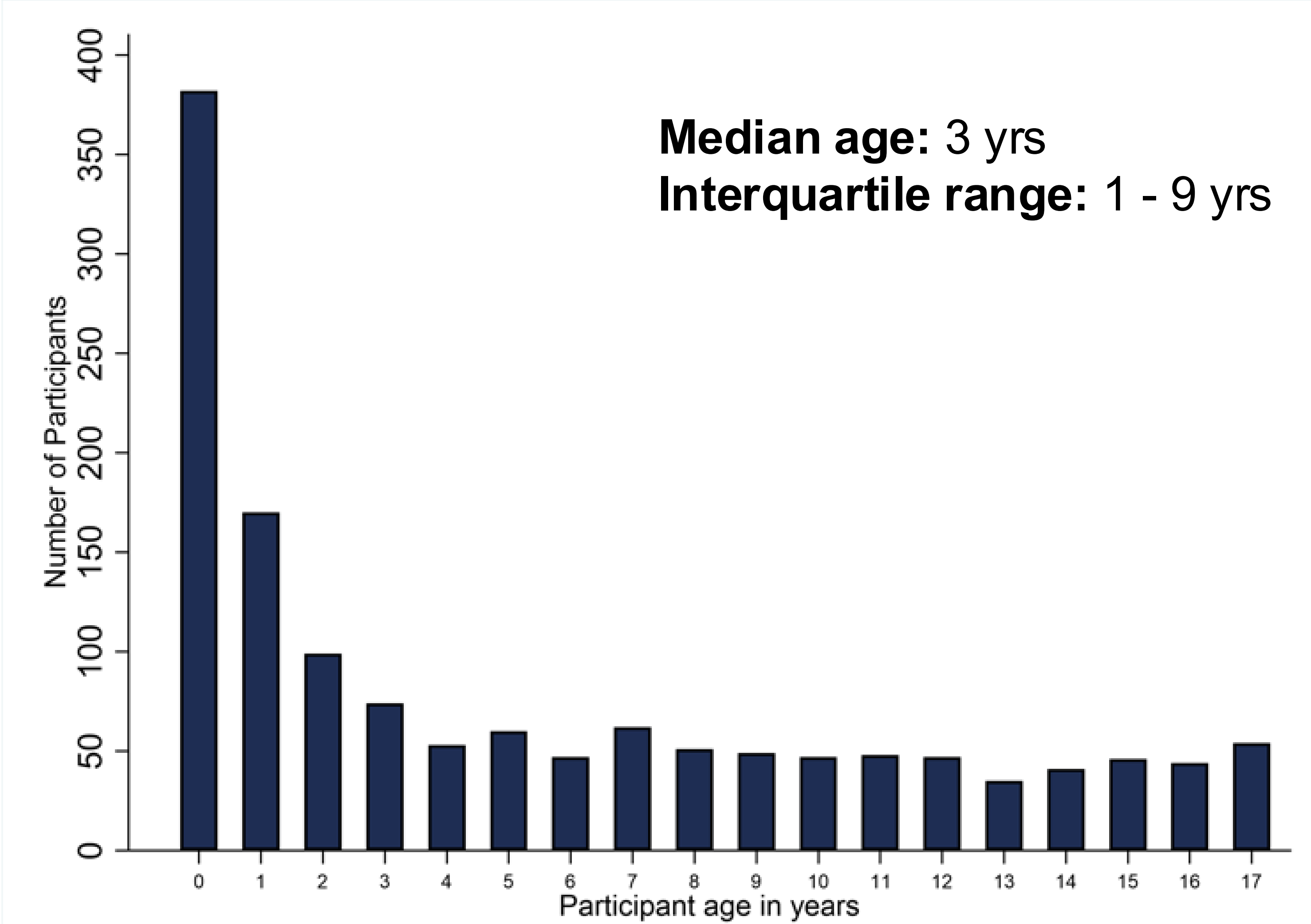
Data from Altman, 1995, Practical Statistics for Medical Research, Table 3.2, p. 23

Histogram of IgM (g/l) in 298 healthy children

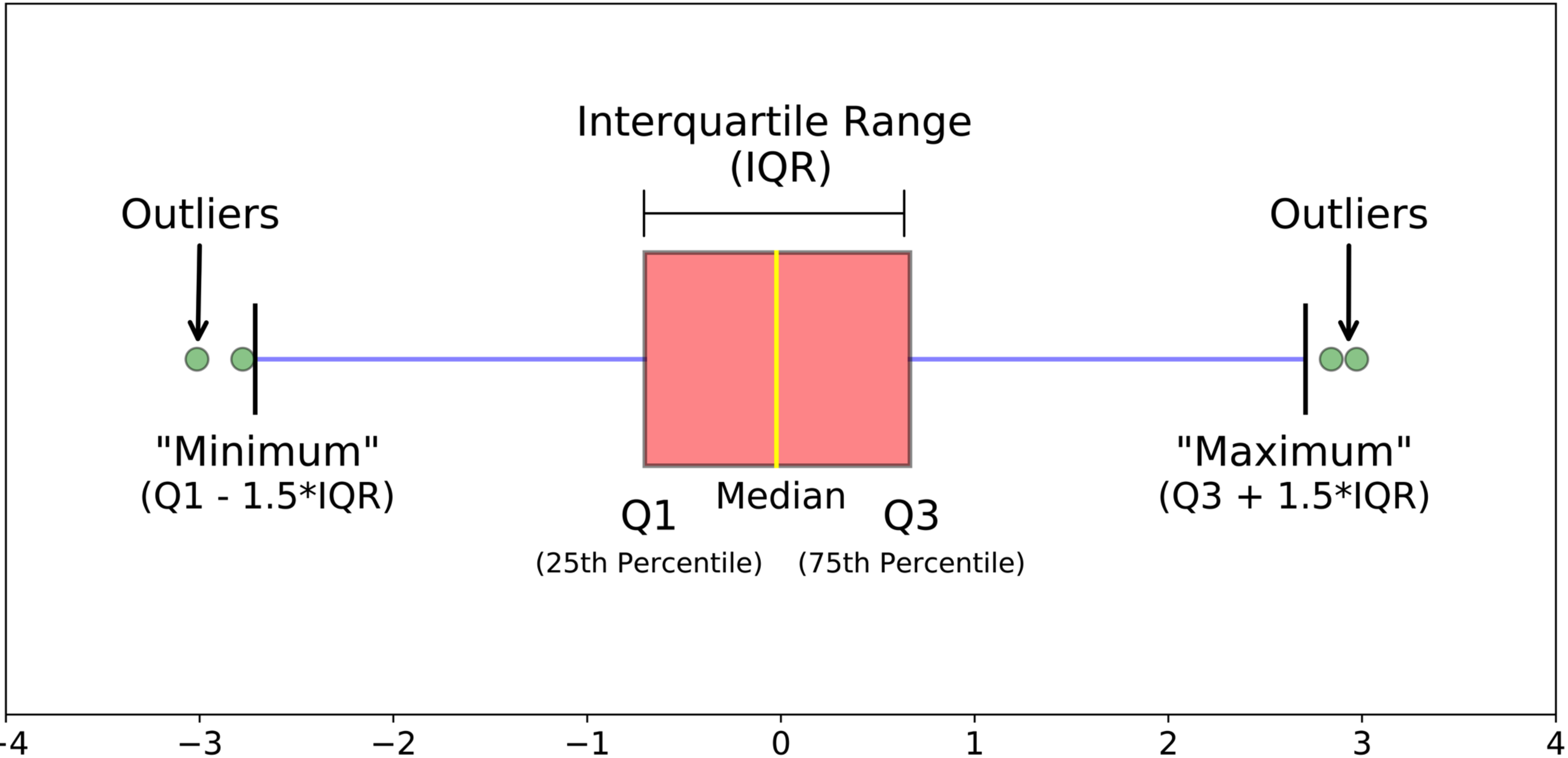


How would you describe this?

Ages of 1398 children with COVID-19 enrolled in the PERN-COVID-19 cohort study



The Box and Whisker Plot: Anatomy



Note: Points above and below the "whiskers" are considered to be outliers. These outliers are not representative of the general distribution of values

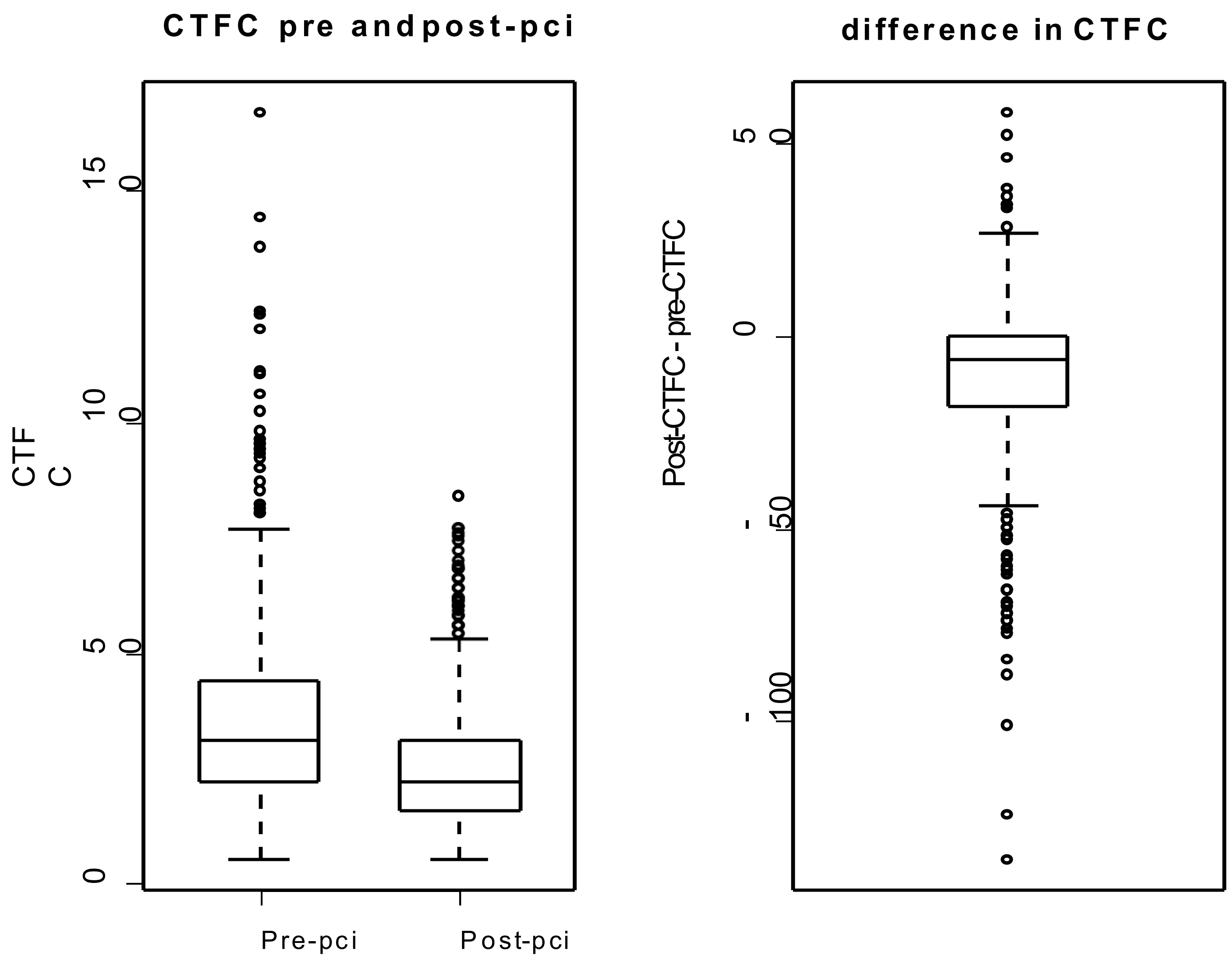
Graphic taken from: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Boxplot Example: CTFC

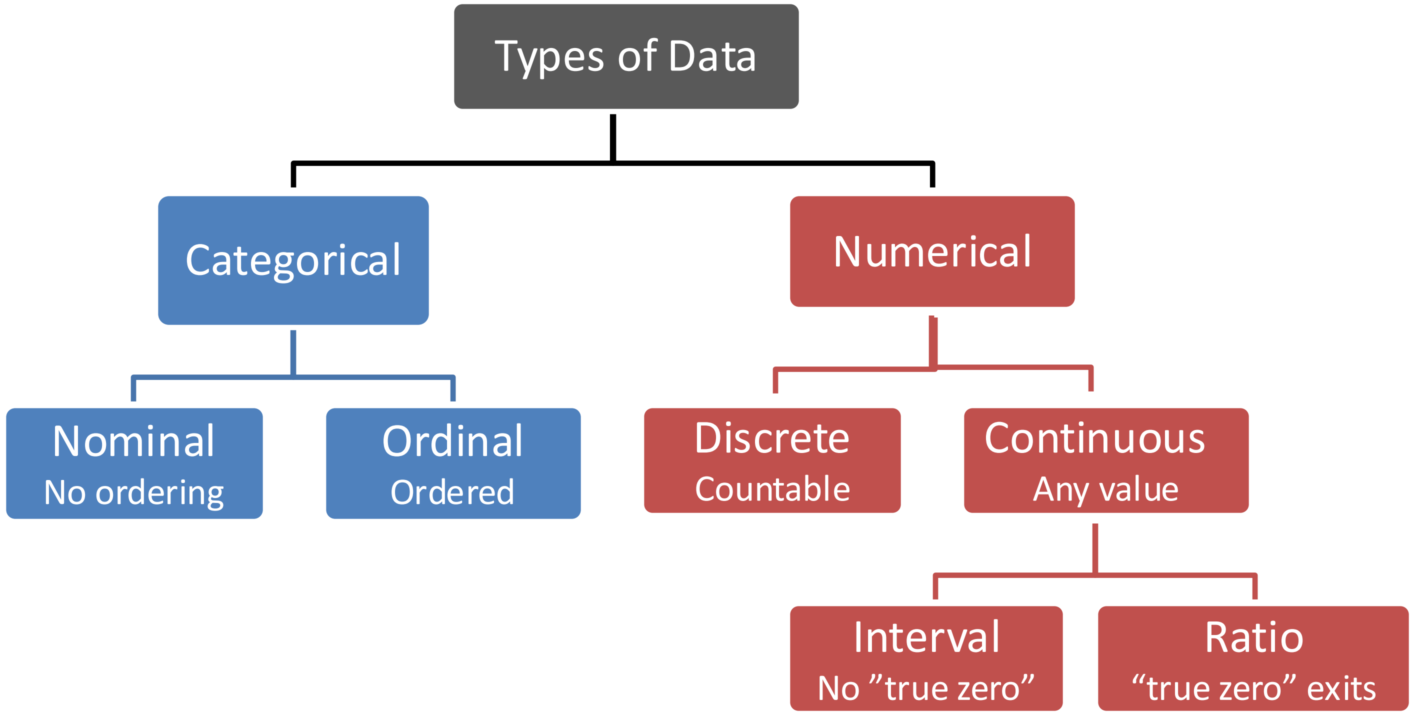
- In coronary angiography, the corrected TIMI frame count (CTFC) is the number of frames required for dye to reach a standardized distal landmark.

- The following boxplots show the CTFC before and after angioplasty in approximately 1000 patients.
  - *Note the skewness in the distributions.*

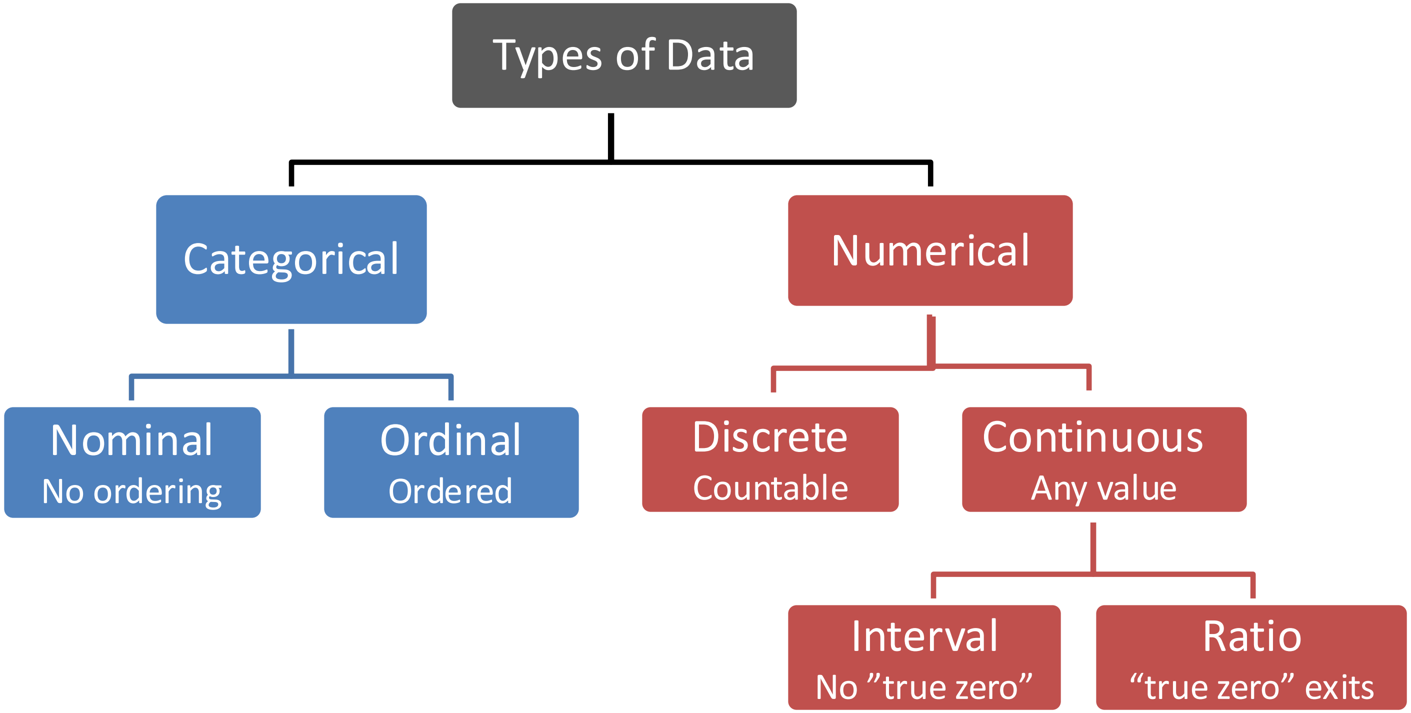
Boxplot Example: CTFC



Types of Data - takeaway

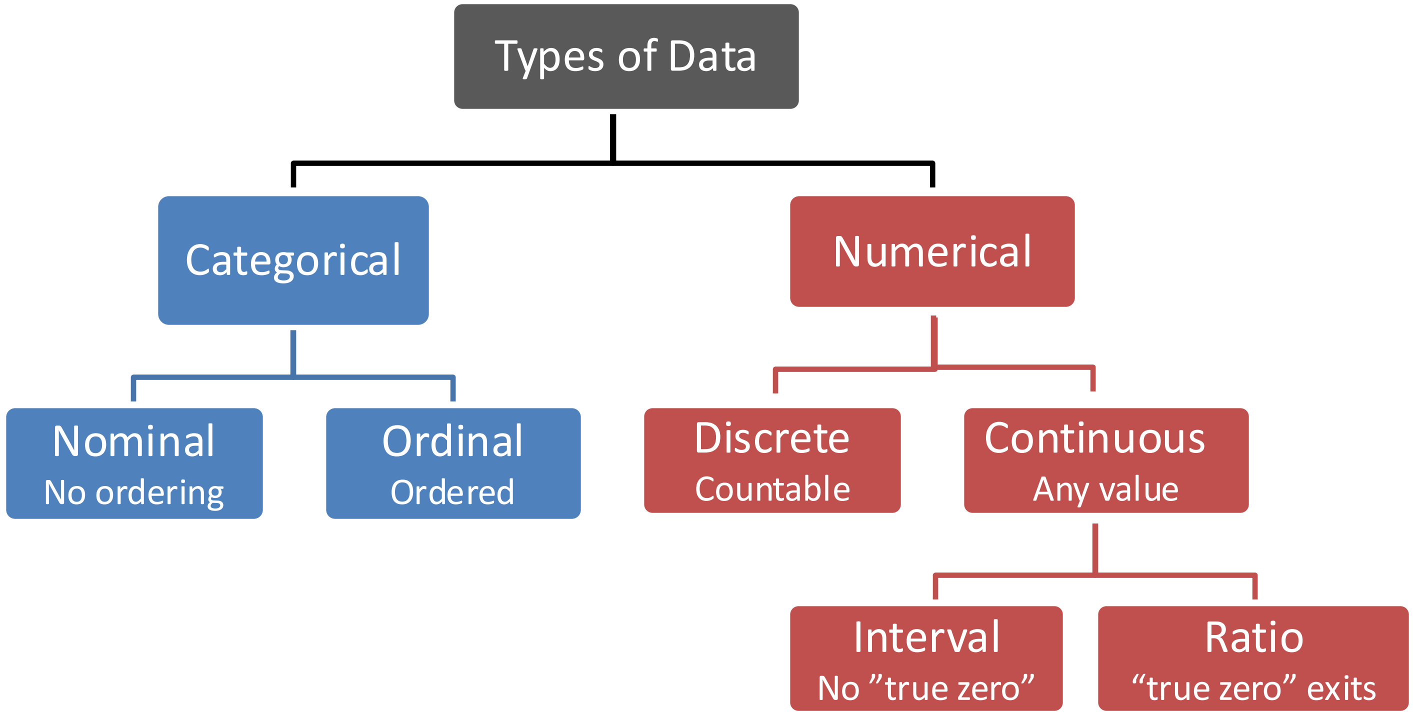


Types of Data - takeaway



- Difference between 2 categories not equal
- Frequency counts with % (proportions)
- Mode
- Bar charts

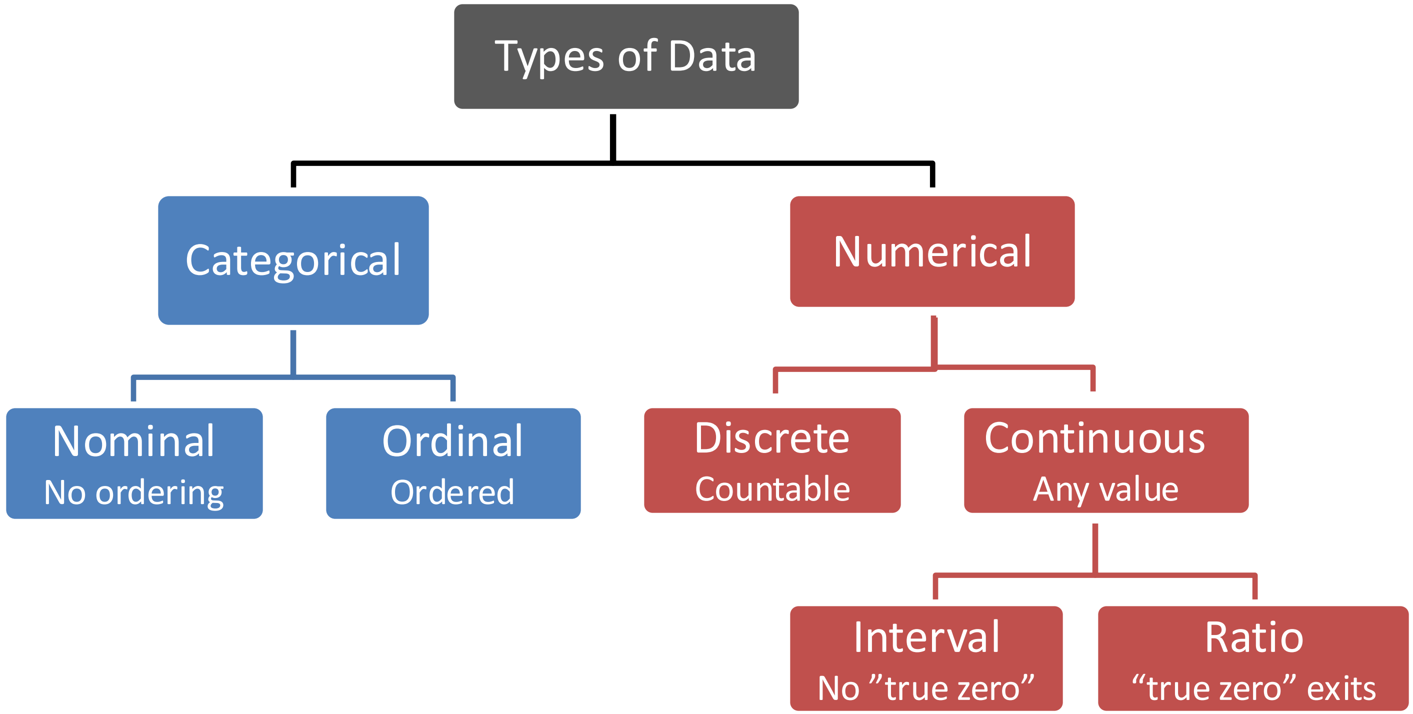
Types of Data - takeaway



- Difference between 2 categories not equal
- Frequency counts with % (proportions)
- Mode
- Bar charts

- Difference between values is meaningful
- Mean, median
- SD, Range/IQR
- Skewness
- Histograms, box plots

Types of Data - takeaway



- Difference between 2 categories not equal
- Frequency counts with % (proportions)
- Mode
- Bar charts

- Difference between values is meaningful
- Mean, median
- SD, Range/IQR
- Skewness
- Histograms, box plots

Graphics should <u>help</u> readers understand the data

Lecture objectives

- **Statistics, study design, and critical thinking about bias are important**

- Types of sampling methods

- Identify what types of data you are (or want to be) working with and describe it appropriately
  - Graphical presentation

- Statistical analysis plan **--> have one before you start**

Your statistical analysis methods

- Definitions
  - E.g. What is a ‘severe outcome’ for a child with COVID-19
- State descriptive statistics used
  - Keep in mind – is your sample representative?
- State statistical test or methods used for comparison of groups (if any) in order of when done
- Cite references for complex or uncommon statistical tests used to analyze the data
  - Check similar papers, consult with a designated study (or department) statistician and epidemiologist

Your statistical analysis methods

- Supplementary appendix for more detail, complex methods, *'personalization'* of models

- Describe how you checked the assumptions of your statistical methods
  - E.g. is your quantitative data normally distributed?
  - E.g. do participants differ by an unadjusted variable (e.g. site)?

- Provide a sample size calculation
  - **Not** done at time of report writing, should have been done (or declared 'convenience') at time of proposal

Types of Measures

- The way (scale of measurement) in which your data is collected, will influence statistical analysis
  - **E.g.** blood pressure – high, normal, low *vs* exact value in mm Hg
  - **E.g.** viral presence (yes/no) *vs* viral load in CT (cycle threshold) values
  - **Challenges:** e.g. quantification of tissue damage? socioeconomic status?
  - **Consider which type of measure is appropriate for answering your research question**
    - Proportion of treatment failures? Number of diarrheal episodes per patient? Change in quality of life pre and post intervention.

Your statistical analysis methods

In reality, prior to undertaking statistical analysis (as designated in your plan)

→ Careful graphical and exploratory analysis

- Histograms, scatter plots, univariate analysis, etc
- Aids in data cleaning and the decision about the appropriateness of planned methods
  - It is not preferred but OK to change your statistical analysis plan! All changes need to be stated and justified in detail.