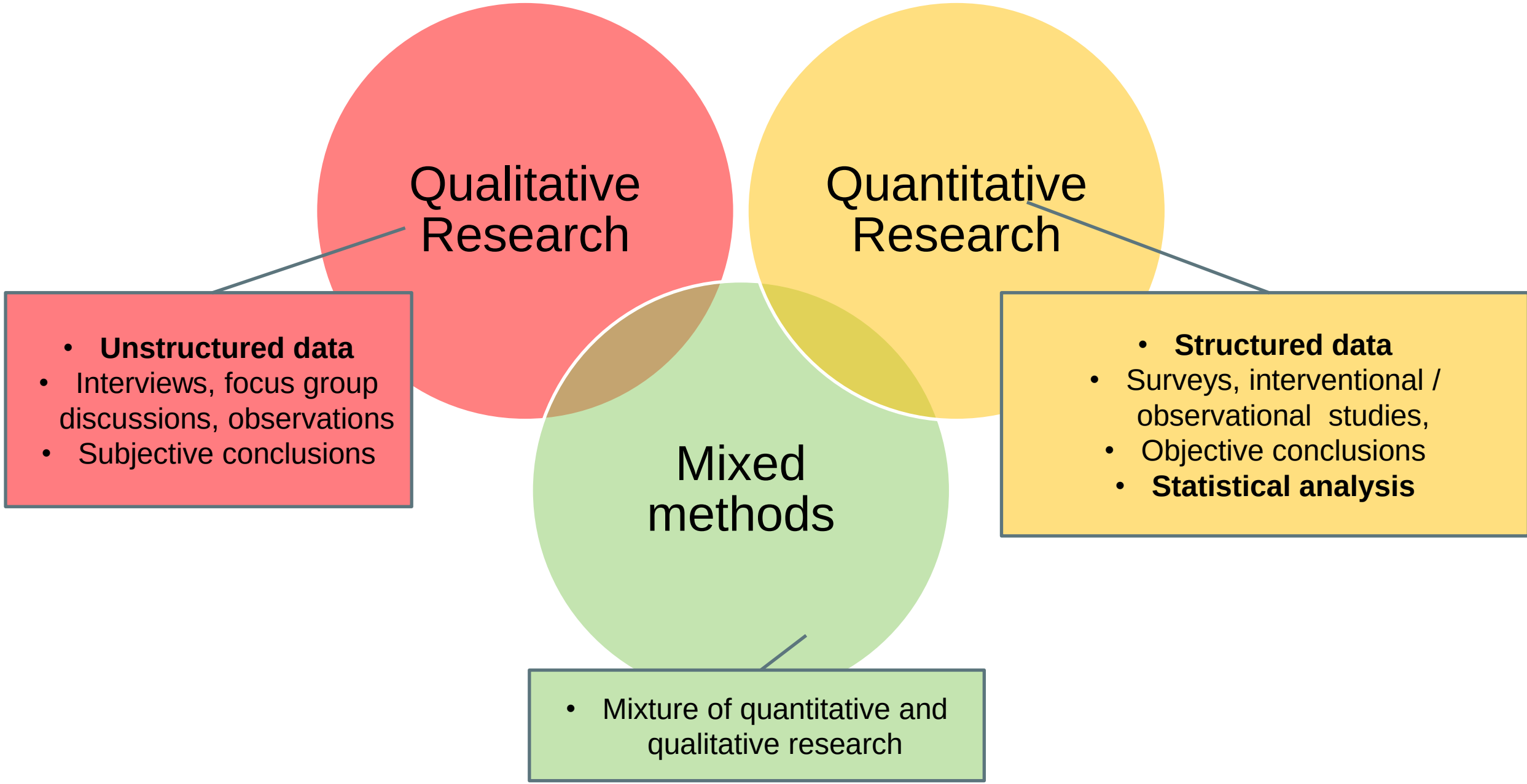
UP-ART Training
Research Methods Part 2: Introduction to Statistical Analysis I

Lizzie Chappell

Siobhan Crichton

Edith Milanzi

ICTM
INSTITUTE OF CLINICAL TRIALS AND METHODOLOGY


Qualitative Research
Quantitative Research
Mixed methods
• Unstructured data
• Interviews, focus group discussions, observations
• Subjective conclusions
• Structured data
• Surveys, interventional / observational studies,
• Objective conclusions
• Statistical analysis
• Mixture of quantitative and qualitative research

Nyogea et al. BMC Infectious Diseases (2015) 15:28
DOI 10.1186/s12879-015-0753-y

BMC Infectious Diseases

RESEARCH ARTICLE Open Access

Determinants of antiretroviral adherence among HIV positive children and teenagers in rural Tanzania: a mixed methods study

Daniel Nyogea[1,2*], Sally Mtenga[1], Lars Henning[1,2,4], Fabian C Franzeck[1,2], Tracy R Glass[2], Emilio Letang[1,2,3], Marcel Tanner[1,2] and Eveline Geubbels[1]

Abstract

Background: Around 3.3 million children worldwide are infected with HIV and 90% of them live in sub-Saharan Africa. Our study aimed to estimate adherence levels and find the determinants, facilitators and barriers of ART adherence among children and teenagers in rural Tanzania.

Methods: We applied a sequential explanatory mixed method design targeting children and teenagers aged 2–19 years residing in Ifakara. We conducted a quantitative cross sectional study followed by a qualitative study combining focus group discussions (FGDs) and in-depth interviews (IDIs). We used pill count to measure adherence and defined optimal adherence as > =80% of pills being taken. We analysed determinants of poor adherence using logistic regression. We held eight FGDs with adolescent boys and girls on ART and with caretakers. We further explored issues emerging in the FGDs in four in-depth interviews with patients and health workers. Qualitative data was analysed using thematic content analysis.

Results: Out of 116 participants available for quantitative analysis, 70% had optimal adherence levels and the average adherence level was 84%. Living with a non-parent caretaker predicted poor adherence status. From the qualitative component, unfavorable school environment, timing of the morning ART dose, treatment longevity, being unaware of HIV status, non-parental (biological) care, preference for traditional medicine (herbs) and forgetfulness were seen to be barriers for optimal adherence.

Conclusion: The study has highlighted specific challenges in ART adherence faced by children and teenagers. Having a biological parent as a caretaker remains a key determinant of adherence among children and teenagers. To achieve optimal adherence, strategies targeting the caretakers, the school environment, and the health system need to be designed.

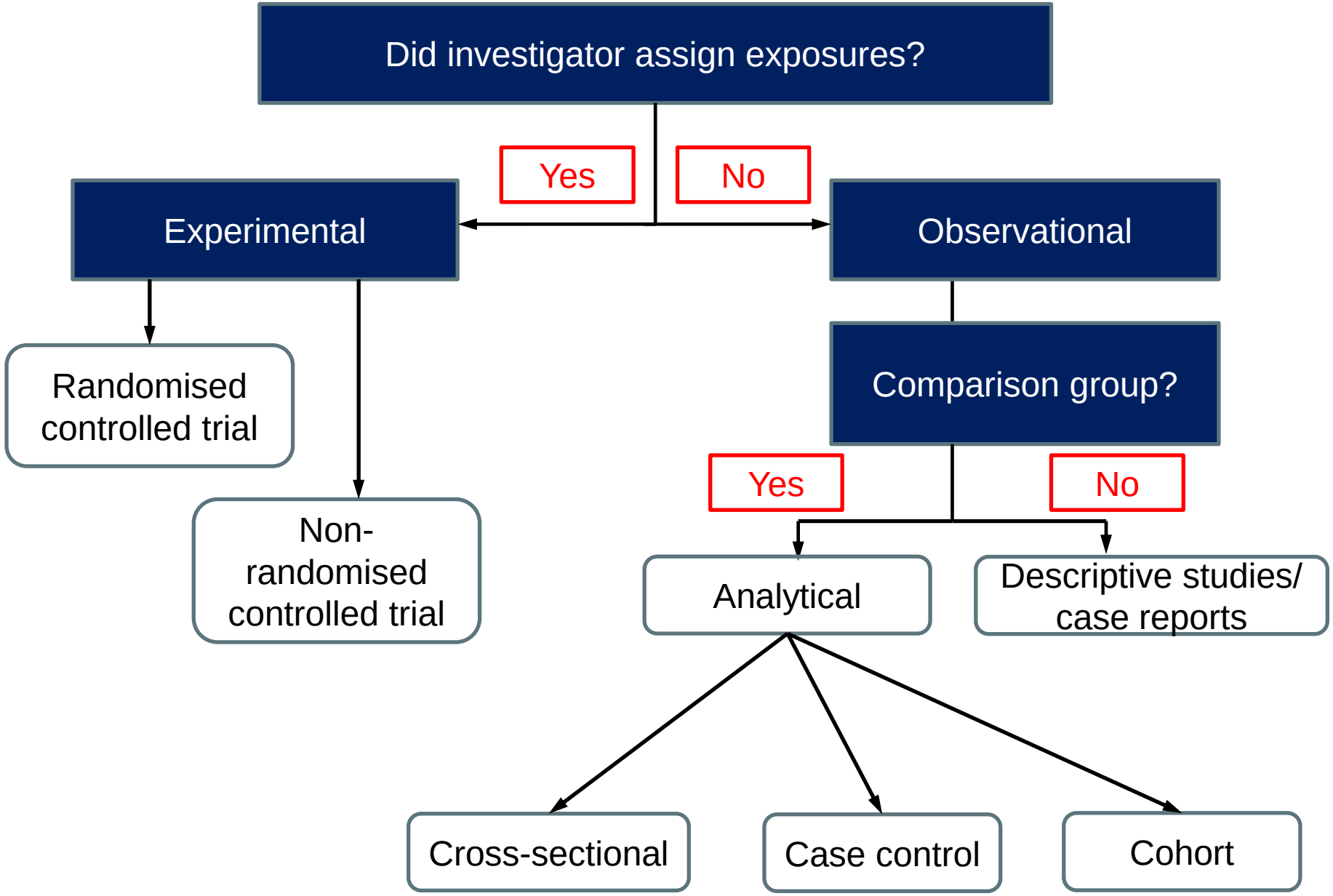
Table 2 Summary of characteristics of participants by adherence categories

Variables	Optimal n	Optimal %	Suboptimal n	Suboptimal %	P-Value
Sex					
Male	46	69	21	31	
Female	35	71	14	29	0.75
Age-school					
Pre-school age 2-5 yrs	23	77	7	23	
>6 yrs never been to school	11	58	8	42	
Primary school age and in primary	21	70	9	30	
Secondary school but in primary	19	70	8	30	
Secondary school age and in secondary	7	70	3	30	0.74
Baseline CD4 + cell count					
Below 350 cells/mm3	33	66	17	34	
Above 350cells/mm3	42	71	17	29	
Missing	6	86	1	14	0.54

Cross sectional study on adherence based on pill count (n=116), 70% had optimal adherence, factors associated with adherence.

N=56 in focus group discussions and 4 in-depth interviews with patients, caregivers and HCW on barriers to adherence

"My mother tells me to take drugs in the morning and I sometimes wake up very early and no food prepared, and when I take medicine (in the morning) I feel nausea but I go to school just like that... and I try not to miss school...". [FGD, Boy].



Adapted from Schulz and Grimes, The Lancet Handbook of Essential Concepts in Clinical Research

Aim of session

To be able to…

- Recognise different types of data, and know how these can be summarised
- Understand the need to summarise uncertainty in our estimates using confidence intervals, and know how to interpret these
- Use and interpret a p-value to test hypotheses about our data
- Understand the relationship between confidence intervals and p-values

Outline

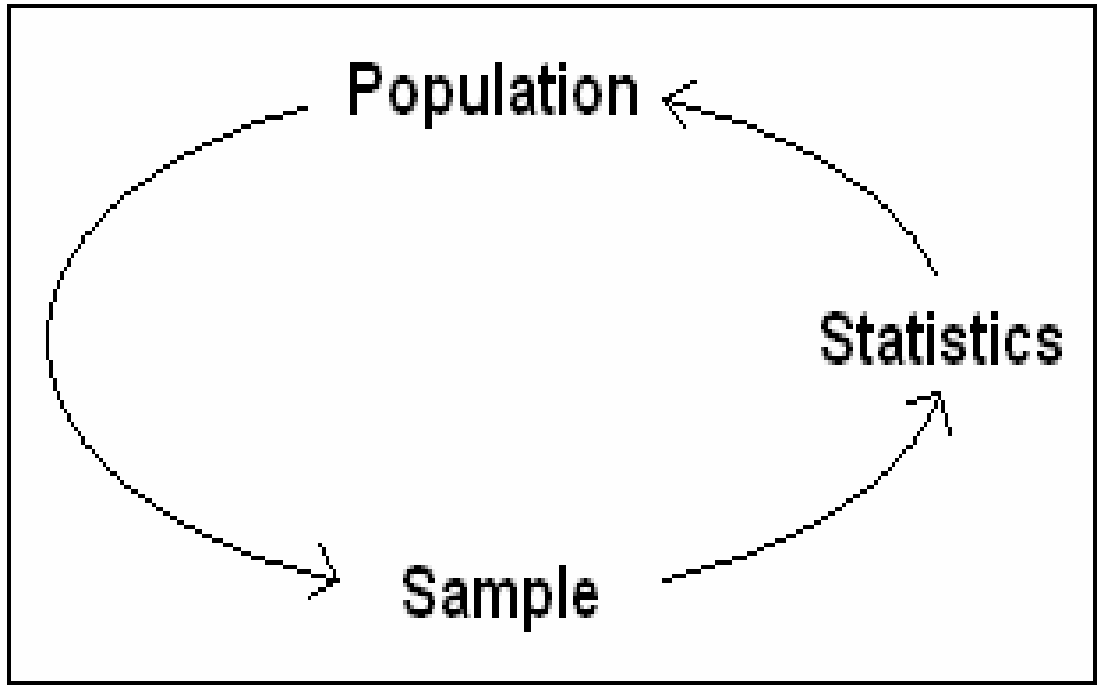
- Descriptive statistics
- Confidence intervals
- Hypothesis testing

DESCRIPTIVE STATISTICS

Section 1: Introduction

Aims:

- Use sample of data to make inference regarding a population.
- Collect, summarise, and present data.
- To understand and explain associations and variation in data.



What do we do?

- What is the scientific question?
- Design an appropriate study and analysis plan.
- Calculate the sample size.
- Collect data - importance of data quality.
- Validation and consistency checks required.
- Appropriate analysis.

Section 2: Types of Data

- A variable is a measurable characteristic or attribute. In quantitative research, we distinguish between two types of variables

Categorical data

- **Binary**
- **Ordinal**
- **Unordered (or nominal)**

Numerical

- **Continuous**
- **Discrete**

Categorical variables

Binary – two categories

e.g. male/female, dead/alive etc.

Ordered (ordinal) – three or more ordered categories

e.g. tumour status, WHO HIV stage (1-4) etc.

Unordered (nominal) – three or more unordered categories

e.g. country of birth, study centre etc.

Numerical variables

Continuous - can take any value (in a given range)

e.g. age, height etc.

Discrete - can only take specific (usually integer) values

e.g. year of birth, number of drugs etc.

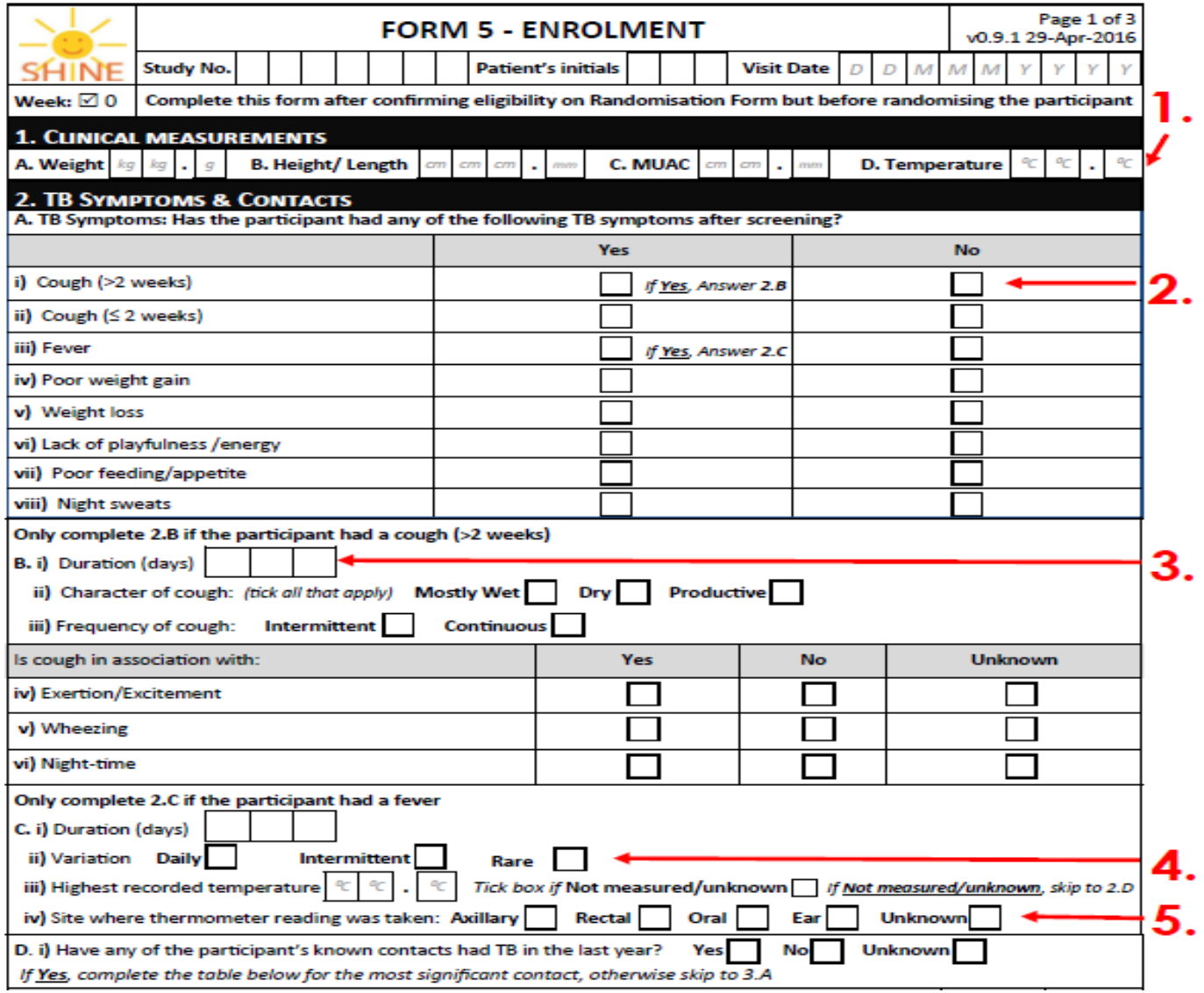
Quiz

Types of Data: Examples

What type of data are the variables highlighted on the CRF?

Types: Binary, Ordinal, Nominal, Continuous and Discrete

Answers available at end of slide set



SHINE

FORM 5 - ENROLMENT

Page 1 of 3
v0.9.1 29-Apr-2016

Study No. | Patient's initials | Visit Date D D M M M Y Y Y Y

Week: ☑ 0 — Complete this form after confirming eligibility on Randomisation Form but before randomising the participant

1. CLINICAL MEASUREMENTS — 1.

A. Weight kg kg . g | B. Height/ Length cm cm cm . mm | C. MUAC cm cm . mm | D. Temperature °C °C . °C

2. TB SYMPTOMS & CONTACTS

A. TB Symptoms: Has the participant had any of the following TB symptoms after screening?

	Yes	No
i) Cough (>2 weeks)	☐ *If Yes, Answer 2.B*	☐ — 2.
ii) Cough (≤ 2 weeks)	☐	☐
iii) Fever	☐ *If Yes, Answer 2.C*	☐
iv) Poor weight gain	☐	☐
v) Weight loss	☐	☐
vi) Lack of playfulness /energy	☐	☐
vii) Poor feeding/appetite	☐	☐
viii) Night sweats	☐	☐

Only complete 2.B if the participant had a cough (>2 weeks)

B. i) Duration (days) ☐☐☐ — 3.

ii) Character of cough: *(tick all that apply)* Mostly Wet ☐ Dry ☐ Productive ☐

iii) Frequency of cough: Intermittent ☐ Continuous ☐

Is cough in association with:	Yes	No	Unknown
iv) Exertion/Excitement	☐	☐	☐
v) Wheezing	☐	☐	☐
vi) Night-time	☐	☐	☐

Only complete 2.C if the participant had a fever

C. i) Duration (days) ☐☐☐

ii) Variation Daily ☐ Intermittent ☐ Rare ☐ — 4.

iii) Highest recorded temperature °C °C . °C *Tick box if* Not measured/unknown ☐ *if Not measured/unknown, skip to 2.D*

iv) Site where thermometer reading was taken: Axillary ☐ Rectal ☐ Oral ☐ Ear ☐ Unknown ☐ — 5.

D. i) Have any of the participant's known contacts had TB in the last year? Yes ☐ No ☐ Unknown ☐

If Yes, complete the table below for the most significant contact, otherwise skip to 3.A

Section 3: Describing and summarising data

Descriptive analysis should be the starting point of any statistical investigation.

Tables and graphical methods.

- exploring distributions
- investigating relationships
- inspection for outliers

Summary measures.

- “typical” value
- spread or range of values

Descriptive statistics and graphical methods : Categorical data

Baseline data tables are useful for deciding if data is representative of the target population

Frequency and relative frequency

Degree of stunting at start of treatment in 4815 children living with HIV

Stunting at start of treatment	Frequency	Relative frequency
Not stunted	2507	52.1%
Stunted	1267	26.3%
Severely stunted	1041	21.6%
Total	4815	100%

Contingency tables

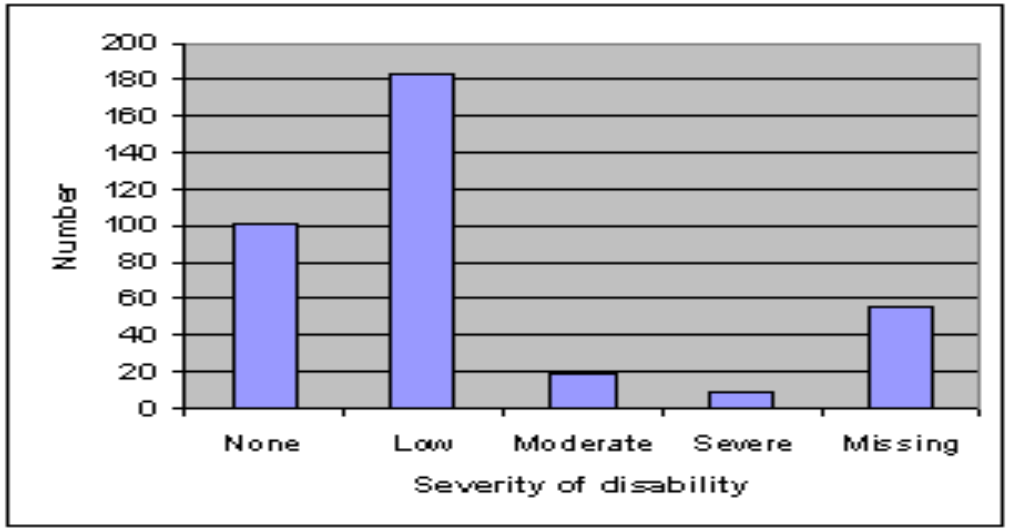
We may wish to look at two categorical variables

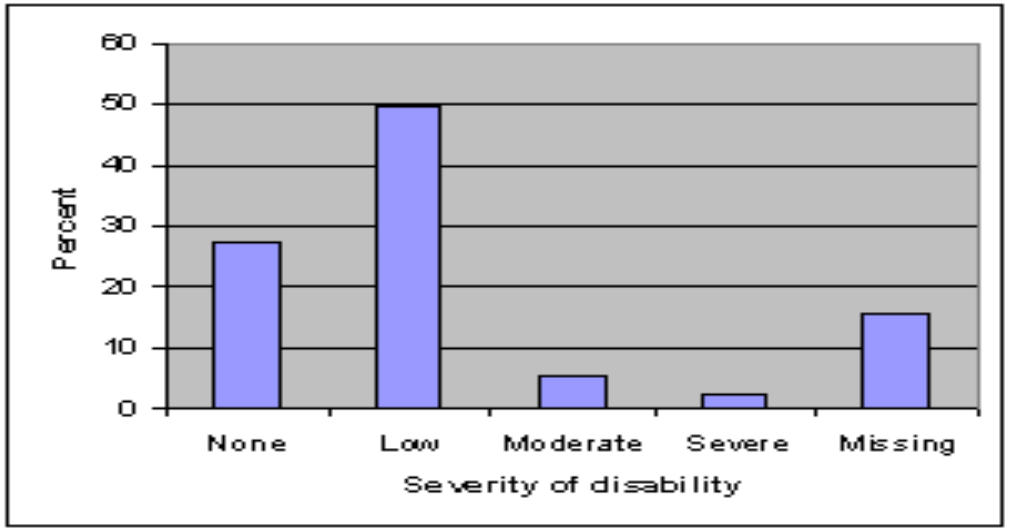
Degree of stunting at start of treatment in 4815 children living with HIV by region

Stunting at start of treatment	Eastern and Central Africa	Botswana and South Africa	Western and rest of Southern Africa	Europe and North America	Latin America	Asia	Total
Not stunted	1178 (47.2%)	429 (52.5%)	189 (61.4%)	419 (82.5%)	93 (5%)	199 (39.4%)	2507 (52.1%)
Stunted	709 (28.4%)	234 (28.6%)	67 (21.8%)	66 (13%)	53 (29.6%)	138 (27.3%)	1267 (26.3%)
Severely stunted	611 (24.5%)	154 (18.8%)	52 (16.9%)	23 (4.5%)	33 (18.4%)	168 (33.3%)	1041 (21.6%)
Total	2498	817	308	508	179	505	4815

Bar charts

Level of manual disability in 368 adults with cerebral palsy





Categories on x-axis, frequencies (percentage) on y-axis.

Height indicates number (percentage) in each group. Bars have same width.

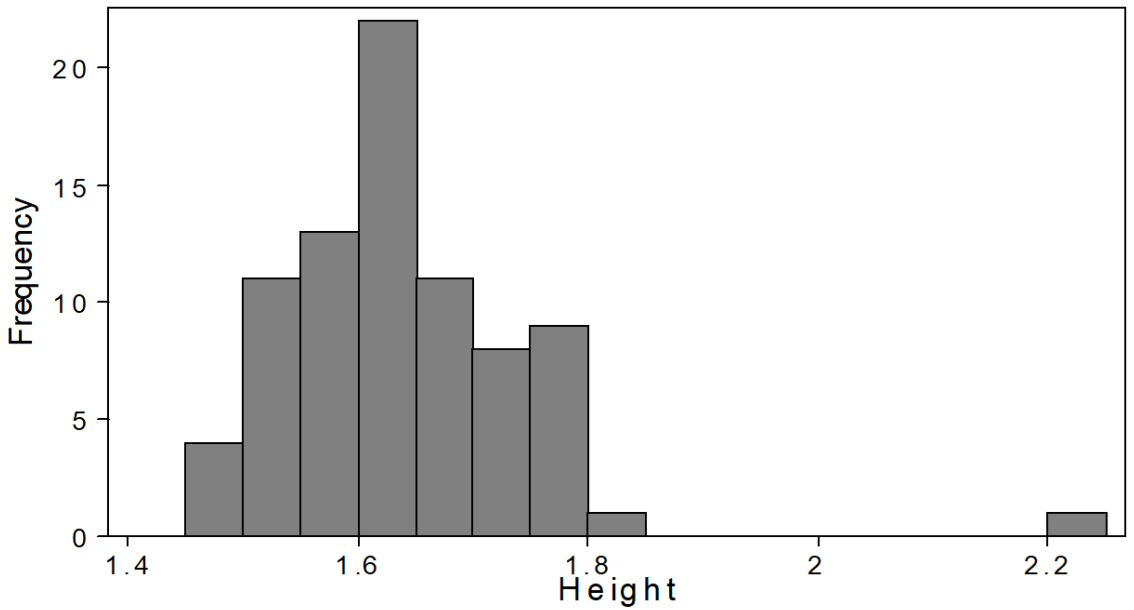
Descriptive statistics and graphical methods: Continuous data

Heights of 80 South Africans.

1.51 1.75 1.52 1.59 1.56 1.73 1.68 1.63 1.61 1.63
1.62 1.73 1.78 1.75 1.57 1.52 1.63 1.61 1.69 1.58
1.64 1.54 1.68 1.68 1.63 1.49 1.53 1.62 1.62 1.58
1.55 1.66 1.77 1.73 1.66 1.65 1.53 1.48 1.70 1.59
1.77 1.64 1.62 1.60 1.76 1.52 1.56 1.68 1.80 1.50
1.58 1.61 1.60 1.66 1.72 1.59 1.57 1.62 1.52 1.67
1.57 1.58 1.75 1.58 1.62 1.73 1.70 1.84 2.20 1.54
1.66 1.60 1.46 1.46 1.70 1.76 1.64 1.68 1.64 1.65

We can summarise this data using numerical or graphical methods as before.

Histograms

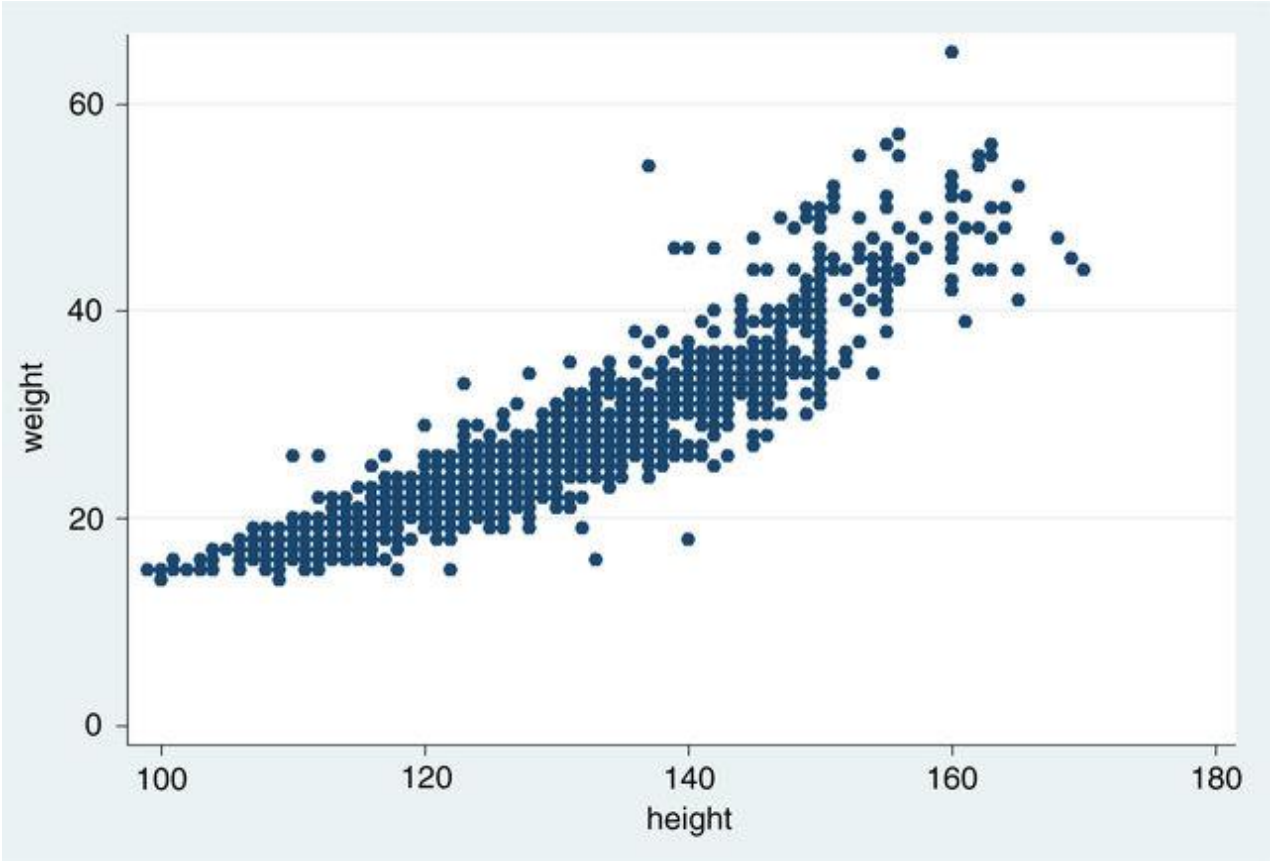


- Used to explore the shape of the distribution
- position of the peak and degree of symmetry
- spread of values about peak

- Also used to spot outliers
- an observation inconsistent with the remaining data
- may have an unduly large influence

Scatter plots

- Use to look for correlation between two continuous variables
- Can visually inspect direction of correlation
- Weight increasing as height increases-> positive correlation



Scatter plot weight and height

Summary measures

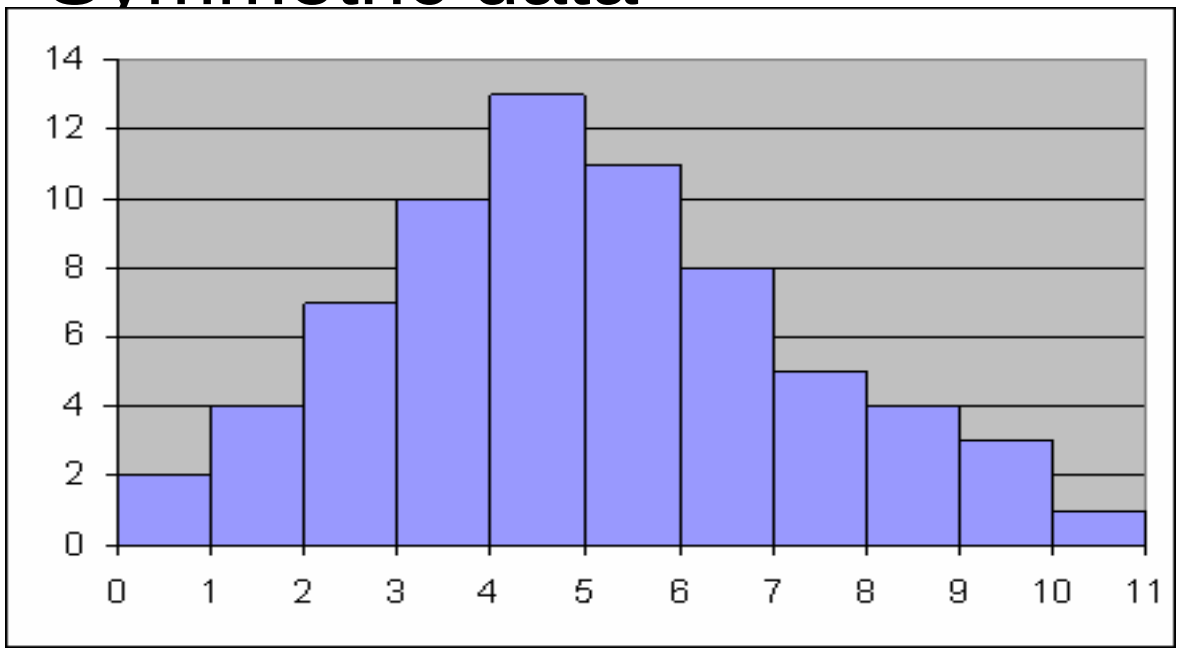
Summary measures of location and spread.

Presentation of these should be based on a consideration of the data.

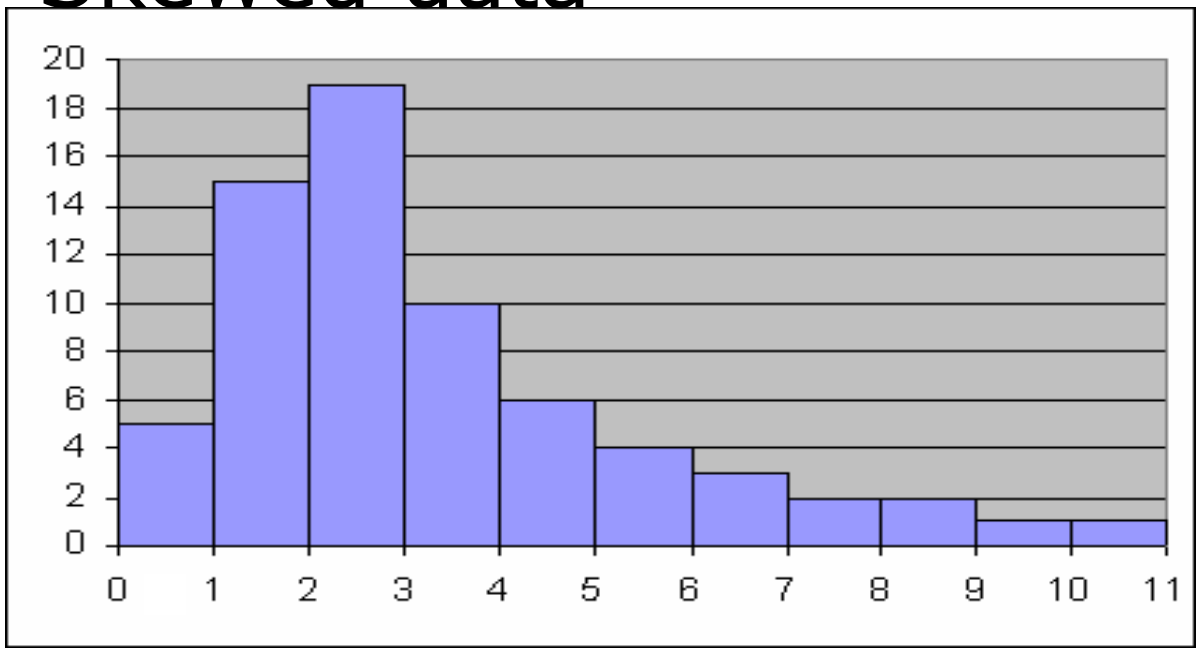
Can also be used to check data.

Symmetric or skewed?

Symmetric data



Skewed data



Positive skew: long tail on right hand side

Negative skew: long tail on left hand side (rare)

Location: Mean or median?

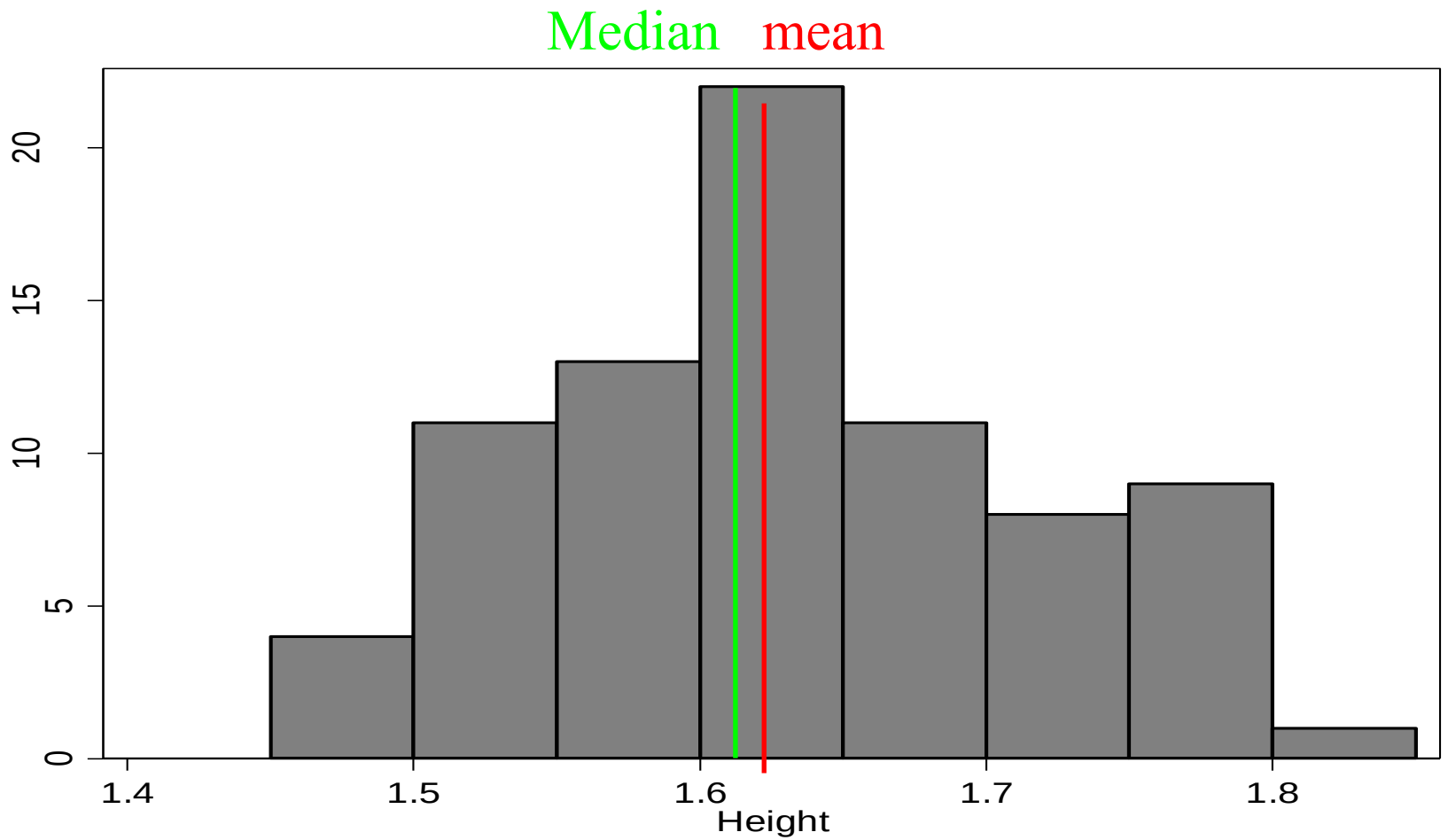
Symmetric data: use arithmetic mean

- mean = sum of all values / number of observations

Skewed data: use median

- median = order observations and take middle value

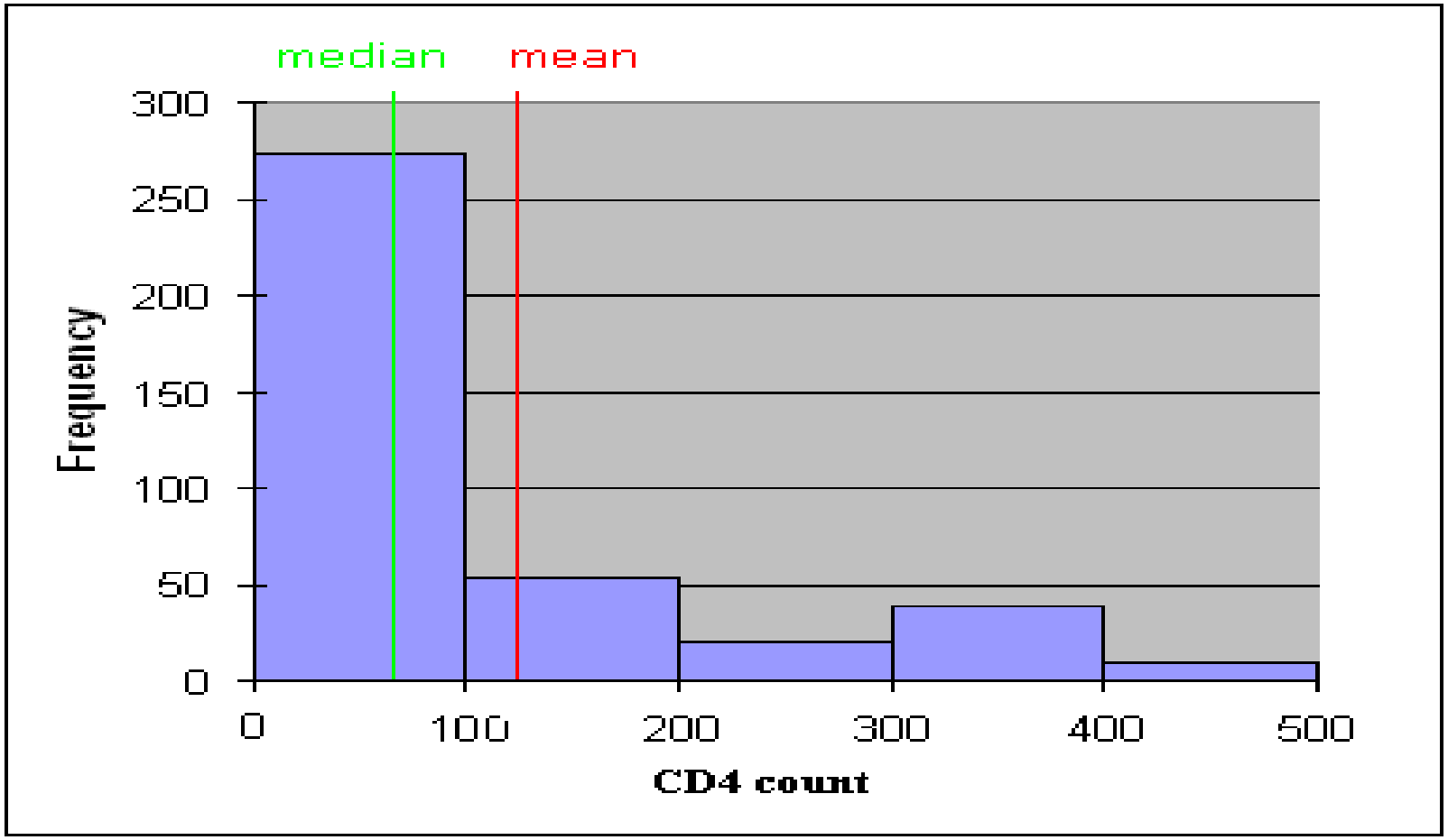
Symmetric data



Mean = 1.63

Median = 1.62

Skewed data



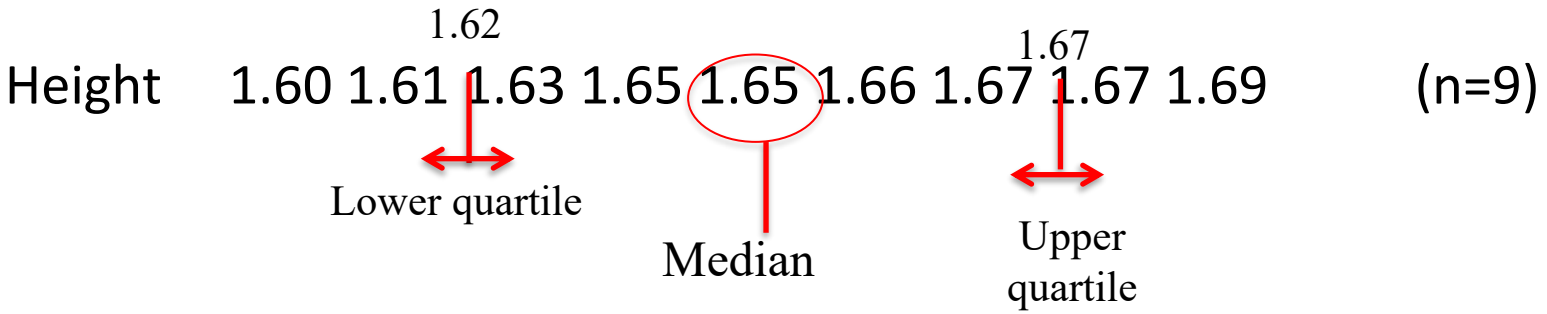
Mean = 126.5

Median = 67

Spread: Centiles and quartiles

The cth centile is the observation below which c% of the observations lie

- median is the 50th centile
- 25th centile known as lower quartile
  - a quarter of the sample lies below the lower quartile
- 75th centile known as upper quartile
  - a quarter of the sample lies above the upper quartile



Spread: Range and IQR

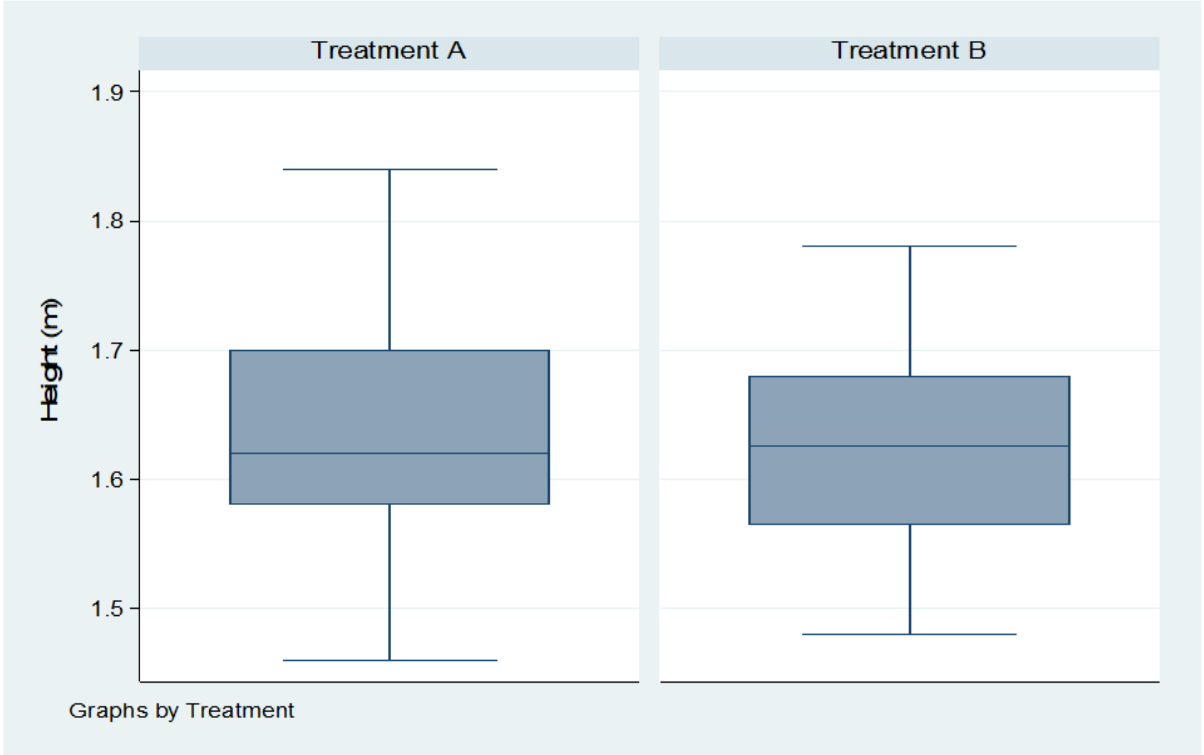
Range
- minimum to maximum value
- e.g. Height range 1.46 to 1.84 metres

Inter-quartile range (IQR)
- 25th to 75th centile
- e.g. Height IQR 1.57 to 1.68 metres

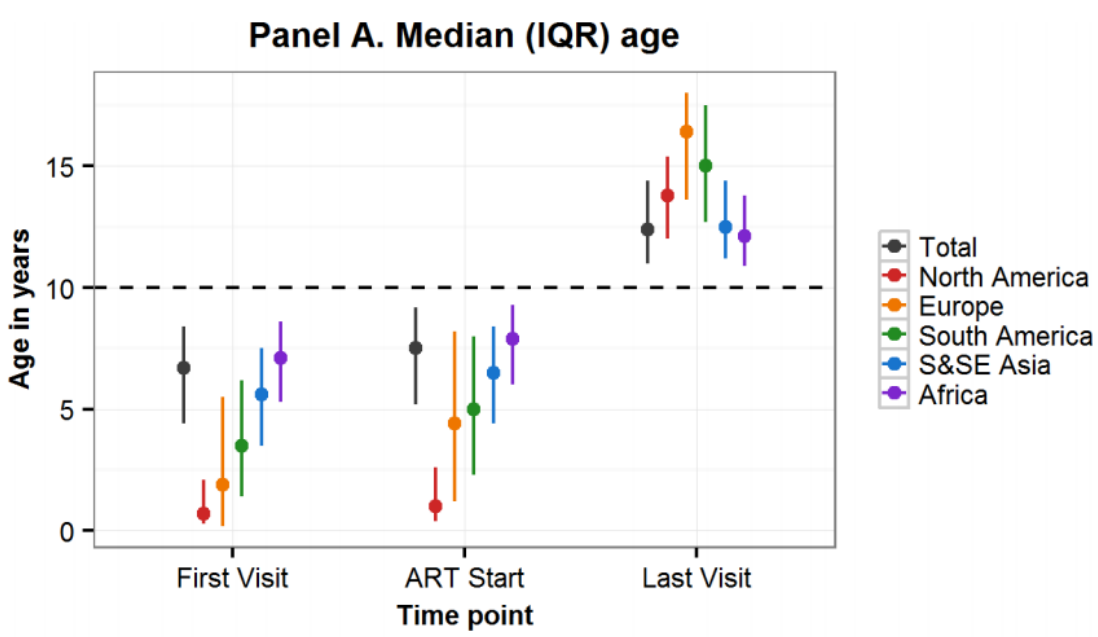
Use median, lower and upper quartiles, and median ± (1.5*IQR).

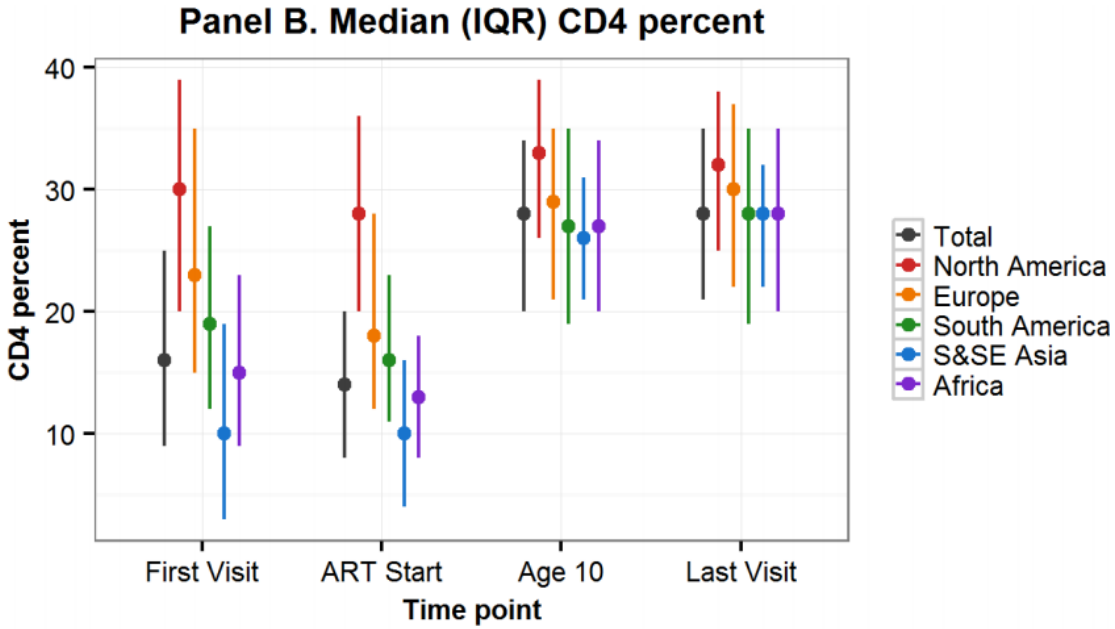
Good for comparing distributions.
Can be visualized by **boxplots**

Box plots for Height



More ways to visualize spread





Spread: Mean and Standard Deviation

Mean

- sum of all values / number of observations

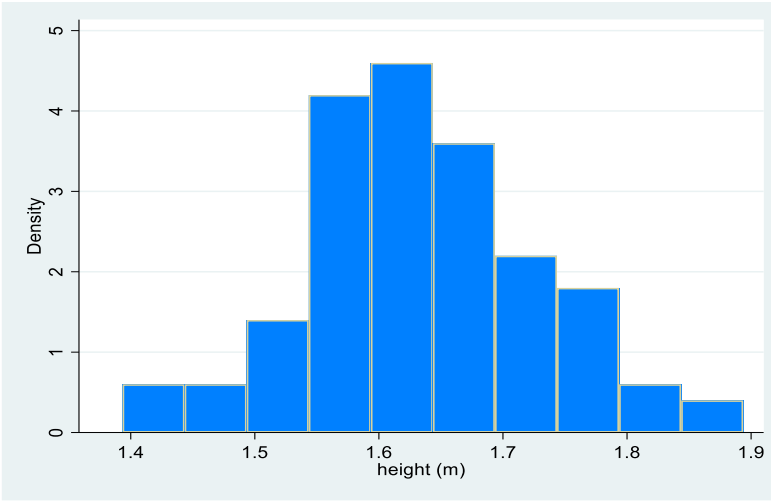
Variance (σ2)

- average squared deviation around the mean
- (each value-mean)2/number of observations
- tells you how tightly the data is clustered around the mean

Standard deviation (σ)

- square root of variance
- same scale as measurements
- Heights: SD = 0.08 metres

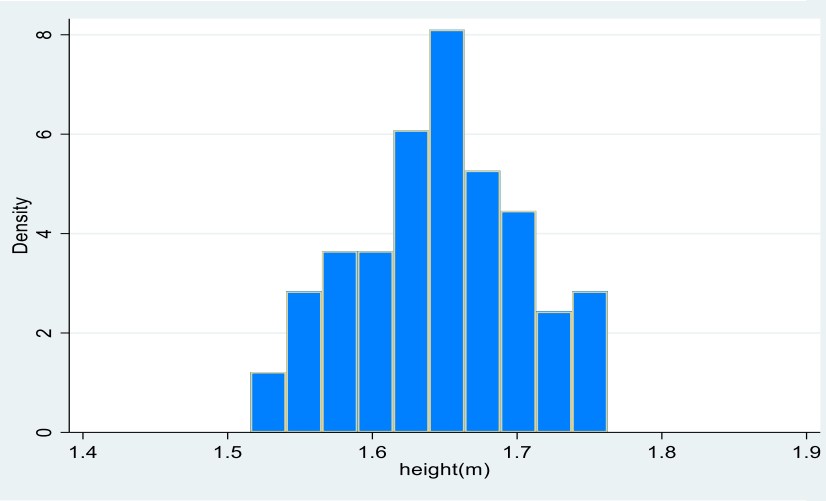
Spread: Mean and Standard Deviation



Mean = 1.65m

SD=0.1m

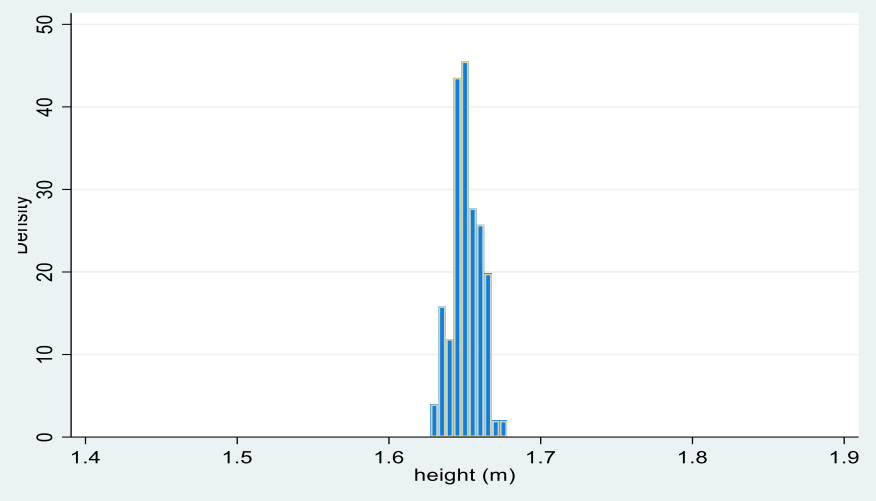
(n=100)



Mean = 1.65m

SD=0.05m

(n=100)



Mean = 1.65m

SD=0.01m

(n=100)

Overview of summary measures

Mean and standard deviation

- uses information from every observation
- can be distorted by outliers or skewed data

Median and inter-quartile range

- better for asymmetrical distributions or outliers
- less efficient use of data
- less easy to handle mathematically

Range

- increases with sample size
- largest and smallest observations most likely to be suspect

Section 4. Statistical distributions

Some statistical methods require distributional assumptions → can usually check these using our data.

- Independence
  - Probability that an event occurs for an individual is unrelated to outcomes of other individuals.

- Normal Distribution
  - what is it and why is it so important?
  - transforming to Normality

Summary

- Why we do statistics!
- Importance of looking at the types of data.
- How we might check data.
- How to start a statistical analysis.
- Making (and verifying) assumptions.

CONFIDENCE INTERVALS

Exposure or outcome?

Exposure = explanatory, independent, risk factor.

Outcome = response, dependent variable.

Important to distinguish between exposure and outcome variables to identify the questions of interest and appropriate methods.

Example: Exposure – Age at diagnosis, sex, ART regimen

Outcome – CD4 count, viral load, blood pressure, weight

Outline of the research process

Form a **research hypothesis / research question**

↓

Design an appropriate study

↓

Conduct the study and collect the data

↓

Carry out a **statistical analysis** of the data

- Estimate parameters
- Draw conclusions

Estimation: From Sample to Population

- Ideally want to know about the whole population
  - Not practical or necessary
  - So take a sample
  - Use information from this sample to make inferences about the population

- Two concepts in estimation:
  1. ‘True’ values: The true – unknowable – values in the entire population
  2. ‘Estimated’ values: Estimated from the sample we have taken

Uncertainty in our estimates

- Imagine taking lots of repeated samples from our population
- Estimate prevalence for each sample
- Our samples will all be slightly different, and so will have different prevalence values
- Uncertainty in our estimated values and thus what the true value is

How to summarise this uncertainty?

- Standard errors

- Confidence intervals

What do we want to estimate?

- Two types of data:
  - Categorical/binary *e.g. prevalence of HIV*
  - Continuous *e.g. blood pressure*

- Two types of measure
  - Estimate a single value *e.g. average blood pressure in a defined population*
  - Estimate a difference between two values *e.g. difference between average blood pressure in the treatment group and in the control group*

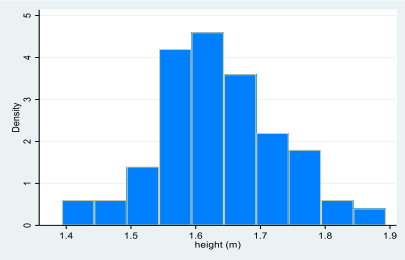
Example: what is the average blood pressure among children living with HIV?

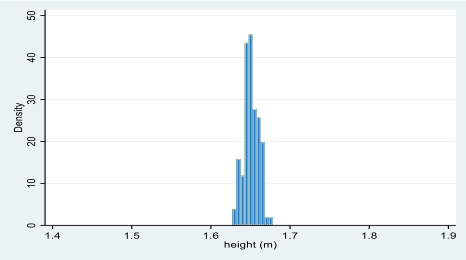
Standard error of the mean

- The standard error measures the uncertainty in the estimate of the mean (*m*)
  - Standard error = *standard deviation* / √*n*

With bigger SD:

- Estimates more uncertain





With bigger N:

- Mean get closer to true value

Confidence Intervals (CI): Definition

- A range of values that are our best guess of where the true population parameter might lie
- Conventionally use **95% confidence intervals**
  - If we took repeated samples from the population, 95% of these CI would be expected to contain the true value
- There is always a chance that the CI will not contain the true population parameter

Calculate a Confidence Interval (CI)

- XX% CI for the true population mean is:

sample mean ± (multiplier x **standard error**)

where the *multiplier* depends on the required level of the confidence interval e.g. 95%, 90% or 99%

- For a **95%** confidence interval: multiplier = 1.96 (based on the standard Normal distribution)

Recap

- The standard error measures the uncertainty in the estimate of the mean

$$\text{Standard error of mean (SE)} = \text{standard deviation} / \sqrt{n}$$

- Confidence interval is the range of values that we are confident includes the true population parameter

$$95\%\ CI = mean \pm (1.96 * standard\ error)$$

Example 1

- **Research question:** What is the mean blood pressure in Ugandan children with HIV?
- **Population:** All children with HIV in Uganda
- **Sample:** Children attending a healthcare facility during a one week study period
- **Data collected:** N = 50, Mean = 100 mmHg, SD = 13.7

Example 1

sample size (N) = 50; mean = 100 mmHg; SD = 13.7 mmHg

1. Calculate the standard error of the mean blood pressure

$$\text{Standard error} = SD/\sqrt{n}$$

$$= 13.7/\text{sqrt}(50) = 1.9$$

2. Calculated the associated CI

$$95\%\ CI = mean \pm 1.96 * SE$$

$$= 100 \pm 1.96 \times 1.9 = 100 \pm 3.8$$

$$= (96.2, 103.8)$$

3. Interpret the CI

"We can be 95% sure that the true mean blood pressure is between 96.2 and 103.8 mmHg"

What do we want to estimate?

- Two types of data:
  - Categorical/binary *e.g. prevalence of HIV*
  - Continuous *e.g. blood pressure*

- Two types of measure
  - Estimate a single value *e.g. average blood pressure in a defined population*
  - Estimate a difference between two values *e.g. difference between average blood pressure in the treatment group and in the control group*

Motivating example: what is the average blood pressure among children living with HIV?

Comparing two populations

- In practice we often want to **compare the results of two groups** rather than just one.

Standard Error of a difference

- The standard error of a difference measures the uncertainty in the estimate of the difference between the means (m_1 – treatment, m_2 – control

- Difference (treatment effect) = $m_1 - m_2$
- SE (d) $= \text{SE}(m_1 - m_2)$

$$= \sqrt{[\text{SE}(m_1)^2 + \text{SE}(m_2)^2]}$$

$$= \sqrt{[(SD_1/\sqrt{N_1})^2 + (SD_2/\sqrt{N_2})^2]}$$

With bigger N:

- Estimate of difference closer to true value

With bigger SD:

- Estimates of difference more uncertain

Calculate a CI of a difference

- XX% CI for the true population difference is:

  difference ± multiplier x **SE (difference)**

  where the *multiplier* depends on the required level of the confidence interval e.g. 95%, 90% or 99%

- For a **95%** confidence interval: multiplier = 1.96 (based on the standard Normal distribution)

Example 2

- **Research question:** Is there a difference in mean blood pressure between girls and boys living with HIV in Uganda?
- **Population:** All children with HIV in Uganda
- **Sample:** Children attending a healthcare facility during a one week study period
- **Data collected**
  - Girls: N = 24, Mean = 98, SD= 13.1, SE = 2.7
  - Boys: N = 26, Mean = 101, SD= 14.2, SE = 2.8

Example 2

Girls: N = 24, Mean = 98, SD= 13.1, SE = 2.7
Boys: N = 26, Mean = 101, SD= 14.2, SE = 2.8

1. Calculate the difference and standard error of the difference of average blood pressure between groups

Diff = 101 – 98 = 3

SE (diff) = $\sqrt{[SE(m_B)^2 + SE(m_G)^2]} = \sqrt{[2.8^2 + 2.7^2]} = 3.9$

2. Calculated the associated CI

95% CI = $diff \pm 1.96*SE$

= $3 \pm 1.96 \times 3.9$

= 3 ± 7.6

= (-4.6, 10.6)

3. Interpretation

"We can be 95% sure that the difference in blood pressure between girls and boys is between -4.6 and 10.6"

What do we want to estimate?

- Two types of data:
  - Categorical/binary *e.g. prevalence of HIV*
  - Continuous *e.g. blood pressure*

- Two types of measure
  - Estimate a single value *e.g. average blood pressure in a defined population*
  - Estimate a difference between two values *e.g. difference between average blood pressure in the treatment group and in the control group*

SE and CI of a single proportion

- e.g. What is the true HIV prevalence (p) in Uganda?

Standard error of a proportion $\sqrt{\frac{p*(1-p)}{n}}$

Confidence interval $p \pm (1.96 * \sqrt{\frac{p*(1-p)}{n}})$

SE and CI of difference in proportions

- e.g. What is the difference in HIV prevalence between men (p_1) and women (p_2) in Uganda?

Standard error of difference $\sqrt{\frac{p_1 *(1-p_1)}{n_1} + \frac{p_2 *(1-p_2)}{n_2}}$

Confidence interval $(p_1 - p_2) \pm (1.96 * \sqrt{\frac{p_1 *(1-p_1)}{n_1} + \frac{p_2 *(1-p_2)}{n_2}})$

Example 3

RESEARCH Open Access

Prevalence and risk factors for hypertension and diabetes among those screened in a refugee settlement in Uganda

Rachel W. Kubiak[1], Elinor M. Sveum[2], Zikama Faustin[3], Timothy Muwonge[4], Hussain Abbas Zaidi[5], Andrew Kambugu[4], Simon Masereka[6], Julius Kasozi[7], Ingrid V. Bassett[8] and Kelli N. O'Laughlin[2,9*]

Hypertension

Overall, 1067 (50%, 95% CI 48.0–52.2%) of participants met criteria for pre-hypertension at the time of their clinic visit and 187 (9%, 95% CI 7.7–10.1%) met criteria for hypertension. The number needed to screen to iden-

Diabetes

Overall, 32 participants met the criteria for diabetes (1.5%, 95% CI 1.1–2.1%). The number needed to screen

Kubiak et al, Confl Health, 2021

Quiz

1. If we increase the sample size, what do we expect to happen to the standard error?
   - Increase / decrease / no change

2. True or false, the estimated mean is always inside the confidence interval we calculate?

3. If we increase the sample size, what will happen to the CI?
   - Get narrower / get wider / no change

Answers available at end of slide set

HYPOTHESIS TESTING

Outline

- Laying out a research hypothesis in statistical terms
- Interpreting a p-value
- Relationship between p-values and confidence intervals

Start with a research hypothesis:

- What we expect or hope
- For example:
  - HIV viral load is lower on treatment A than treatment B
  - Adverse events on treatment A are more common in infants than in older children

Durand *et al. BMC Infectious Diseases* (2017) 17:611
DOI 10.1186/s12879-017-2692-2

BMC Infectious Diseases

STUDY PROTOCOL Open Access

The Canadian HIV and aging cohort study - determinants of increased risk of cardio-vascular diseases in HIV-infected individuals: rationale and study protocol

Madeleine Durand[1*], Carl Chartrand-Lefebvre[2], Jean-Guy Baril[3], Sylvie Trottier[3], Benoit Trottier[3], Marianne Harris[4], Sharon Walmsley[5], Brian Conway[5], Alexander Wong[6], Jean-Pierre Routy[7], Colin Kovacs[8], Paul A. MacPherson[9], Kenneth Marc Monteith[9], Samer Mansour[9], George Thanassoulis[10], Michal Abrahamowicz[11], Zhitong Zhu[11], Christos Tsoukas[12], Petronela Ancuta[13], Nicole Bernard[14], Cécile L. Tremblay[13] and For the investigators of the Canadian HIV and Aging Cohort Study

Abstract

Background: With potent antiretroviral drugs, HIV infection is becoming a chronic disease. Emergence of comorbidities, particularly cardiovascular disease (CVD) has become a leading concern for patients living with the infection. We hypothesized that the chronic and persistent inflammation and immune activation associated with HIV disease leads to accelerated aging, characterized by CVD. This will translate into higher incidence rates of CVD in HIV infected participants, when compared to HIV negative participants, after adjustment for traditional CVD risk factors. When characterized further using cardiovascular imaging, biomarkers, immunological and genetic profiles, CVD associated with HIV will show different characteristics compared to CVD in HIV-negative individuals.

Study hypothesis

"The chronic and persistent inflammation associated with HIV disease leads to accelerated aging, characterized by premature CVD, altered metabolism and immune senescence. ***This will translate into higher incidence rates of CVD in HIV infected participants, when compared to HIV negative participants, after adjustment for traditional CVD risk factors****."*

Received: 15 May 2018 | Revised: 12 November 2018 | Accepted: 10 April 2019 | First published online: 29 April 2019

DOI: 10.1002/ijgo.12823

CLINICAL ARTICLE

Obstetrics

WILEY GYNECOLOGY OBSTETRICS FIGO

HIV serostatus, viral load, and midtrimester cervical length in a Zambian prenatal cohort

Joan T. Price[1,2,*] | Bellington Vwalika[3] | Jennifer Winston[1] | Andrew Kumwenda[3] | Mwansa K. Lubeya[3] | Katelyn J. Rittenhouse[2] | Elizabeth Stringer[1] | Margaret P. Kasaro[2] | Jeffrey S.A. Stringer[1]

[1] Department of Obstetrics and Gynecology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2] University of North Carolina Global Projects Zambia, Lusaka, Zambia

[3] Department of Obstetrics and Gynaecology, University of Zambia School of Medicine, Lusaka, Zambia

*Correspondence

Joan T. Price, Division of Global Women's Health, Department of Obstetrics and Gynecology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Email: Joan_Price@med.unc.edu

Funding Information

Global Alliance to Prevent Prematurity and Stillbirth; Center for AIDS Research; National Institutes of Health

Abstract

Objective: To evaluate whether maternal HIV serostatus and plasma viral load (VL) are associated with midtrimester cervical length (CL).

Methods: The Zambian Preterm Birth Prevention Study (ZAPPS) is an ongoing prospective cohort that began enrolling in Lusaka in August 2015. Pregnant women undergo ultrasound to determine gestational age and return for CL measurement at 16–28 weeks. We evaluated crude and adjusted associations between dichotomous indicators and short cervix (≤2.5 cm) via logistic regression, and between VL and CL as a continuous variable via linear regression.

Results: This analysis includes 1171 women enrolled between August 2015 and September 2017. Of 294 (25.1%) HIV-positive women, 275 (93.5%) had viral load performed close to CL measurement; of these, 148 (53.8%) had undetectable virus. Median CL was 3.6 cm (IQR 3.5–4.0) and was similar in HIV-infected (3.7 cm, IQR 3.5–4.0) versus uninfected (3.6 cm, IQR 3.5–4.0) participants (P=0.273). The odds of short CL were similar by HIV serostatus (OR 0.64; P=0.298) and detectable VL among those infected

"We hypothesized that cervical length would be shorter among women with HIV *such that the risk of HIV- associated PTB could be at least partly attributable to shortened cervix"*

Statistical hypothesis:

- **Take the research hypothesis…**
  - What we expect or hope

- **…and turn it into a statistical hypothesis**
  - Something that can be statistically tested
  - Needs to be phrased as a **null hypothesis** and an **alternative hypothesis**

The null hypothesis: H_0

- The hypothesis that needs to be **rejected** in order to confirm the research hypothesis

  (i.e. the **opposite** of the research hypothesis)

- **Research hypothesis**
  - Treatment A is better than treatment B
- **Null hypothesis**
  - The difference between the two groups is equal to zero
    - H_0: $m_A=m_B$
    - H_0: $m_A-m_B=0$

The alternative hypothesis: H_1

- The opposite of H_0
  - Specifies a way in which H_0 may be false

- **Null hypothesis**
  - H_0: $m_A - m_B = 0$

- **Alternative hypothesis**: mean difference ≠ 0
  - H_1: $m_A \neq m_B$
  - H_1: $m_A - m_B \neq 0$

Example: comparing blood pressure

- **Research hypothesis:**
  - Boys have higher blood pressure than girls
- **Statistical hypotheses:**
  - H_0: $m_B - m_G = 0$ (ie the difference in mean blood pressure in each group is 0)
  - H_1: $m_B - m_G \neq 0$ (ie the difference in mean blood pressure in each group is not 0)

Hypothesis testing

- Objective is to determine whether there is **sufficient evidence** to **reject H_0** in favour of H_1

- Results in a **probability statement** about the likelihood of the observed data **given that H_0 is true (p-value)**

- Many standard tests:
  - **t-test** for means
  - **χ^2-test** (chi-squared) for proportions
  - **log-rank test** for survival times

P - value

- The p-value is the probability that you could observe the results that you have observed **if H_0 were true**.

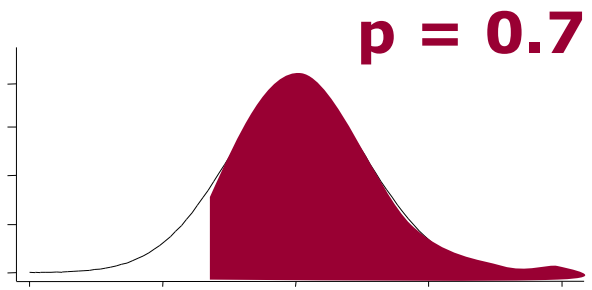
- If very unlikely (low p-value), then the assumption that H_0 is true is probably incorrect
  - **Reject H_0 in favour of H_1** 

- If likely (not a low p-value), then there is no reason to think that the assumption that H_0 is true is incorrect
  - **Cannot reject H_0** 

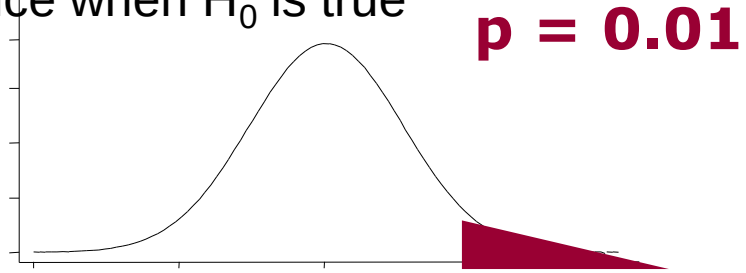
P-values

- **The probability of observing the sampled data (or more extreme) when H_0 is true.**

- Large p-value (e.g. p = 0.7)
  - Data could occur often when H_0 is true
  - Insufficient evidence to reject H_0
  - **Warning: Not evidence that H_0 is true**
  - ‘non-significant’
  - Weak / no evidence



- Small p-value (e.g. p = 0.01)
  - H_0 appears implausible since these data would rarely arise by chance when H_0 is true
  - Reject H_0 in favour of H_1
  - ‘significant’
  - Evidence

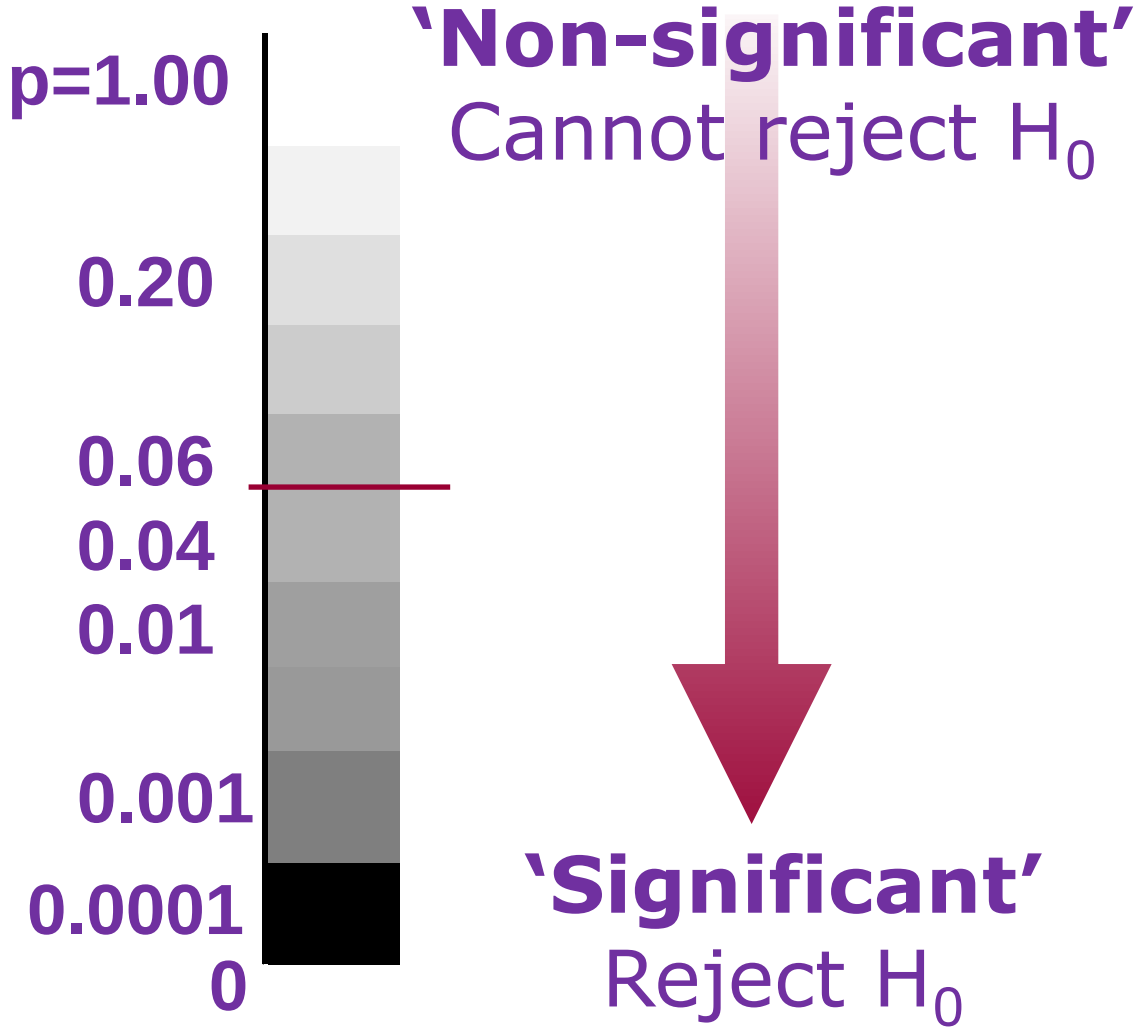


Where is the “significance” threshold?

- Conventional to use a significance level of **0.05**

- p>0.05
  - Not significant
  - No evidence
  - **Cannot reject H_0**
  - **Warning: we can never accept H_0**

- p <0.05
  - Significant
  - Strong evidence
  - **Reject H_0**



Use of p-values in practice

- **Recommendations**
  - Strict use of cut-offs is not advisable
    - In reality p=0.049 is not much different to p=0.051!
  - Report actual p-values
  - Always report quantitative measure with **confidence interval**

- **Warning**
  - ‘Significant’ p-values
    - **Not necessarily a clinically important effect**
  - ‘Non-significant’ p-values
    - **Not necessarily evidence of no effect**

Example: comparing blood pressure

Group	Number	Mean blood pressure	SE(mean)
Girls	24	98 mmHg	2.8
Boys	26	101 mmHg	2.7

- **Results:**
  - $d = m_B - m_G = 3$
  - SE(d) = 3.9
- **95% CI:**
  - $d \pm 1.96 \times \text{SE}(d)$
  - $= 3 \pm 1.96 \times 3.9$
  - = (-4.6, 10.6) mmHg

- **Hypothesis test (t-test)**
  - t = d / SE(d)
  - t = 3 / 3.9
  - t = 0.77
- **p-value = 0.402**

No significant difference!

Example: comparing blood pressure

Group	Number	Mean blood pressure	SE(mean)
Girls	24	98 mmHg	2.8
Boys	26	101 mmHg	2.7

- **Results:**
  - $d = m_B - m_G = 3$
  - SE(d) = 3.9
- **95% CI:**
  - d
  - [illegible]
  - [illegible].6) mmHg
- **Hypothesis test (**
  - t =
- **p-value = 0.402**

No significant difference!

There was no evidence of a difference in blood pressure between boys and girls (mean difference 3mmHg (95% CI -4.6, 10.6), p=0.402)

Example: comparing blood pressure

- **What if the results had been different?**
  - Estimated difference = 6 mmHg
  - 95% CI = (0.5, 11.5) mmHg
  - p=0.031 → 3.1% chance of seeing a difference this large if the truth was sex differences
  - Therefore, based on our data, **there is sufficient evidence to conclude that blood pressure is different in boys and girls**

Example: comparing blood pressure

- **What if the results had been different?**
  - Estimated difference = 6 mmHg
  - 95% CI =
  - ce of seeing a difference this large if the truth erences
  - Therefore, based on our data, **there is <u>sufficient</u> evidence to conclude that blood pressure is different in boys and girls**

There was a significant difference in blood pressure by sex with blood pressure on average 6mmHg (95%CI 0.5,11.5) higher in boys than girls (p=0.031).

Link to confidence intervals

95% CI includes value being tested in H_0 ⟷ P-value > 0.05 Not enough evidence to reject H_0

95% CI does not include value being tested in H_0 ⟷ P-value < 0.05 H_0 rejected

Example: blood pressure

Equivalent

CI = -4.6, 10.6 → includes 0 mmHg

p-value = 0.403 → cannot reject H_0

Summary: confidence intervals and hypothesis testing

- Close relationship:
  - $p < 0.05$ ⇔ 95% CI does not contain value in H_0
  - $p < 0.01$ ⇔ 99% CI does not contain value in H_0

- Both give **same evidence** for or against H_0
  - **Different perspectives on same approach**

- **However:**
  - Hypothesis tests only give probability statements about H_0

  - Confidence intervals provide a quantitative measure of the interval likely to contain the unknown parameter

Quiz

Comorbidity is more common and occurs earlier in persons living with HIV than in HIV-uninfected matched controls, aged 50 years and older: A cross-sectional study

Rafael Aguiar Maciel[a], Helena Moreira Klück[b], Madeleine Durand[c], Eduardo Sprinz[b,*]

[a] Medical Sciences Post-Graduation Program, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[b] School of Medicine, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[c] Université de Montréal, Québec, Canada

ARTICLE INFO

Article history:
Received 13 January 2018
Received in revised form 2 February 2018
Accepted 15 February 2018
Corresponding Editor: Eskild Petersen, Aarhus, Denmark

Keywords:
HIV
AIDS
Multimorbidity
Comorbidities
Aging
Developing countries
Brazil

ABSTRACT

Objectives: At present, data are limited on the comorbidity profiles associated with aging people with HIV in the developing world, where most such people live. The aim of this study was to compare the disease burden between older HIV-positive subjects and HIV-negative matched controls in Brazil.
Methods: This was a cross-sectional analysis of the South Brazilian HIV Cohort. Individuals aged 50 years and older were enrolled at Hospital de Clínicas de Porto Alegre and matched with HIV-negative controls from the primary practice unit of the same hospital. Multimorbidity (the presence of two or more comorbid conditions) and the number of non-infectious comorbidities were compared. Poisson regression was used to identify factors associated with multimorbidity.
Results: A total of 208 HIV-positive subjects were matched to 208 HIV-negative controls. Overall, the median age was 57 years and 56% were male. The prevalence of multimorbidity was higher in HIV-positive subjects than in HIV-negative controls (63% vs. 43%, $p < 0.001$), and the median number of comorbidities was 2, compared to 1 in controls ($p < 0.001$). The duration of HIV infection ($p = 0.02$) and time on treatment in years ($p = 0.015$) were associated with greater multimorbidity in HIV-positive persons.
Conclusions: In this large cohort from the developing world, multimorbidity was found to be more common in HIV-positive subjects than in HIV-negative controls. The duration of HIV and time on antiretrovirals were associated with multimorbidity.

- Cross-sectional analysis using data from South Brazilian HIV Cohort
- Individuals aged 50+ years enrolled from a single site and matched with HIV-negative controls from the primary practice unit of the same hospital.
- **Multimorbidity** (the presence of two or more comorbid conditions) compared in those with and without HIV

What is the null hypothesis being tested in this study?

NB: p+ = proportion of people living with HIV with multimorbidity, P- = proportion of people without HIV with multimorbidity

A H0 p+ > p- (i.e. Proportion of patients with multimorbidity is higher in those with HIV than those without)
B H0 p+ = p- (i.e. Proportion of patients with multimorbidity is the same in those with HIV and those without)
C H0 p+ =/= p- (i.e. Proportion of patients with multimorbidity is not the same in those with HIV and those without)

Answers available at end of slide set

Quiz

Table 3
Comparison of the burden of comorbidities between HIV-positive patients and non-HIV controls.

	HIV ($n=208$)	Non-HIV ($n=208$)	p-Value
Cardiovascular disease, n (%)	20 (9.6)	20 (12.5)	0.435
Kidney disease, n (%)	35 (16.8)	14 (6.7)	0.002
Hepatic disease, n (%)	53 (25.5)	14 (6.7)	<0.001
Diabetes, n (%)	47 (22.6)	59 (28.4)	0.216
Hypertension, n (%)	129 (62.0)	145 (69.7)	0.121
Neoplasia, n (%)	22 (10.6)	13 (6.3)	0.157
Bone disease, n (%)	110 (52.9)	21 (10.1)	<0.001
Multimorbidity, n (%, 95% CI)	133 (63.9, 57–70)	90 (43.3, 37–52)	<0.001
Mean number of comorbidities			
General	2	1	<0.001
50–55 years	1.8	0.9	<0.001[a]
56–60 years	2	1.5	
61–65 years	2	1.6	
>65 years	2.2	2	

What des the p-value for multimorbidity tell us about the statistical hypotheses ?

A There is no evidence against the null hypothesis
B The null hypothesis is true
C The alternative hypothesis is true
D There is sufficient evidence to reject the null hypothesis

Which of the following best describes this result:

A The prevalence of multimorbidity was significantly higher in HIV patients: 63% (95% CI 57–70%) vs. 43% (95% CI 37–52%), $p < 0.001$
B Multimorbidity was significantly higher in HIV patients: 133 patients with HIV had multimorbidity compared to 90 patients without HIV, $p < 0.001$
C There was evidence that the average number of comorbidities was higher in patients with HIV than those without, $p<0.001$

20% more patients with HIV had multimorbidity than those without HIV. A CI for this difference was not reported in the table but given the p-value is <0.001, which of the following would be a plausible confidence interval?

A difference = 20% (95%CI -9% to 49%)
B difference = 20% (95%CI 11% to 29%)
C difference = 20% (95%CI 0% to 40%)

Answers available at end of slide set

Quiz questions and answers

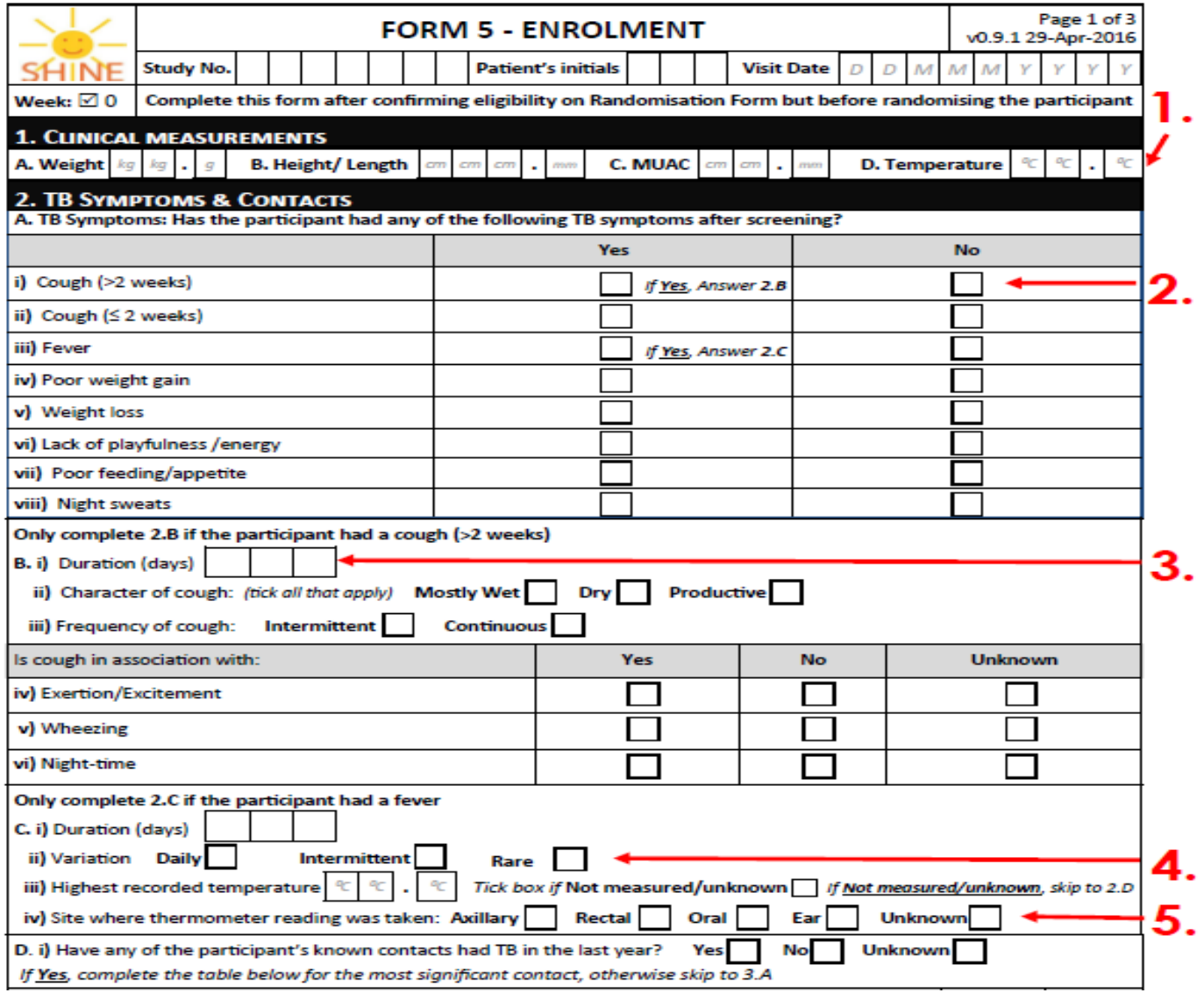
Quiz

Types of Data: Examples

What type of data are the variables highlighted on the CRF?

Types: Binary, Ordinal, Nominal, Continuous and Discrete

1: Continuous
2: Binary
3: Discrete
4: Ordinal
5: Nominal



SHINE

FORM 5 - ENROLMENT

Page 1 of 3
v0.9.1 29-Apr-2016

Study No. | Patient's initials | Visit Date (D D M M M Y Y Y Y)

Week: ☑ 0 — Complete this form after confirming eligibility on Randomisation Form but before randomising the participant

1. CLINICAL MEASUREMENTS — 1.

A. Weight (kg kg . g) | B. Height/ Length (cm cm cm . mm) | C. MUAC (cm cm . mm) | D. Temperature (°C °C . °C)

2. TB SYMPTOMS & CONTACTS

A. TB Symptoms: Has the participant had any of the following TB symptoms after screening?

	Yes	No
i) Cough (>2 weeks)	☐ *If Yes, Answer 2.B*	☐ — 2.
ii) Cough (≤ 2 weeks)	☐	☐
iii) Fever	☐ *If Yes, Answer 2.C*	☐
iv) Poor weight gain	☐	☐
v) Weight loss	☐	☐
vi) Lack of playfulness /energy	☐	☐
vii) Poor feeding/appetite	☐	☐
viii) Night sweats	☐	☐

Only complete 2.B if the participant had a cough (>2 weeks)

B. i) Duration (days) ☐☐☐ — 3.

ii) Character of cough: *(tick all that apply)* Mostly Wet ☐ Dry ☐ Productive ☐

iii) Frequency of cough: Intermittent ☐ Continuous ☐

Is cough in association with:	Yes	No	Unknown
iv) Exertion/Excitement	☐	☐	☐
v) Wheezing	☐	☐	☐
vi) Night-time	☐	☐	☐

Only complete 2.C if the participant had a fever

C. i) Duration (days) ☐☐☐

ii) Variation Daily ☐ Intermittent ☐ Rare ☐ — 4.

iii) Highest recorded temperature (°C °C . °C) *Tick box if* Not measured/unknown ☐ *if Not measured/unknown, skip to 2.D*

iv) Site where thermometer reading was taken: Axillary ☐ Rectal ☐ Oral ☐ Ear ☐ Unknown ☐ — 5.

D. i) Have any of the participant's known contacts had TB in the last year? Yes ☐ No ☐ Unknown ☐

If Yes, complete the table below for the most significant contact, otherwise skip to 3.A

Quiz

1. If we increase the sample size, what do we expect to happen to the standard error?
   - Increase / decrease / no change

2. True or false, the estimated mean is always inside the confidence interval we calculate? True

3. If we increase the sample size, what will happen to the CI?
   - Get narrower / get wider / no change

Quiz

Comorbidity is more common and occurs earlier in persons living with HIV than in HIV-uninfected matched controls, aged 50 years and older: A cross-sectional study

Rafael Aguiar Maciel[a], Helena Moreira Klück[b], Madeleine Durand[c], Eduardo Sprinz[b,*]

[a] Medical Sciences Post-Graduation Program, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[b] School of Medicine, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[c] Université de Montréal, Québec, Canada

ARTICLE INFO

Article history:
Received 13 January 2018
Received in revised form 2 February 2018
Accepted 15 February 2018
Corresponding Editor: Eskild Petersen, Aarhus, Denmark

Keywords:
HIV
AIDS
Multimorbidity
Comorbidities
Aging
Developing countries
Brazil

ABSTRACT

Objectives: At present, data are limited on the comorbidity profiles associated with aging people with HIV in the developing world, where most such people live. The aim of this study was to compare the disease burden between older HIV-positive subjects and HIV-negative matched controls in Brazil.
Methods: This was a cross-sectional analysis of the South Brazilian HIV Cohort. Individuals aged 50 years and older were enrolled at Hospital de Clínicas de Porto Alegre and matched with HIV-negative controls from the primary practice unit of the same hospital. Multimorbidity (the presence of two or more comorbid conditions) and the number of non-infectious comorbidities were compared. Poisson regression was used to identify factors associated with multimorbidity.
Results: A total of 208 HIV-positive subjects were matched to 208 HIV-negative controls. Overall, the median age was 57 years and 56% were male. The prevalence of multimorbidity was higher in HIV-positive subjects than in HIV-negative controls (63% vs. 43%, $p < 0.001$), and the median number of comorbidities was 2, compared to 1 in controls ($p < 0.001$). The duration of HIV infection ($p = 0.02$) and time on treatment in years ($p = 0.015$) were associated with greater multimorbidity in HIV-positive persons.
Conclusions: In this large cohort from the developing world, multimorbidity was found to be more common in HIV-positive subjects than in HIV-negative controls. The duration of HIV and time on antiretrovirals were associated with multimorbidity.

- Cross-sectional analysis using data from South Brazilian HIV Cohort
- Individuals aged 50+ years enrolled from a single site and matched with HIV-negative controls from the primary practice unit of the same hospital.
- **Multimorbidity** (the presence of two or more comorbid conditions) compared in those with and without HIV

What is the null hypothesis being tested in this study?

NB: p+ = proportion of people living with HIV with multimorbidity, P- = proportion of people without HIV with multimorbidity

A H0 p+ > p- (i.e. Proportion of patients with multimorbidity is higher in those with HIV than those without)
B H0 p+ = p- (i.e. Proportion of patients with multimorbidity is the same in those with HIV and those without)
C H0 p+ =/= p- (i.e. Proportion of patients with multimorbidity is not the same in those with HIV and those without)

Quiz

Table 3
Comparison of the burden of comorbidities between HIV-positive patients and non-HIV controls.

	HIV ($n=208$)	Non-HIV ($n=208$)	p-Value
Cardiovascular disease, n (%)	20 (9.6)	20 (12.5)	0.435
Kidney disease, n (%)	35 (16.8)	14 (6.7)	0.002
Hepatic disease, n (%)	53 (25.5)	14 (6.7)	<0.001
Diabetes, n (%)	47 (22.6)	59 (28.4)	0.216
Hypertension, n (%)	129 (62.0)	145 (69.7)	0.121
Neoplasia, n (%)	22 (10.6)	13 (6.3)	0.157
Bone disease, n (%)	110 (52.9)	21 (10.1)	<0.001
Multimorbidity, n (%, 95% CI)	133 (63.9, 57–70)	90 (43.3, 37–52)	<0.001
Mean number of comorbidities			
General	2	1	<0.001
50–55 years	1.8	0.9	<0.001[a]
56–60 years	2	1.5	
61–65 years	2	1.6	
>65 years	2.2	2	

What des the p-value for multimorbidity tell us about the statistical hypotheses ?

A There is no evidence against the null hypothesis
B The null hypothesis is true
C The alternative hypothesis is true
D There is sufficient evidence to reject the null hypothesis

Which of the following best describes this result:

A The prevalence of multimorbidity was significantly higher in HIV patients: 63% (95% CI 57–70%) vs. 43% (95% CI 37–52%), $p < 0.001$
B Multimorbidity was significantly higher in HIV patients: 133 patients with HIV had multimorbidity compared to 90 patients without HIV, $p < 0.001$
C There was evidence that the average number of comorbidities was higher in patients with HIV than those without, $p<0.001$

20% more patients with HIV had multimorbidity than those without HIV. A CI for this difference was not reported in the table but given the p-value is <0.001, which of the following would be a plausible confidence interval?

A difference = 20% (95%CI -9% to 49%)
B difference = 20% (95%CI 11% to 29%)
C difference = 20% (95%CI 0% to 40%)