Introduction to Medical Statistics 2026

Handouts of sessions 1-3, 9 and 10, March 23-27, 2026

Ronald Geskus and the biostatistics crew; rgeskus@oucru.org.
Oxford University Clinical Research Unit
Hospital for Tropical Diseases, Ho chi Minh City, Viet Nam

.1

Contents of handouts. The handouts contain materials for the first three and the last two sessions. The slides as used in the clasroom lecturs are in separate pdf files, but they are included in the handouts as well. The other sessions have the slides in powerpoint format.

Contents

.2

Part I
Data; Descriptive Analysis

Medical research starts with a study question. We collect data in order to answer that question. We will explain the structure and characteristics of a data set. We emphasize the importance of having your data in tidy format. A tidy data set consists of variables that each measure a characteristic. Variables can be of different types. We explain how best to describe and summarize the distribution of a variable, which depends on its type.

1 Data

1.1 Data: structure

Most data come in a rectangular format, and contain observations on variables. Preferably, data is in *tidy* format.

Some definitions

Data set: observations from a study, typically in rectangular format
Observation: set of measurements on a unit (patients, animals, farms) at one specific time (rows)
Variable: measurable characteristic; its value typically varies over observations (columns)

.3

Example

Trial in patients with dengue shock

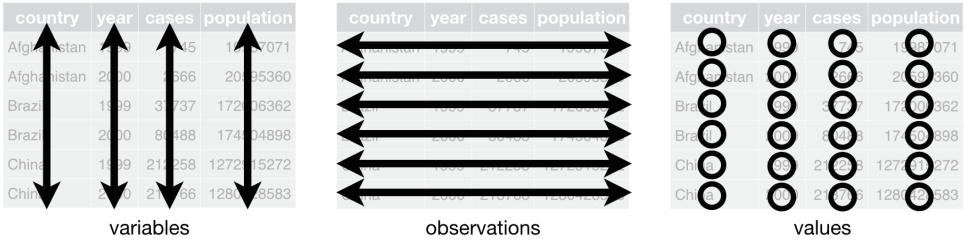
SUBJID	fluid	Age	Sex	Hct	PLT	hospdays	Reshock
01-0007	Colloid	5	F	49.0	24.0	7	No
01-0008	Blood	11	F	54.0	27.0	8	No
01-0009	Colloid	8	M	43.0	47.0	6	No
01-0010	Crystalloid	15	F	45.5	18.9	7	No
01-0011	Crystalloid	13	M	49.0	24.0	4	No
01-0012	Colloid	8	M	40.0	34.0	3	No

Rows: observations, one row per patient (each with different SUBJID)
Columns: variables
Cells: values

Inconsistent naming of variables!

.4

Data structure



.5

Tidy data

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

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

.6

We give an example of tidy and non-tidy (messy) data representation. It presents WHO data on the number of TB cases per country per year. In the tidy format, each column represents one characteristic, each row is one observation (specific year in specific country) and each cell contains a single number. In the first of the two messy formats, the `count` column contains information on number of cases and population size. In the second, the `rate` column combines both number of cases and population size into a single value.

Tidy and messy data

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

Compare with messy format:

```
country      year type           count      |      country      year rate
Afghanistan  1999 cases            745      |      Afghanistan  1999 745/19987071
Afghanistan  1999 population  19987071      |      Afghanistan  2000 2666/20595360
Afghanistan  2000 cases           2666      |      Brazil       1999 37737/172006362
Afghanistan  2000 population  20595360      |      Brazil       2000 80488/174504898
Brazil       1999 cases           37737      |      China        1999 212258/1272915272
Brazil       1999 population 172006362      |      China        2000 213766/1280428583
  with 6 more rows                          |
```

.7

Tidy data: advantages

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

.8

Dataset; further suggestions

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

.9

The dataset in slide 4 doesn't have a consistent use of capitals. Do you think that the names are clear enough to understand what the variables represent?

1.2 Variables: type

Variables can be of different type. The type of the variable has a large impact on the appropriate type of analysis. The main distinction is between numeric and categorical variables.

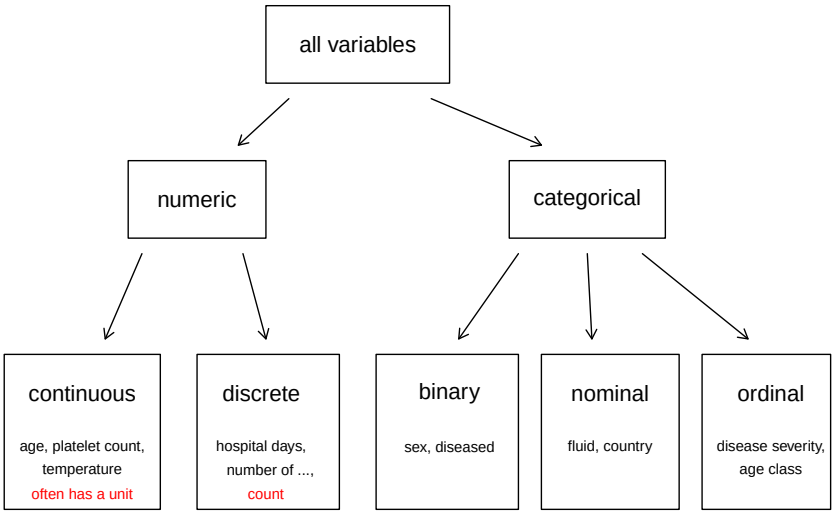
Types of variables; example

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

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

.10

Types of variables



.11

Categorical variables are often coded as numerical. The example in slide 13 shows that this can give serious errors in the analysis.

Missing data should be prevented as much as possible. Missingness may be selective and leaving observations with missing data out of the analysis may bias results. Unfortunately, missing data are often inevitable. The amount of missing observations per variable of interest should be clearly reported. Also, missing values should be coded properly. In the past, missings were sometimes coded as a number that clearly differs from the rest of the data, e.g. 999. This is not recommended.

Coding variable values

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

.12

A paper retracted

https://jamanetwork.com/journals/jama/fullarticle/2752474

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

.13

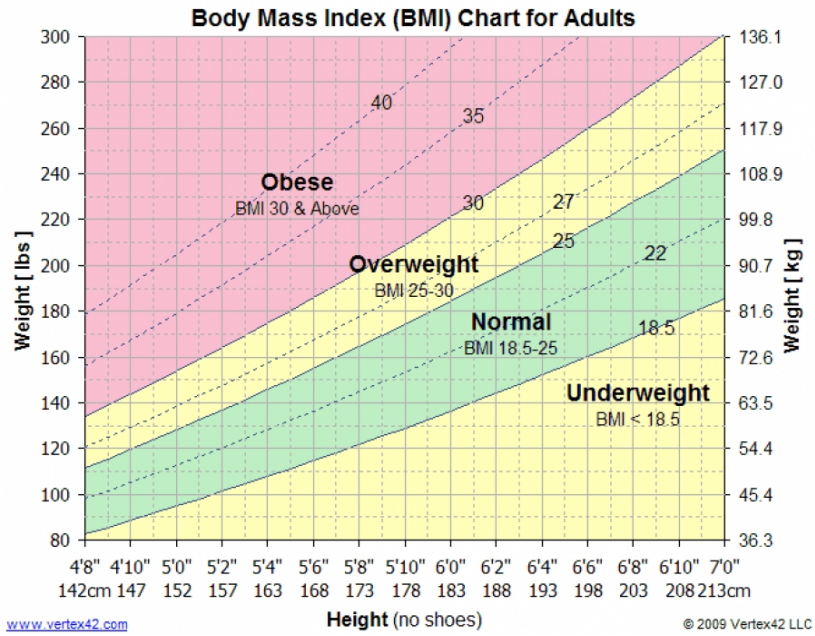
Derived variables

Variables often transformed to simplify display or analysis:

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

.14

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



.15

Exposure and outcome variables

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

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

.16

Spot the exposure / outcome?

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

.17

Spot the exposure / outcome?

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

.18

2 Variables: numerical summaries

Most papers contain a numerical summary of the important "baseline" variables, usually as the first table[1].

[1] There's even an R package to create baseline tables called `tableone`.

2.1 Categorical variables

Categorical variables with few levels are best summarized via the number and percentage of observations with per level. In R it may be useful to transform such categorical variables into a `factor` class, which gives control on the ordering of the levels. Categorical variables with many levels often only serve to discriminate between rows and don't need to be summarized (e.g. `SUBJID` in slide 4).

Presentation

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

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

- Category values often called *levels*

.19

2.2 Numerical variables

Numerical variables can usually take many possible values. Reporting every different observed value is not very insightful. The most important characteristics are the center of the values ("location") and the amount of variation in the values ("spread").

Location and spread

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

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

.20

Location: arithmetic mean

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

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

- Usually just called "mean"

.21

Location: median

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

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

.22

Location: mean versus median

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

.23

In general the median is recommended for skewed variables. However, for a discrete numeric variable with only few values, the mean may be more informative. Here is an example:

number of reshocks	0	1	2	3
group A	60	40	0	0
group B	60	15	0	25

The median is zero in both groups, but the distribution is clearly different. The means are for group A 0.4; for group B 0.9.

Mean versus median, clinical relevance

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

.24

Spread (dispersion, variation)

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

.25

Measure spread I: variance/standard deviation

- Variance: square each deviation from mean, and average

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

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

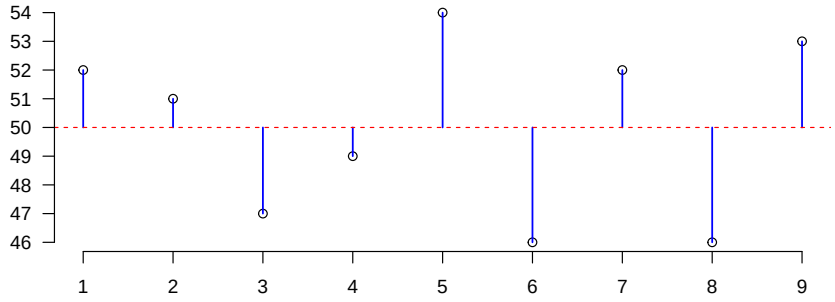
- Standard deviation (sd): square root of variance

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

.26

Deviation (from mean)

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



- Positive and negative deviations from mean always cancel out

.27

Better measure: "standard deviation"

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

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

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

.28

Measure spread II: range, interquartile range

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

.29

The 1st and 3rd quartile in itself are measures of location, but in combination they give information on spread.

Exercise: medians and IQRs

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

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

.30

Measures of spread: comparison

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

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

    $\bar{x} - 2 \times$ sd *gives negative value*: 5.1-6.4=-1.3
  - sensitive to outliers
- Quartiles or IQR
  - can always be used to quantify spread
- Minimum, maximum, range
  - not best measure; range increases with sample size
  - only used as additional information

.31

3 Variables: graphical summaries

For a single categorical variable, there is often little added value above the numerical summary. But when we want to compare categorical variables by subgroup, then a graphical representation can be useful to detetct patterns (see class on Exploratory Data Analysis). For a single numerical variable, graphs often give additional information over the numerical summary.

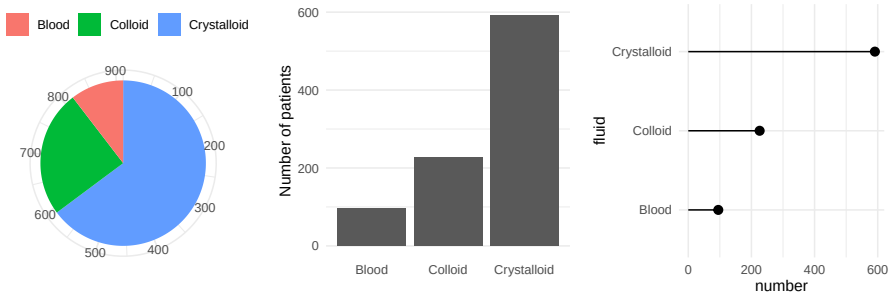
Main graph types for a single variable

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

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

.32

Pie, bar and dotplot



.33

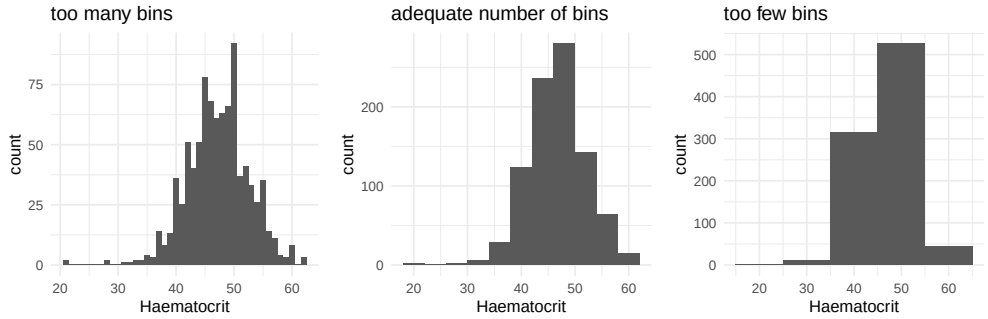
Histogram

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

.34

We show the impact of the choice of number of bins:

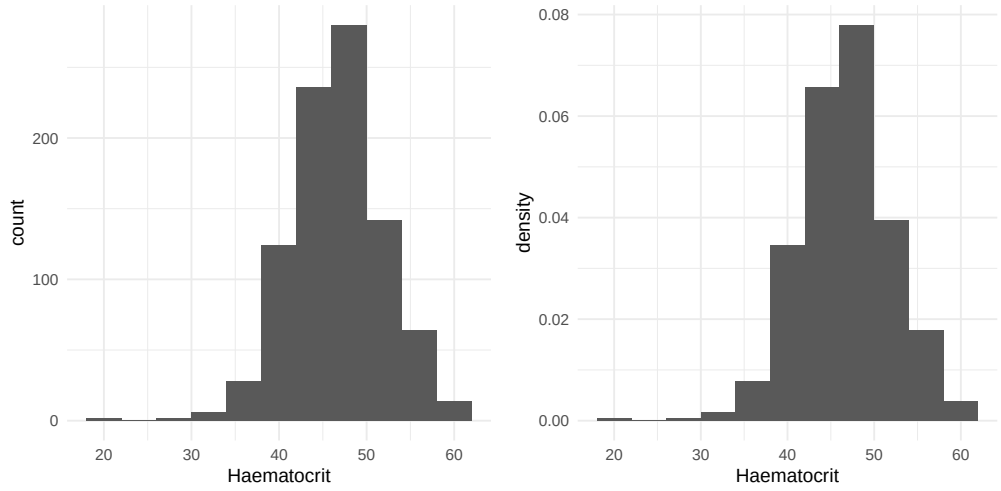
Histograms: different chosen binwidths



.35

In the next slide, we contrast count with the standardized value:

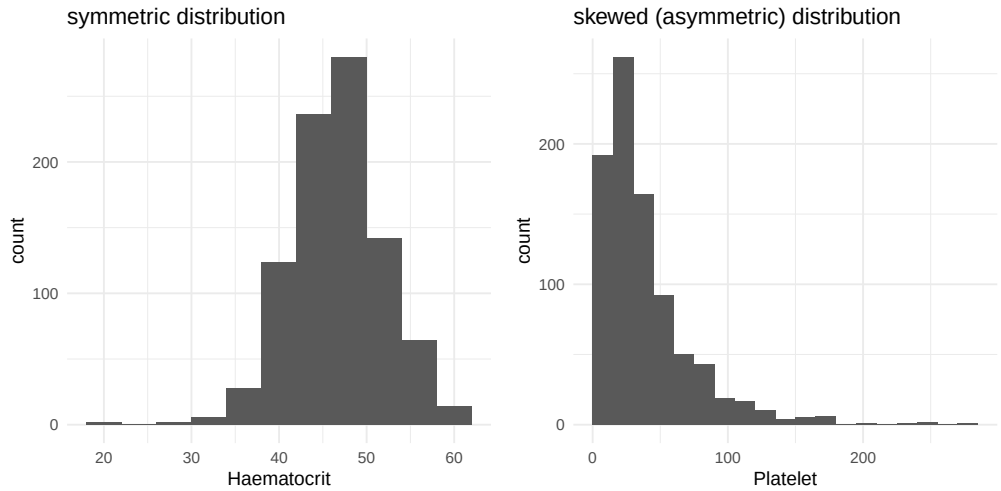
Histogram: count versus frequency/area



.36

The histogram allows us to see skewness of the distribution of the data:

Histogram: distributions with different shape

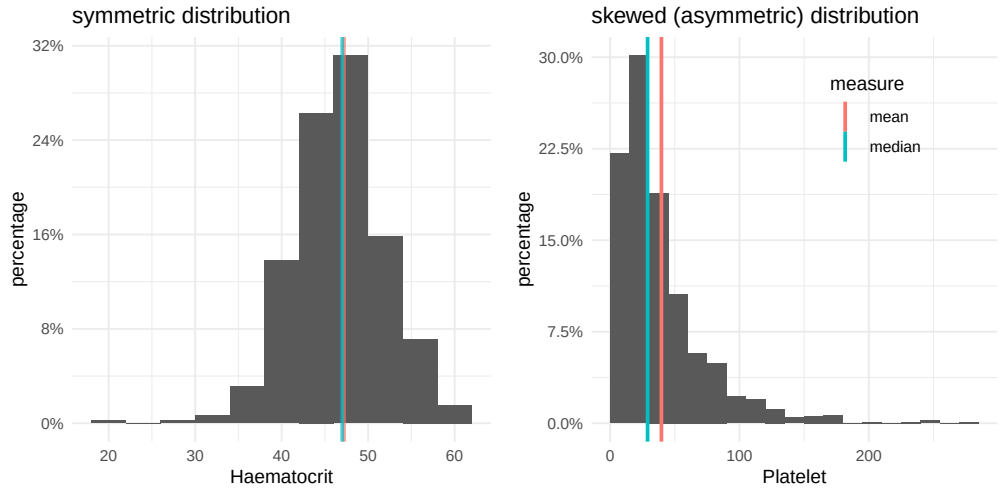


.37

As expected, the mean is larger than the mean for platelet count, which has a distribution that is skewed to the right. Note that here we changed the type of

standardisation. Instead of the area, now the height of the bar represents relative frequency of that grouped value.

Histogram: relation with location measures

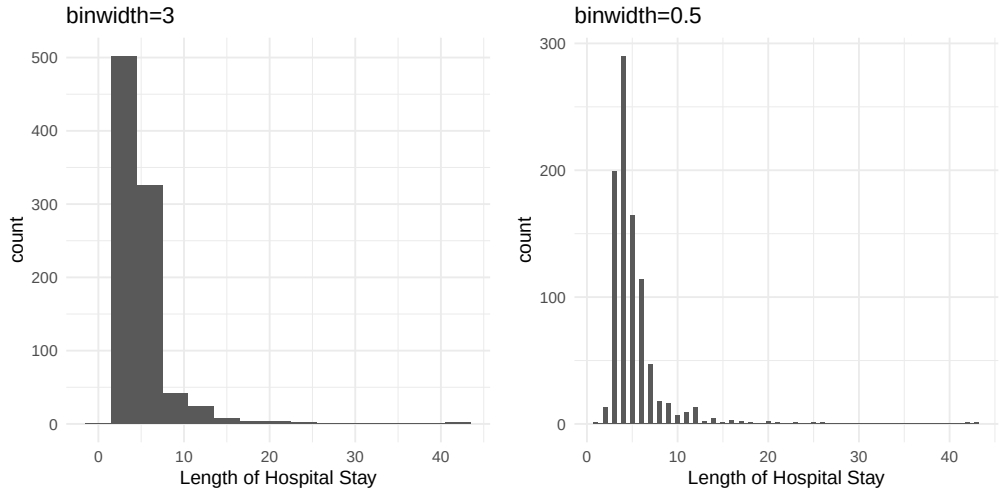


.38

With discrete numerical values it is sometimes more informative to have one bar for each value:

Histogram: discrete numeric variable

Make each value a separate bar

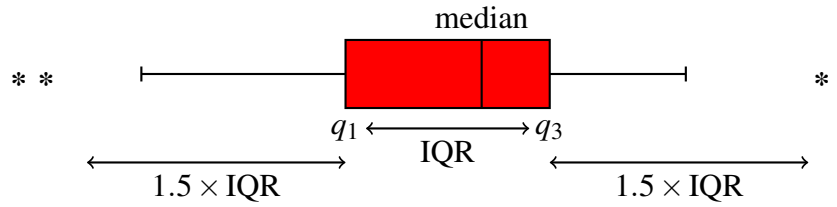


.39

There are some variations to the histogram, like the frequency polygon and the density plot. They allow for a more concise display of information, which is useful if we want to show and compare the distribution of a variable by subgroup (see slides 73 and 75).

Another very popular graph to describe the distribution of a numerical variable is the boxplot. It gives a more concise representation of the values, at the cost of loss of information.

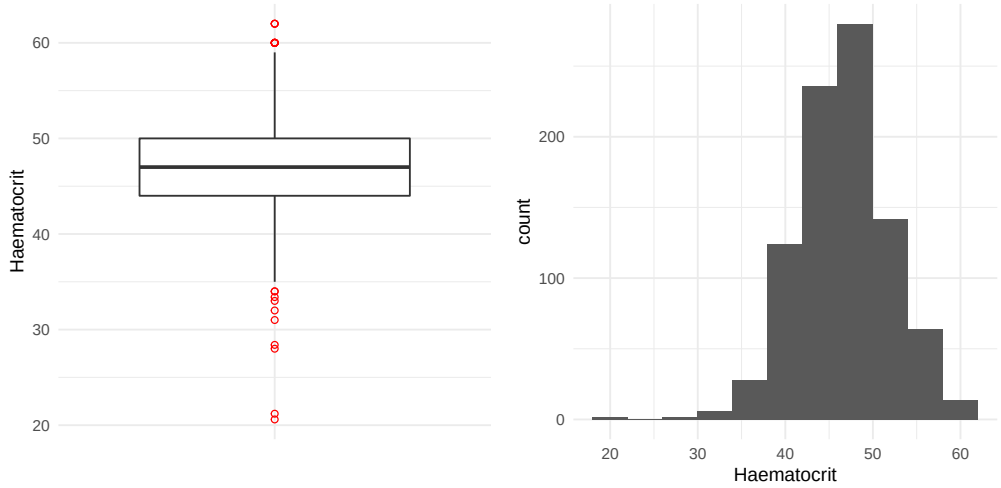
Boxplot



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

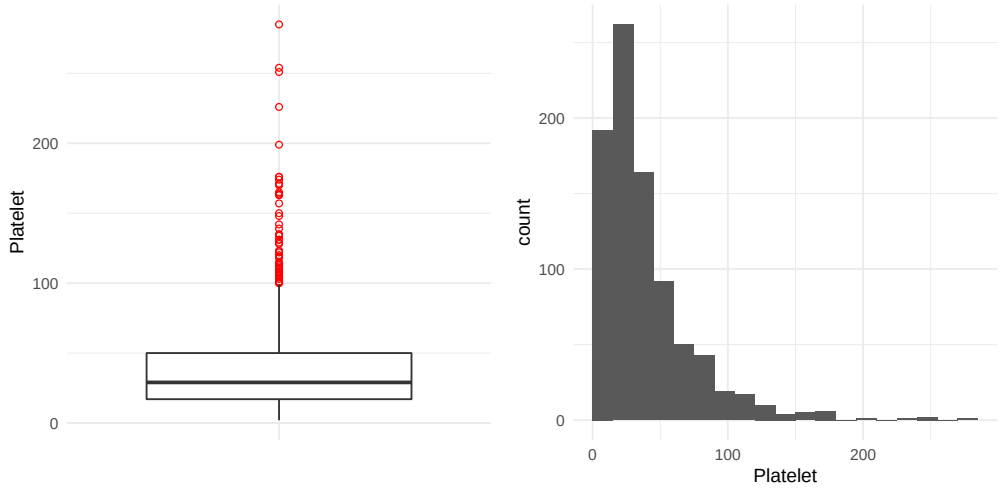
.40

Boxplot versus histogram: symmetric distribution



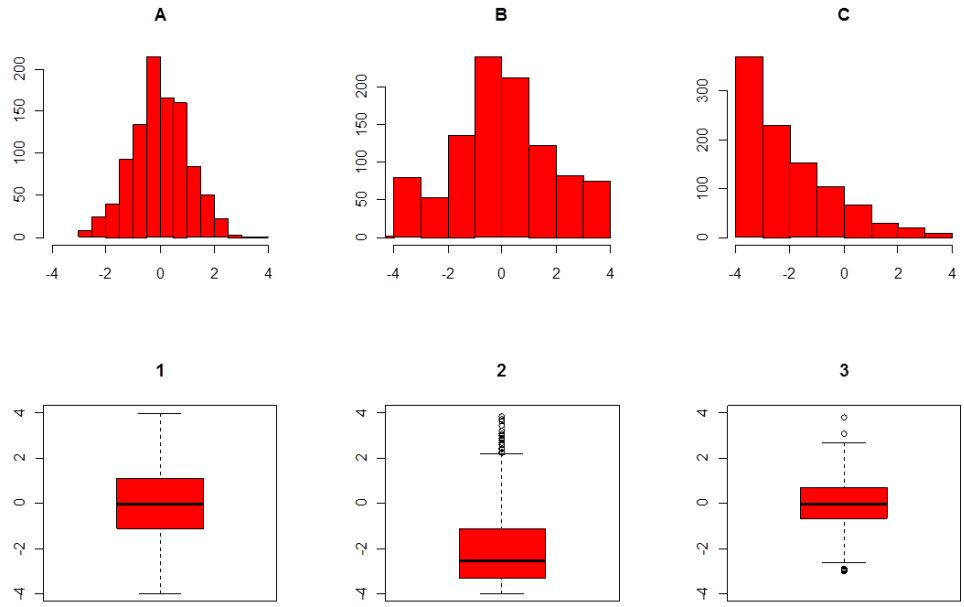
.41

Boxplot versus histogram: asymmetric distribution



.42

Quiz: match histogram and boxplots

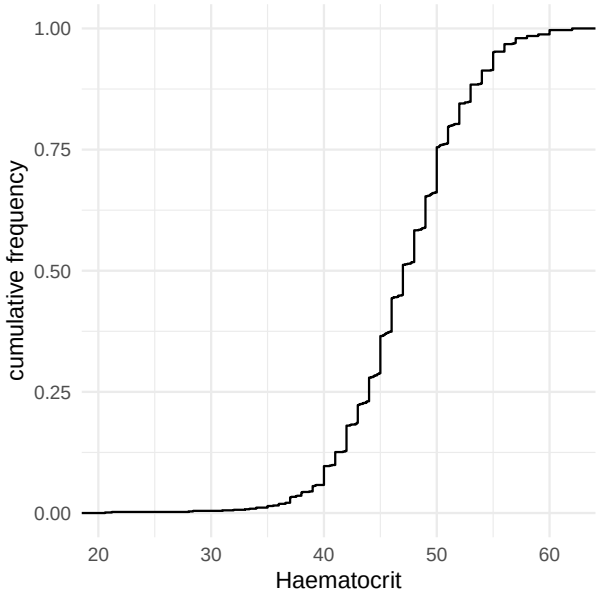


.43

There are variations of the boxplot such as the violin plot and the pirate plot that provide more information (see e.g. slide 78). It is often useful to add the individual measurements as well, especially if the number of observations is fairly low.

The empirical cumulative distribution function plots the fraction of values $\leq x$ for each value x of the variable, yielding a curve going up from 0 to 1 as x increases in value. Because it gives a full description of the values of a continuous variable, it may show some patterns that would otherwise go unnoticed. For example, the ECDF in the next slide shows that most of the values of haematocrit are integers (the large jumps), but some decimal values occur as well (the small jumps in-between).

Empirical Cumulative Distribution Function (ECDF)



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

.44

However, a histogram gives more detailed information on the distribution and is more appropriate to determine skewness. However, there are variations of the box plot that show more information, such as the violin plot and raincloud plot (see the practical).

3.1 Transformation of variables

If a continuous variable is highly skewed, it is often better to transform the values. The most frequent transformation is the logarithm. If the distribution of a variable is skewed to the right, log-transformed values are often much more symmetric (but not always!).

Variable transformation: logarithm

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

.45

An issue may be the interpretation of the transformed values (slide 47). If the log-transformed values are hard to interpret, we can back-transform the mean of the log-transformed values. This doesn't give the arithmetic mean, but the geometric mean. In a graph, we can always use the original values as labels (slide 48).

Geometric mean

- Mean of log-transformed values, e.g. 10-log

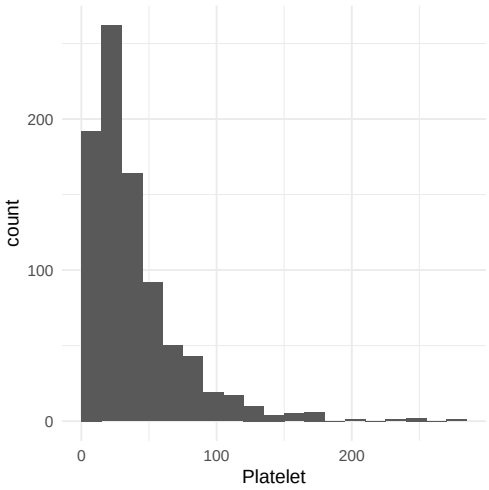
$$m = \frac{\sum_{i=1}^{n} \log_{10}(x_i)}{n}$$

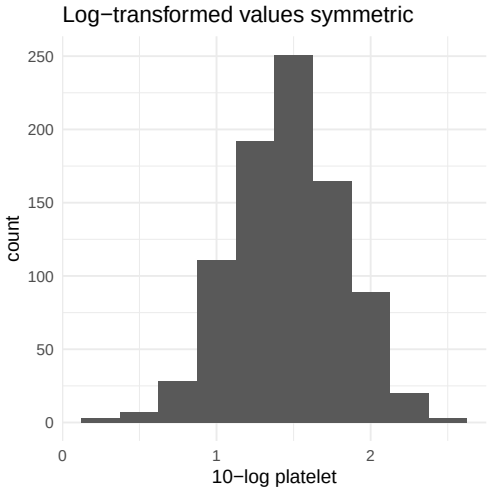
- May be hard to interpret $\to$ transform back to original scale by taking 10^m
- This gives the *geometric mean*, not the arithmetic mean

$$(x_1 \times x_2 \times \ldots \times x_n)^{(1/n)}$$

.46

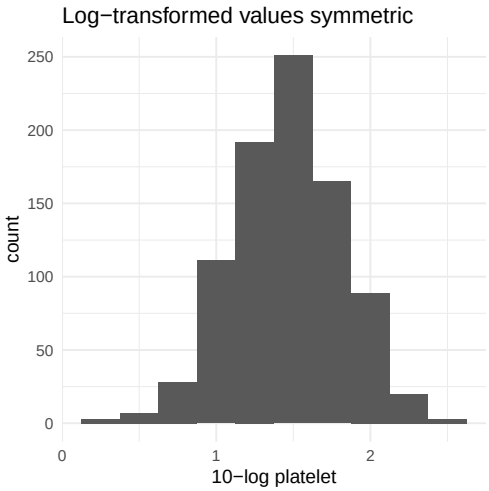
Data skewed to the right (long tail)

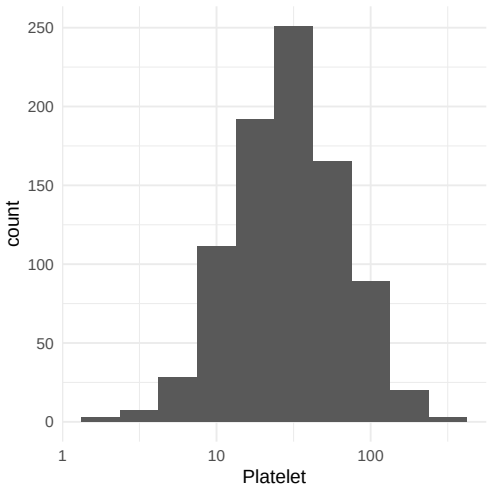




.47

May use original values in labels





.48

Summary descriptive analysis

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

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

.49

References

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

.50

4 Practicals; Introduction to R

How to make the exercises

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

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

.51

The R program

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

.52

R: functions

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

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

.53

R: extensions

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

.54

R: objects

- Everything is an *object* (≈ files in operating system):
  Most important ones: data sets; functions
- You decide about the name of objects you create
  RStudio: overview of objects created in Environment tab
- Output of a function can be assigned to an object or to an element of an object:
  `DataColloid <- subset(myData, fluid=="Colloid")`
- We can also assign to a column in data set
  Example: transform values 1 and 2 into categorical:
  `cmTbm$sex <- factor(cmTbm$sex, labels=c("M","F"))`
- Objects created can be used for further analyses

.55

Writing code: use R script or R Markdown file

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

.56

Further information

Ask LLM, or some suggestions:

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

.57

Part II
Exploratory Data Analysis

5 Exploratory Data Analysis (EDA)

In this class we focus on summarizing and visualizing relationships between variables. It can be an important part of your scientific investigations, leading to new insights on patterns in your data.

Wikipedia (March 2026)

Exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily *EDA is for seeing what the data can tell beyond the formal modeling* and thereby contrasts with traditional hypothesis testing, in which a model is supposed to be selected before the data is seen. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

.58

Exploratory Data Analysis (EDA)

1. Descriptive analysis (IDA): inspect and summarize data characteristics
   - variables: what type, which distribution Is there a need to transform numeric variable?
   - find errors, check for peculiarities, such as missing values and outliers
2. EDA: find patterns and relationships between variables
   - *numerical or graphical summary of variables by subgroup*
3. Purpose EDA
   - *suggest hypotheses for further research*
   - *suggest appropriate modeling approach*
     example: WHO European Health Report data 2007
4. Often using graphics

   > The greatest value of a picture is when it forces us to notice what we never expected to see (John W. Tukey)

.59

Most papers contain a table that summarizes the baseline variables that play a role in answering the study question, often split up by subgroups. Slide 60 shows part of the baseline table of the SARS-CoV-2 Recovery Trial. Note that age is summarized in two different ways, as mean and standard deviation, and in age groups. For the categorical variables, numbers and columnwise percentages are given. The discrete variables "number of days" are summarized by median and 25- and 75 percentiles (which they call IQR). For days since hospitalization, the mean value might be more informative than the median.

Baseline table

- Most papers include a numerical description of all variables that are relevant for a study (often in "Table 1")
- Such summaries often reported by subgroup
  Example: Recovery Trial Dexamethasone in Covid-19 patients

Table 1. Characteristics of the Patients at Baseline, According to Treatment Assignment and Level of Respiratory Support.*

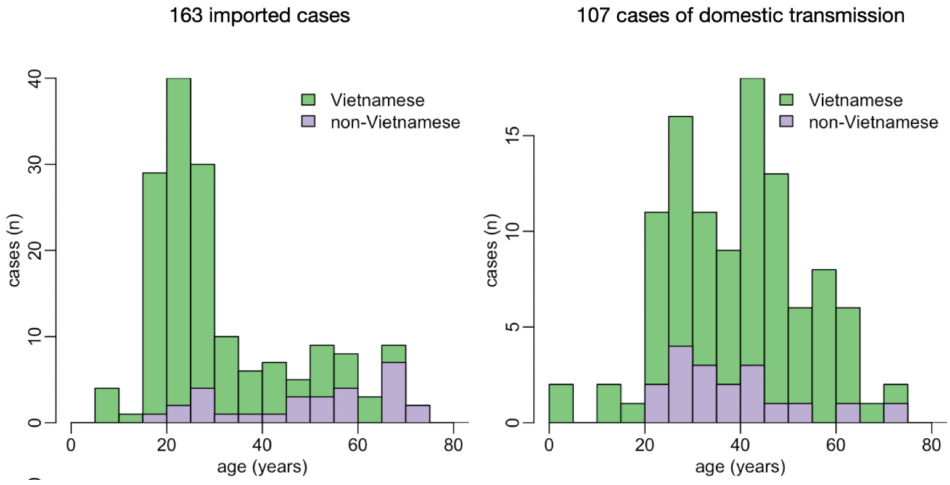
Characteristic	Treatment Assignment		Respiratory Support Received at Randomization		
	Dexamethasone (N=2104)	Usual Care (N=4321)	No Receipt of Oxygen (N=1535)	Oxygen Only (N=3883)	Invasive Mechanical Ventilation (N=1007)
Age†					
Mean — yr	66.9±15.4	65.8±15.8	69.4±17.5	66.7±15.3	59.1±11.4
Distribution — no. (%)					
<70 yr	1141 (54)	2504 (58)	659 (43)	2148 (55)	838 (83)
70 to 79 yr	469 (22)	859 (20)	338 (22)	837 (22)	153 (15)
≥80 yr	494 (23)	958 (22)	538 (35)	898 (23)	16 (2)
Sex — no. (%)					
Male	1338 (64)	2749 (64)	891 (58)	2462 (63)	734 (73)
Female‡	766 (36)	1572 (36)	644 (42)	1421 (37)	273 (27)
Median no. of days since symptom onset (IQR)§	8 (5–13)	9 (5–13)	6 (3–10)	9 (5–12)	13 (8–18)
Median no. of days since hospitalization (IQR)	2 (1–5)	2 (1–5)	2 (1–6)	2 (1–4)	5 (3–9)

.60

Slide 61 shows a graphical summary from the paper on the first 100 days of the SARS-CoV-2 control in Vietnam. It gives a histogram of the number of cases by 5-year age group, stacked by nationality, with two separate panels/facets for location of transmission. We can see that most infected individuals had Vietnamese nationality, and the imported cases tended to be younger.

Stacked barchart (numbers; 2 facets/panels)

SARS-CoV-2 infections in Vietnam (January-April 2020)



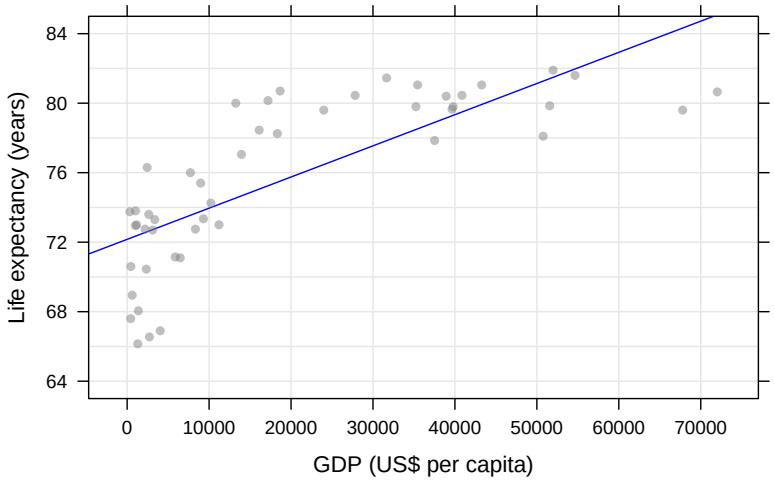
.61

EDA also serves to suggest an appropriate modeling approach. In slide 62 we show data on Gross Domestic Product and Life Expectancy in 2007 (WHO European Health Report 2009[2]). It is immediately clear that the relation is not linear, even though the WHO fitted a linear curve through the data.

[2]Figure 2.17 in https://iris.who.int/handle/10665/107272

WHO European Health Report data 2007

Linear trend not correct, more subtle pattern present



.62

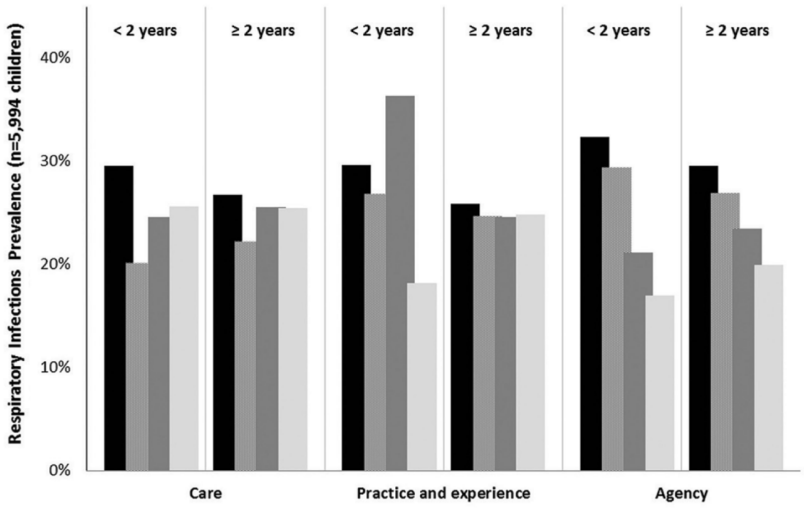
6 Graphs: two or more categorical variables

In Section 3 we noted that visualizing a single categorical variable often has little added value compared to a numerical summary. This may be different if we want to explore relationships between two or more categorical variables. The bar chart is the most frequently used visualisation, either in the form of a stacked barchart (all levels stacked) or a dodged barchart (with the levels next to each other).

Slide 63 shows an example of a dodged barchart, from a paper on diarrhea and acute respiratory tract infections in Indonesian children. It shows the prevalence of acute respiratory tract infections in children in Indonesia by age and three maternal factors. Note that each of the maternal factors are ordered, from low (black) to high (light grey). High maternal agency lower the risk of respiratory infection, but for the rest there are no clear patterns.

Dodged barchart (percentages)

Prevalence respiratory infections by age group and 3 maternal factors (4 levels, low to high)



.63

However, often there are better ways than a barchart to visualize such data. An alternative is the dotplot, which uses less ink to display the same amount of information and allow for more concise display of information. It has the added advantage that it can overlay information from subgroups. The next slides show an example of mortality (number of deaths per 1000 inhabitants) in 1940 in the US state of Virginia by age group, gender and environment[3]. A grouped dotplot may be the most concise and informative display of information.

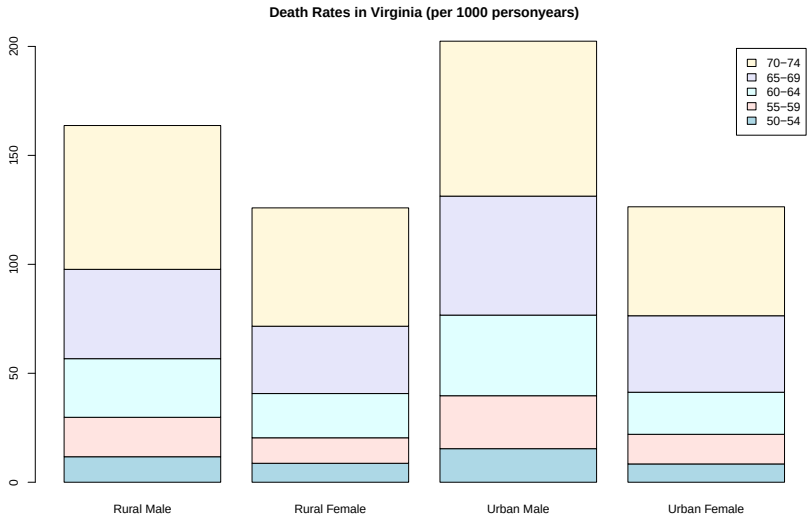
Mortality per 1000 inhabitants in Virginia in 1940

	Rural Male	Rural Female	Urban Male	Urban Female
50-54	11.70	8.70	15.40	8.40
55-59	18.10	11.70	24.30	13.60
60-64	26.90	20.30	37.00	19.30
65-69	41.00	30.90	54.60	35.10
70-74	66.00	54.30	71.10	50.00

.64

We first plot the barchart with numbers by age group stacked. It makes the actual mortality numbers hard by age group to read.

Stacked barchart

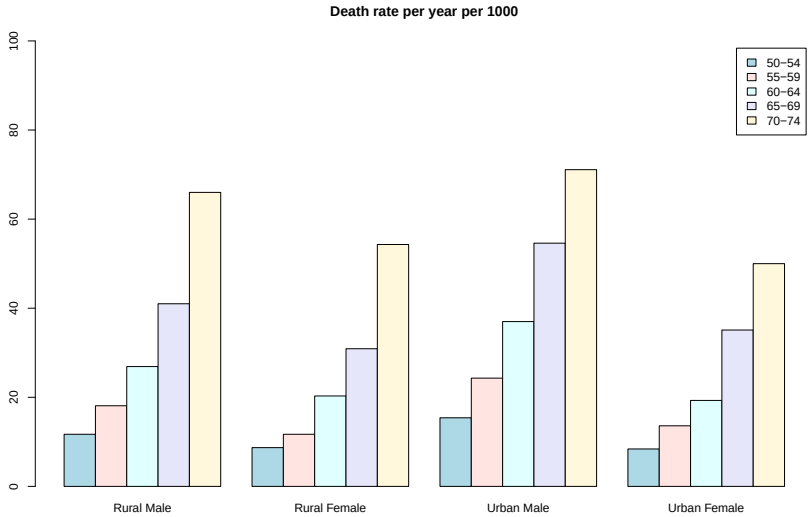


.65

We can use the dodged barchart instead. It more clearly shows the increasing mortality with increase in age.

[3]It is included in R as the example data set `VADeaths`. Note that the data is not in tidy format, and we need to transform the data before we can make the graphs.

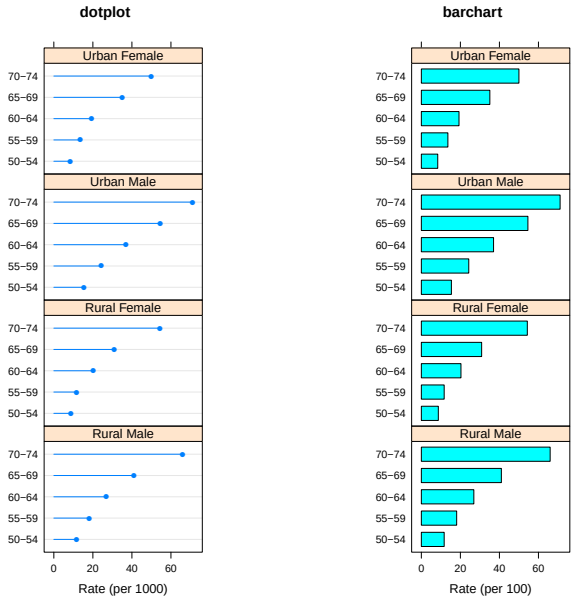
Dodged barchart



.66

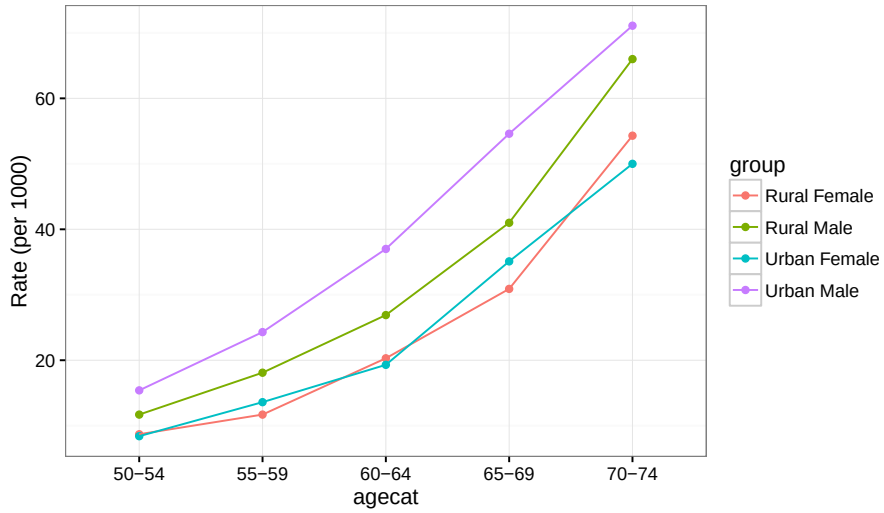
In the next slide, we show the dotplot as alternative (both with vertical rather than horizontal orientation). Here, the dotplot does not have much added value. However, in slide 68 we show an alternative dotplot with dots connected that belong to the same subgroup. It contains all the information in a single insightful figure. For example, we immediately see that males have higher mortality in all age groups, and especially so in the urban group.

Dotplot versus barchart with facets



.67

Grouped dotplot



.68

If we have several binary (yes-no) variables, the Venn diagram has been to most frequently used plot type. It has the disadvantage that it becomes very noisy with a lot of variables. An more insightful alterative type that recently gained much attention is the upset plot. We show an example.

Many categorical variables

Example

Three different covid-19 symptoms; may occur together.
Eight different combinations

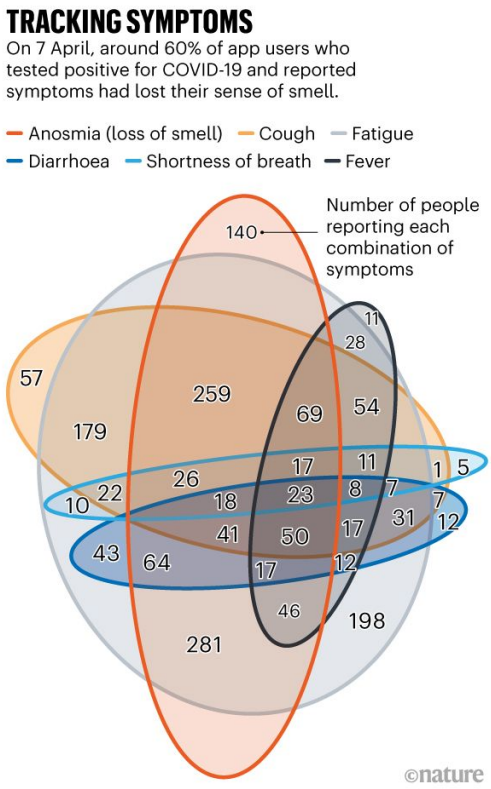
Anosmia	Fatigue	Cough	Nr of symptoms
No	No	No	0
Yes	No	No	1
No	Yes	No	1
No	No	Yes	1
Yes	Yes	No	2
Yes	No	Yes	2
No	Yes	Yes	2
Yes	Yes	Yes	3

With 6 symptoms there are 64 possible combinations

.69

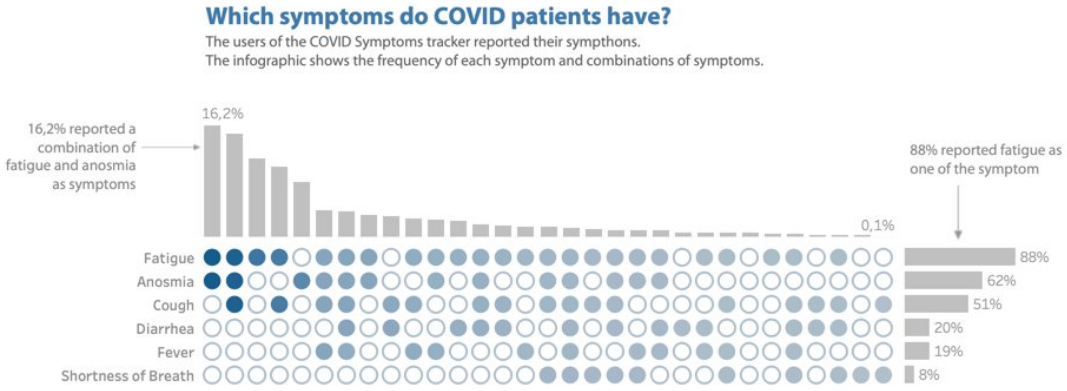
The standard Venn diagram becomes very hard to interpret with 6 different symptom combinations. The upset plot is a novel way of reporting that gives more insights.

Covid-19 symptoms: Venn diagram



.70

Covid-19 symptoms: upset plot



.71

7 Numeric variable by groups

If we want to graphically summarize one numeric variable by one or more categorical variables, we can mostly use the same methods as described in Section 3. Barcharts (called dynamite plots in this setting) are strongly discouraged (see e.g. the paper "Show the data, don't conceal them").

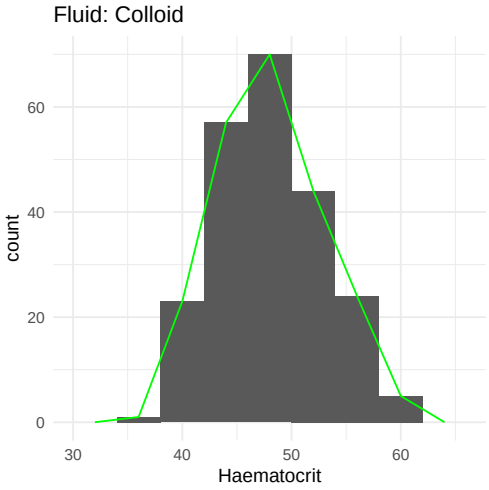
Single continuous variable by subgroup

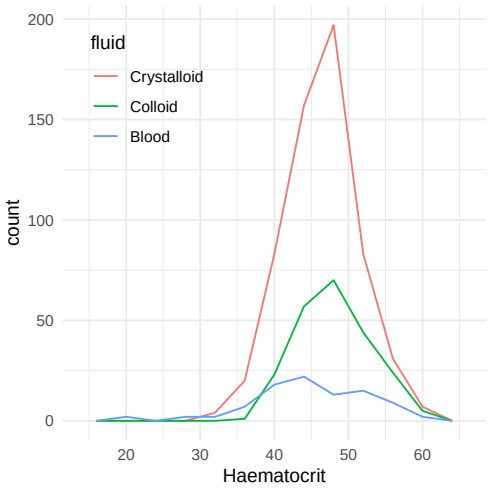
- Histogram, *frequency polygon, density*, ridgeplot
- Boxplot, *violin plot, raincloud plot*
  Don't use dynamite plots
- *Cumulative frequency* ("empirical cumulative distribution function", `ecdf`)

.72

A histogram is not suitable for subgroup comparisons, because we cannot overlap histograms without losing information. However, there are histogram-style plots that can be used instead. A frequency polygon conveys exactly the same information as a histogram, but uses less ink because it does not have the solid bars. When comparing groups, think whether you want to report the absolute counts or the standardized values in which the values are divided by the total number per subgroup. The latter may give a better impression of the distribution of values by subgroup (see slide 74). We can also use a smoothed version of the frequency polygon, which is sometimes more insightful (slide 75).

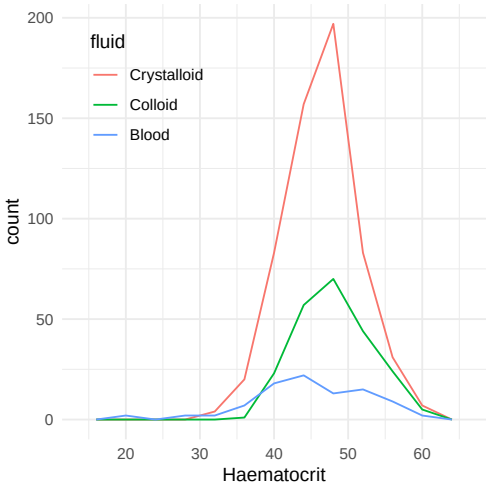
Histogram style for groups: "frequency polygon"

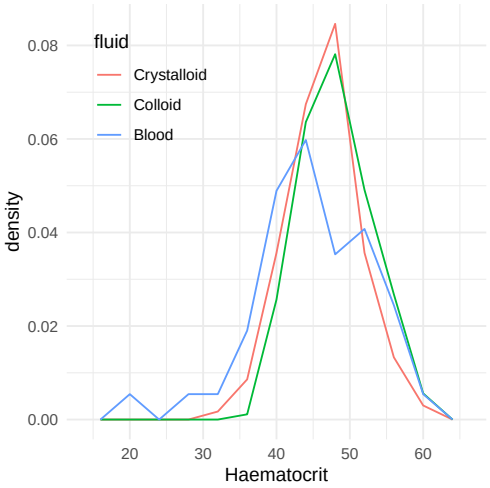




.73

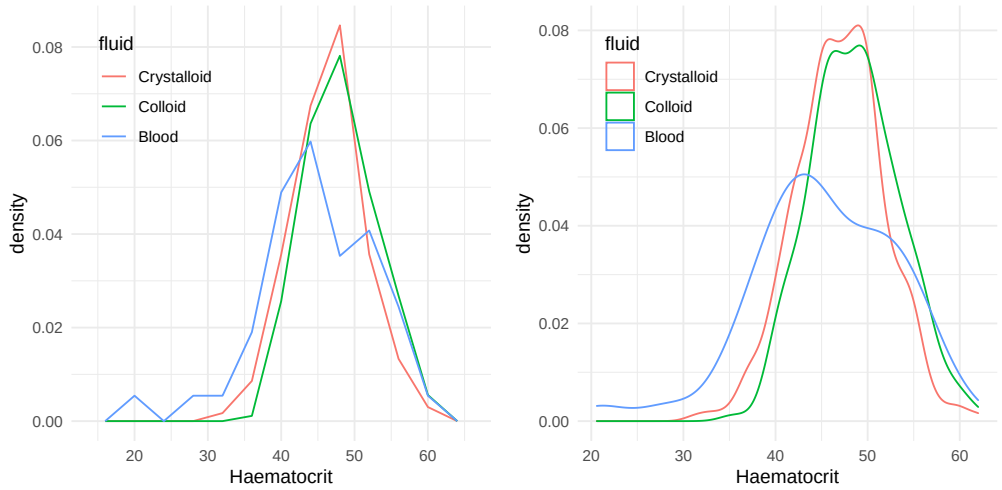
Comparison easier with standardized frequency polygon





.74

Smooth standardized frequency polygon: density plot

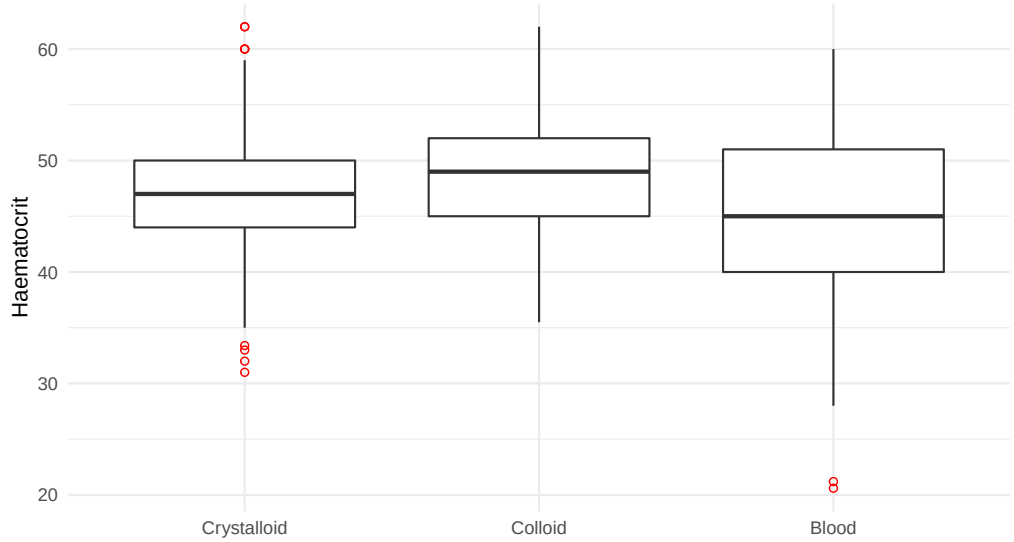


.75

Instead of overlapping densities, one can also use a ridgeplot. A ridgeplot is very infomative if the categorical variable has many levels.

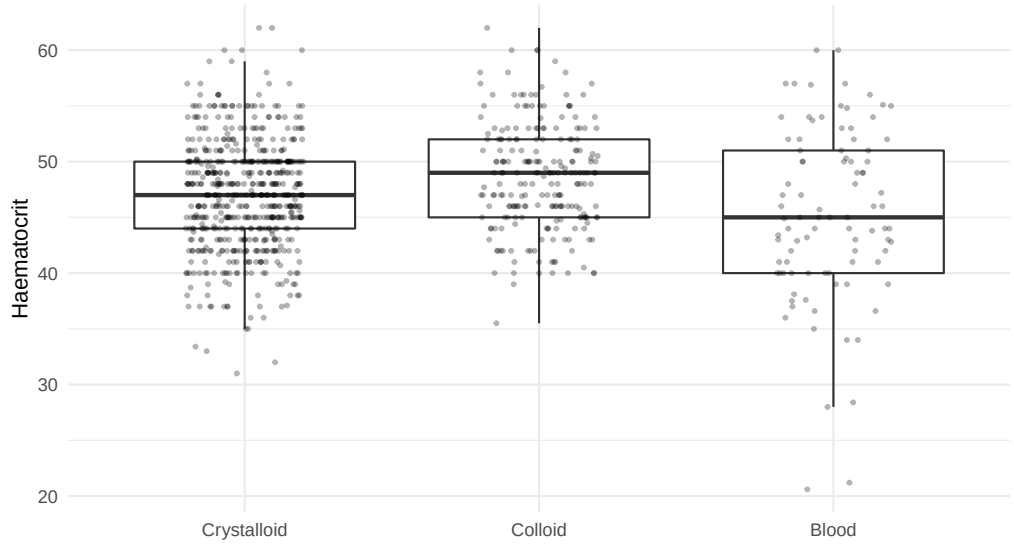
An alternative to histogram style plots is the boxplot. It allows for comparison of subgroups by plotting the boxes next to each other. However, the boxplot itself leaves out a lot of detail. Usually it is easy to add the individual values to the boxplot. Also, the are alternatives that combine the boxplot and the smoothed histogram ideas as in a violin plot or a raincloud plot.

Boxplot for multiple groups



.76

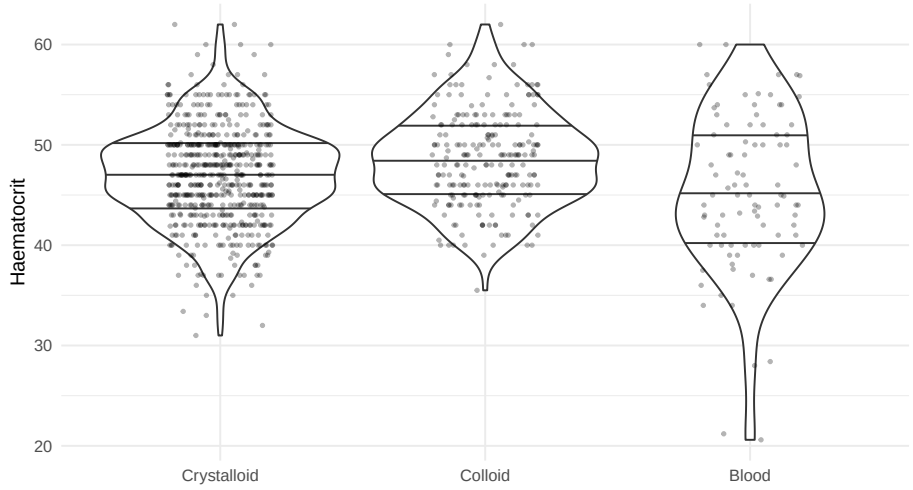
Boxplot, individual values added



.77

The violin plot combines a boxplot with the density plot.

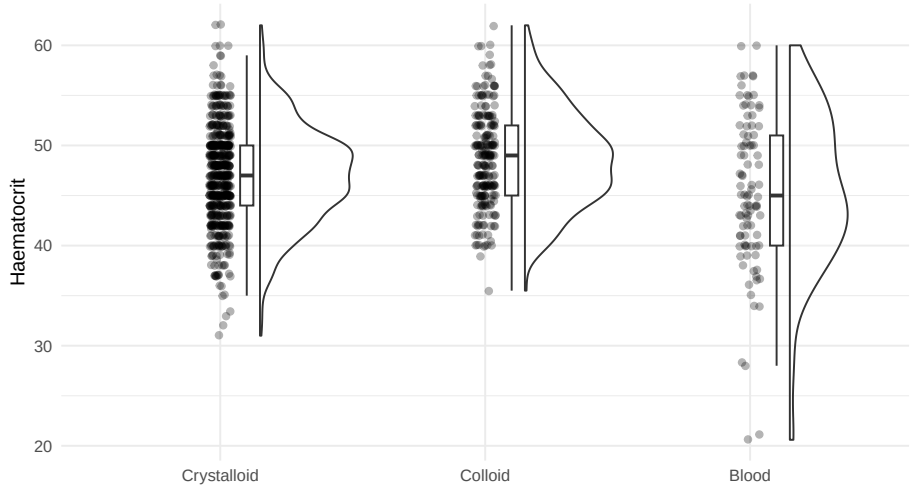
Violin plot



.78

The raincloud plot has recently been suggested as a combination of boxplot, violin plot and individual points.

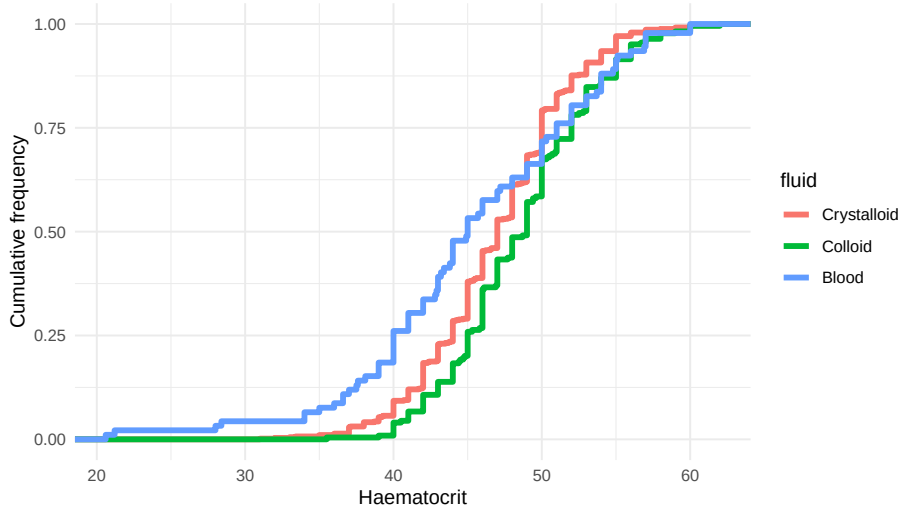
Raincloud plot



.79

The ECDF plot is an alternative. It clearly shows that haematocrit values are lower in those who were given blood. But for the rest it may be a bit harder to interpret:

ECDF plot



.80

8 Two numeric variables

A scatterplot is the most straightforward visualization of the relation between two numeric variables. The correlation coefficient is often used to summarize the strength of relation between them in a single number. Pearson's correlation coefficient is most often used, and is recommended if both variables have a fairly symmetric distribution and the relationship between both variables is fairly linear. Otherwise Spearman's correlation coefficient is often more informative. If there are additional categorical variables to stratify by, then one can separate them by colors or point type ("o" or "+" etcetera) to distinguish them in the scatterplot. Or one can use separate panels/facets.

Scatterplot; correlation coefficient ρ

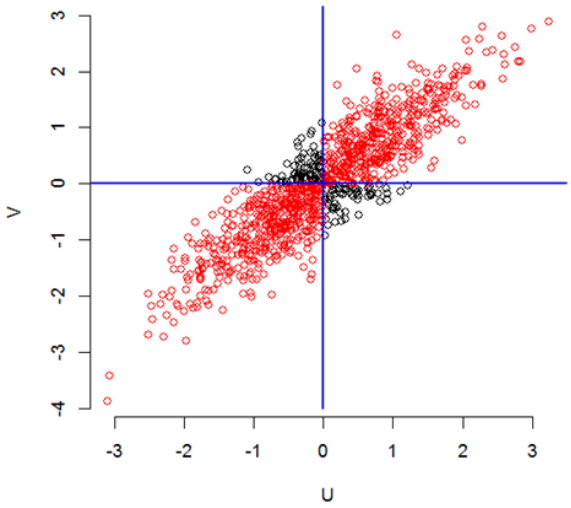
- Strength of relation between numerical variables
  - standardized; not dependent on unit that is used

$$-1 \leq \rho \leq 1$$

  - positive correlation: if value of one variable increases, value of the other also tends to increase
  - negative correlation: if value of one variable increases, value of the other tends to decrease
- Pearson's correlation coefficient
  - degree of *linear* relationship
  - $\rho = -1$ or $\rho = 1$: perfectly linear relationship
    All points on a straight line
  - $\rho = 0$: no linear relationship. Relationship can still be nonlinear

.81

Calculation of Pearson's correlation



$$u_i = \frac{x_i - \bar{x}}{\mathrm{sd}_x}$$

$$v_i = \frac{y_i - \bar{y}}{\mathrm{sd}_y}$$

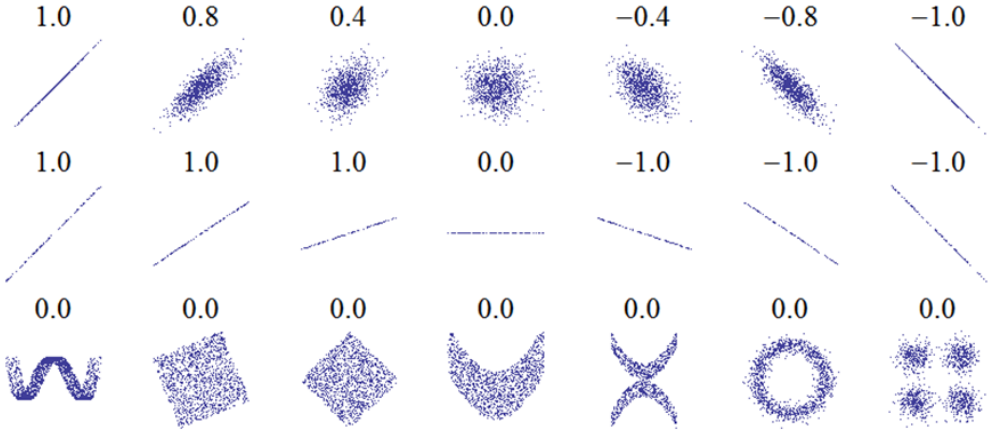
$$\rho = \frac{1}{N-1} \sum_{i=1}^{N} u_i \times v_i$$

N : sample size

.82

As we see in slide 83, quite different nonlinear patterns can give rise to the same correlation of 0 (for both Pearson and Spearman).

Pearson's correlation coefficient; examples



.83

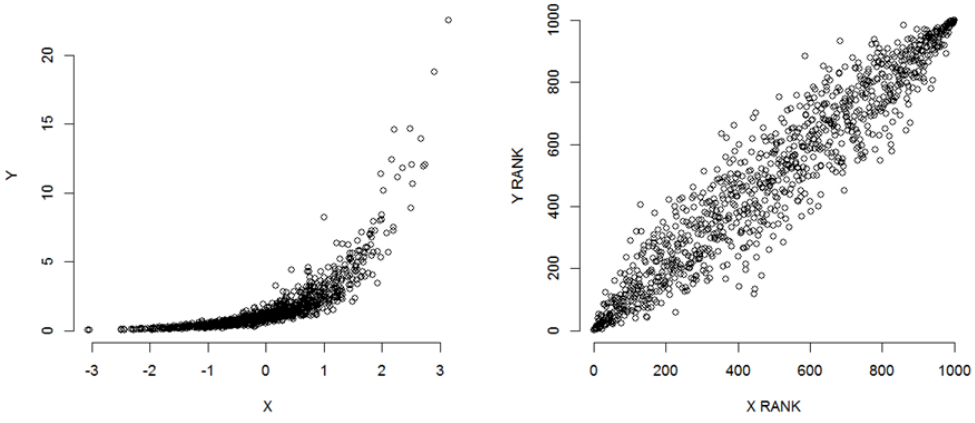
Spearman's rank correlation

- Calculation
  - assign to each value of variable x_i its rank:
    smallest value rank 1, second smallest rank 2 etc.
  - assign to each value of variable y_i its rank
  - compute Pearson's correlation between ranks
- Quantifies degree of monotone relationship
  - rank correlation +1:
    the relationship is positive and perfectly monotone
    (the larger x, the larger y)
  - rank correlation -1:
    the relationship is negative and perfectly monotone
    (the larger x, the smaller y)

.84

In the next slide, the variables have a nonlinear relation. Also, the Y variable has a skewed distribution. If we plot the ranks, the relationship is very linear.

Pearson's versus Spearman's; example



Pearson's correlation = 0.76

Spearman's correlation = 0.95

.85

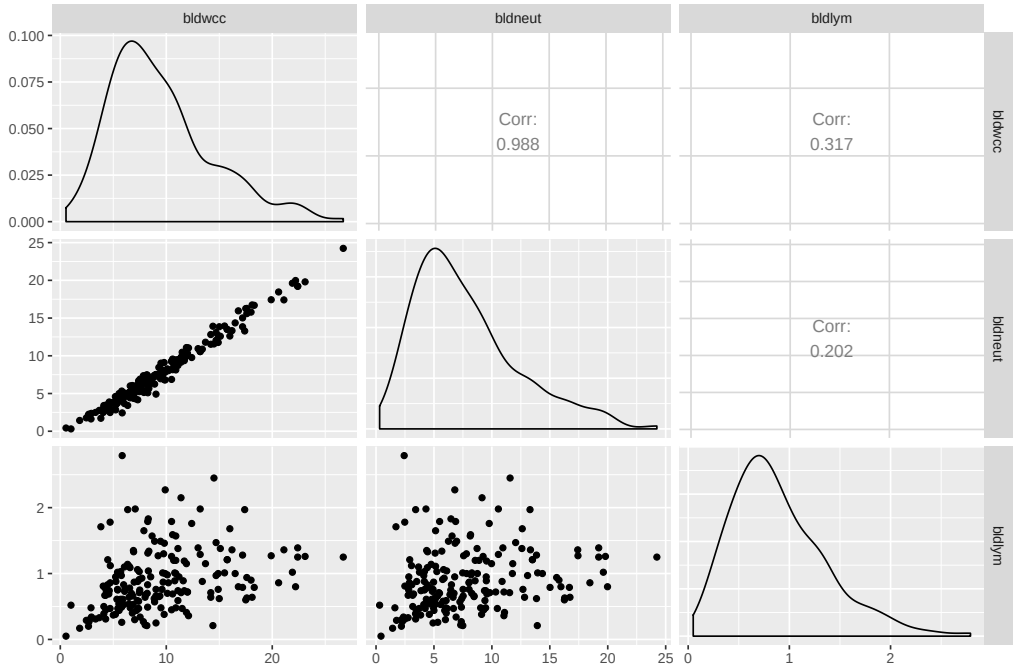
Pearson's versus Spearman's; when?

- Pearson's correlation appropriate if both x and y have approximately symmetric distribution and the scatterplot shows a roughly elliptical pattern
- Otherwise
  - Try to make them symmeric/elliptic via transformation
  - Use Spearman's correlation

.86

If we have several variables for which we want to visualize the pairwise relationships, then a scatterplot matrix may be a good option. As example, we plot the three cell count variables in the TBM data set that were measured in blood.

Pairwise comparison: scatterplot matrix



.87

9 Data visualization in R

R graphics

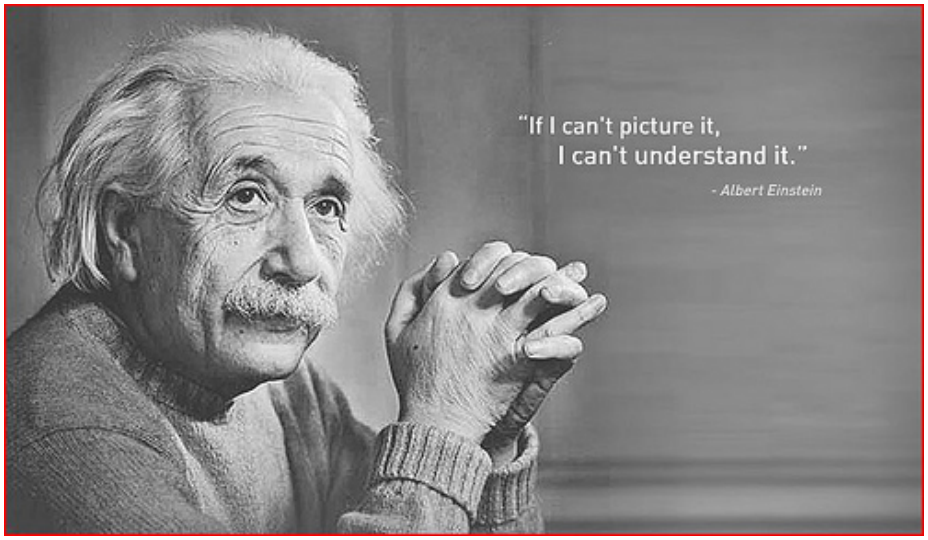
- Base graphics: standard plotting commands
  `plot` used most
  others: `hist, boxplot, ...`
- *`ggplot2`*: based on "the grammar of graphics"
- interactive plots (`rgl`, `plotly`)
- We mostly use `ggplot2` in this course

.88

The ggplot2 package

- Website https://ggplot2.tidyverse.org
  documentation at https://ggplot2.tidyverse.org/reference
- Nice GUI (R package or Web application): https://dreamrs.github.io/esquisse
- Extensions: https://exts.ggplot2.tidyverse.org
- Books
  - ggplot2: Elegant Graphics for Data Analysis
  - Data Visualization
  - R Graphics Cookbook

.89

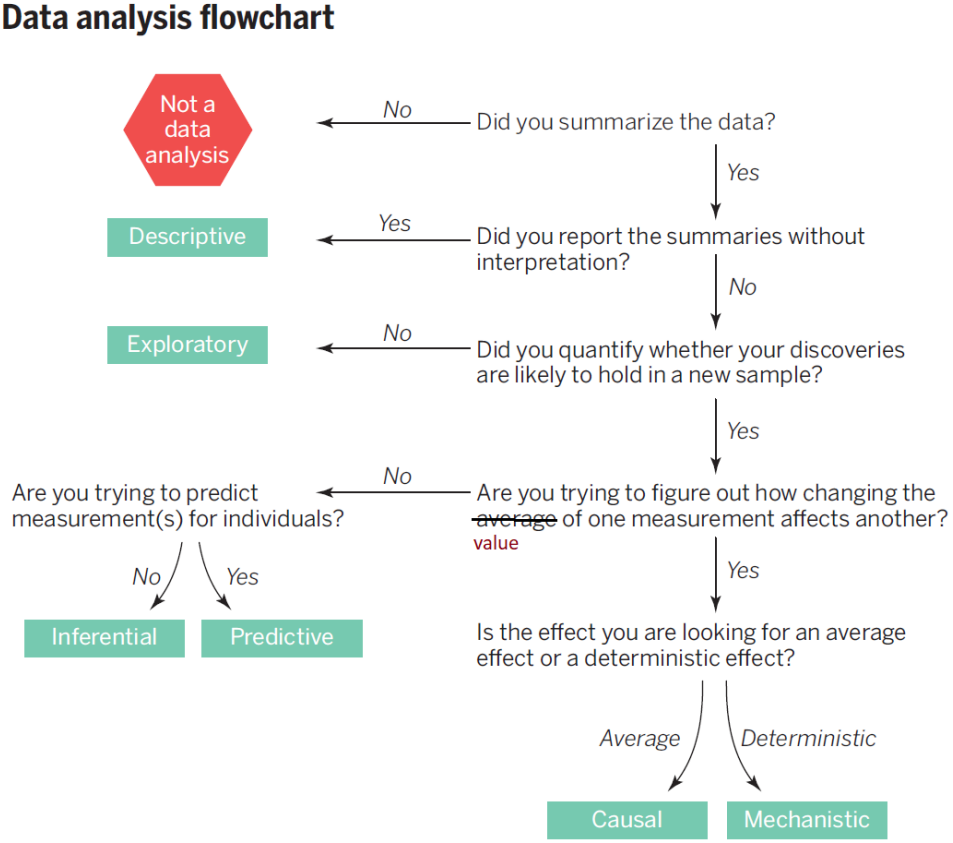


.90

Part III
Statistical Analysis: Main Concepts and Principles; Binomial Distribution

10 Types of study questions

Until now, we focused on ways to summarize variables and visualize the relation between variables. Typically, our decisions with respect to the relations of interest are driven by a specific research question. In a recent paper in Science, Leek and Peng make a distinction between six types of data analysis. They split up the EDA in the definition from Wikipedia that we gave on page 22 into descriptive (which is more or less the same as IDA) and exploratory data analysis[4].



.91

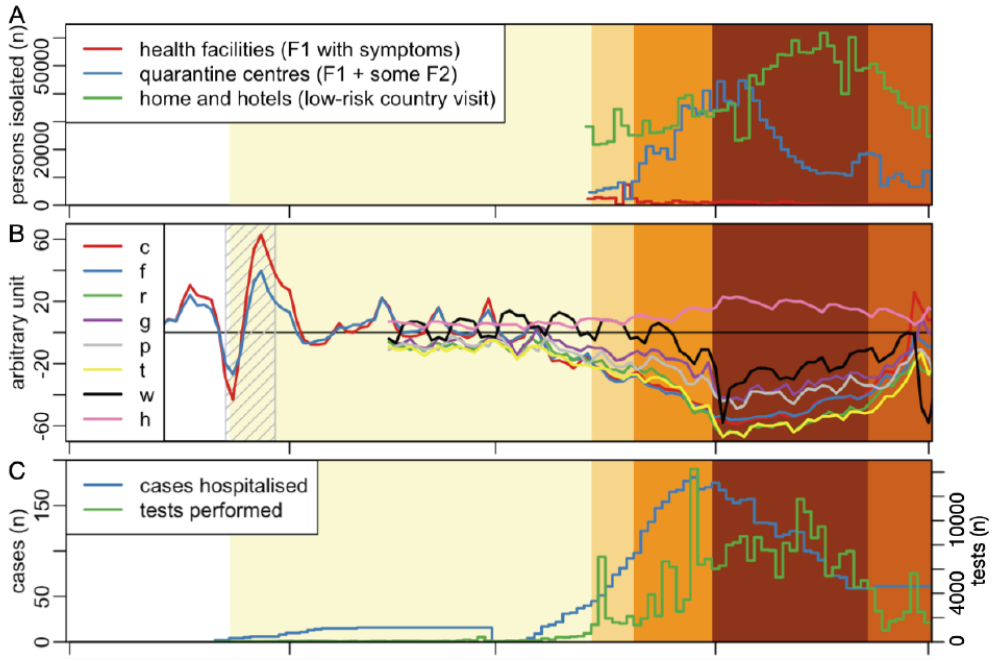
A descriptive analysis summarizes data without giving further interpretation. We did this in Sections 2 and 3 when we described how to summarize the distribution of variables in our data set.

An exploratory data analysis describes relationships between variables in the data to generate ideas or hypotheses or to suggest the appropriate modeling approach. We did this in Section 5. Another example describing and relating several variables at the same time is the figure in slide 92[5].

[4] The wikipedia definition is more in line with the paper To Explain or to Predict?

[5] Again from the paper on the first 100 days of the SARS-CoV-2 control in Vietnam

Exploratory: first 100 days of SARS-CoV-2 in Vietnam

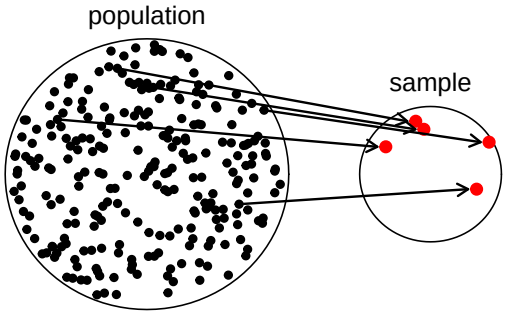


.92

There is a clear distinction between descriptive/exploratory analyses and the other four types. What does it mean to ask "Did you quantify whether your discoveries are likely to hold in a new sample"? If we quantify whether the relationships we discovered are likely to hold in a new sample, we want to generalize our discoveries to a larger population and see our sample as a representative dataset of that population. In a new representative sample from the same population we expect to observe a similar relationship. We collect this representative sample in order to draw conclusions about that population.

"Discovery likely to hold in a new sample"?

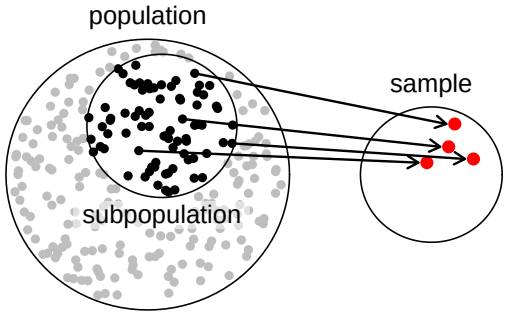
- Data (=sample) from (much) larger *target population* (often hypothetical, of "infinite size")



- Assumption: *representative* sample, reflects population.
  Result from sample *can be generalized to population*, i.e.
  - variable characteristics: sample ≈ population
  - relation between variables: in sample ≈ in population
- *Similar result for a new representative sample, but some differences due to sampling variation*

.93

Generalize to which population?



- Data: patients with dengue shock treated at HTD 2014-17
- Which population does the data represent?
  - all patients with dengue shock treated at HTD
  - all patients with dengue shock in Viet Nam
  - all patients with dengue shock, including patients who may present in the future and at other locations

.94

There are three types of study questions about some characteristics in a population that we want to conclude on based on a sample. For these study questions, statistical methods play an essential role. The fourth, mechanistic study questions, try to quantify deterministic relationships. They are frequent in physics, but hardly ever make sense in medicine.

We further discuss the other three, inferential, predictive and causal study questions. Although there is a philosophical difference, causal analyses are also called explanatory[6].

As an example, we give some options to phrase results in English.

Inferential, predictive or causal?

Relation between variable X and outcome Y

Language: association, risk factor, predictor, effect, cause?

- *Smoking is associated with lung cancer*
- *Smoking is a risk factor for lung cancer*
- *Smoking is a predictor of lung cancer*
- *Smoking has an effect on lung cancer*
- *Smoking can cause lung cancer*

.95

None of the above formulations is incorrect. However, if we replace "smoking" by "alcohol consumption" or "carrying a lighter", some phrases become nonsense. What has changed? Association and risk factor are general terms; they do not explicitly assume a direct causal machanism. The same holds for predictors, although results from a prediction model will be much more robust if there is a causal mechanism. When we use the word "effect", then a causal mechanism is implicated.

[6] Randomized controlled trials are the gold standard to test for the *causal* effect of an intervention (treatment). Efficacy can be shown without having a clear explanation of why the intervention works.

Association does not imply causation

- If x and y are associated, this may be because
  - x influences (or "causes") y
  - y influences (or "causes") x
  - both x and y are determined by one or more other variables
  - by chance
- See Chocolate Consumption, Cognitive Function, and Nobel Laureates and website on spurious correlations
- Causality is not a statistical concept. It requires use of background knowledge

.96

Example: tuberculous meningitis (TBM)

Relation between one or more variables and an outcome

- *Causal (well-defined hypothesis; "effect")*
  Does dexamethasone decrease mortality? (intervention)
  Role of Leukotriene A4 hydrolase (LTA4H) genotype in disease process (etiology)
- *Inferential (risk factors; "association/correlation")*
  What are the risk factors for mortality? No formal causal structure specified
- *Predictive* (personalized medicine; "prognostic/diagnostic value")
  Prognostic: probability of dying within 12 months based on individual characteristics
  Diagnostic: probability to have TBM based on individual characteristics

.97

Sometimes all variables under consideration play an equal role. However, more often one has some direction in mind. In prognostic and causal study questions this direction is explicit; in inferential and diagnostic study questions the direction may be more implicit and reversed.

Relationships between variables

- Without clear direction (correlation, association)
  - Risk factors for outcome (inferential)
  - Diagnosis based on test results and patient characteristics
- With direction $E \rightarrow D$
  - Risk factors for outcome (inferential)
  - Effect of exposure(s) E on outcome (causal)
  - Predict future outcome based on characteristics (prognostic)
- Role of variables in directional analysis
  - Dependent variable D, also called outcome, response, event
  - Independent variables E, also called covariates, covariables, predictors, risk factors, exposures, explanatory variables

.98

This course

- Focus on methods for inferential and causal questions
  - quantify relationship between variables and *average* value of outcome
  - p-value and confidence interval main concepts
  - Often via *regression models*
- Predictive questions: relationship between variables and outcome at *individual* level
  - predictive questions can be answered via regression models, but also via machine learning techniques
- Focus on frequentist estimation techniques (alternative: Bayesian statistics)

.99

11 Sampling variation

Statistics is the science that develops methods to summarize data, to construct models for answering inferential, predictive and causal study questions, and to validate the quality of these models in answering the study question. An important component of statistical science is to interpret the results while taking into account the uncertainty due to the fact that it is only based on a sample from the population.

Sample versus population: randomness

- Statistical science: develop methods to obtain *estimate* from sample that approximates true value in population
- Sample data and estimate will always be
  - (slightly) different from data and estimate from new sample
  - (slightly) different from values in population
- Statistical science: develop methods to quantify *uncertainty and variation* in estimates
- Simplest setting: estimating mean value of a variable in the population

.100

Sample versus population: terminology

- *Parameter*: characteristic of a population
  - Fixed value, but unknown
  - Notation: often Greek letter
    - π for event probability: probability of dying after disease onset/hospitalisation (Covid-19, dengue, TBM,...)
    - μ for mean value of numeric variable: mean viral load at time t after infection or disease onset
- *Estimate*: calculated from observations in sample
  - Assumed to represent (be close to) population value (parameter)
  - Value known in sample, but varies per sample
  - Notation: often with "hat", e.g. $\hat{p}$ for probability; $\hat{\bar{x}}$ for mean value; $\hat{\beta}$ for estimate of parameter in regression model
- More general term: *statistic*. Any function of the data that is used to infer on population characteristic In hypothesis testing we use a "test statistic"

.101

In order to know how far the estimate in our sample is from the population value, we need to quantify the amount of variation that we would observe in our estimate if we repeated the experiment many times. This is what is called the sampling distribution. In hypothesis testing we use a related quantity that is called a test statistic, for which the same principles hold: what is the variation we can observe (in this case under the null hypothesis). The general term that covers both estimate and test statistic is *statistic*.

Sampling variability

- $\hat{p}$ not exactly equal to π
  $\hat{\bar{x}}$ not exactly equal to μ
- new random sample from target population
  $\rightarrow$ new set of individuals
  $\rightarrow$ new value of estimate $\hat{p}$ or $\hat{\bar{x}}$
- How accurate is our (single) estimate?
  sampling distribution: variation in estimate if we repeat the experiment (random sampling and computing estimate) many times

.102

Distribution of estimator of population parameter

- Distribution of statistic if we repeatedly draw random samples from the population
- On average (over many experiments), value of estimator should be *close to* population parameter
  - preferably equal (i.e. *unbiased*)
- Standard deviation describes variation/spread of any distribution
  - standard deviation of statistic is called *standard error*
- If sample size is large then
  - statistic often follows a normal distribution (next class)
  - the amount of variation decreases and the estimate approaches population value

.103

12 Binary Variable; Estimating Proportions

We start with the estimation of the event probability based on the values of a binary variable. The estimator is the proportion. If we define the outcomes of the binary variable as 0 and 1, then the proportion is the number of ones divided by the sample size, which is in fact also a mean value. We can easily describe the distribution of the number of ones; it is called a binomial distribution. Before we return to estimation, we first introduce this distribution.

12.1 Binomial distribution

Binary variable

female/male; alive/death; no/yes; fail/success; 0/1

Examples

- negative NS1 ELISA test in patients with DENV infection
- development of severe dengue in patients with DENV infection

Setting

- Sample of n observations of a binary variable
- Count number of times X that "success" value occurs (negative NS1 ELISA, severe dengue)

.104

The binomial distribution

Definition

The variation in *number of "successes"* X in sample of n individuals. Notation: $X \sim B(n, p)$

- n: total number of observations/individuals
- p: probability of "success" per observation/individual
- X can be any whole number between 0 and n

The observations should meet these requirements:

1. The outcomes of all n observations are **independent** of each other.
2. All n observations have the **same probability of "success"**: p

.105

Example

Flip a fair coin

- A coin is flipped 10 times. Outcome: head or tail
- The variable X is the number of heads among those 10 flips, our count of "successes"
- For each flip, the probability of success, "head", is known as $p = 0.5$
- Number X of heads has the binomial distribution $B(n = 10, p = 0.5)$

.106

We describe and visualize the variation in the observed number of "successes" based on a sample of n individuals, for a specific "success" probability p. As before, important characteristics of distribution are the mean and the variance. For the binimial distribution, both are completely determined by the value of p.

Binomial distribution: mean and variance of X

The mean μ and variance σ^2 of the binomial distribution for a count X are defined by the population success probability p via:

$$\mu = np$$
$$\text{variance} = np(1-p) \rightarrow$$
$$\text{sd} = \sqrt{np(1-p)}$$

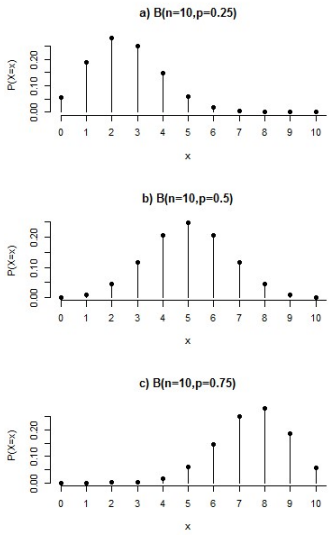
Effect of changing p when n is fixed:

a $n = 10$, $p = 0.25$

b $n = 10$, $p = 0.5$

c $n = 10$, $p = 0.75$

Binomial distribution skewed when p different from 0.5 (especially if n is small).



.107

If we "flip the coint" three times, we can write down all outcomes and their probabilities.

Binomial distribution for $n = 3$

Examples

$n = 3$ patients, each can either have a success (S) or a failure (F)

Pt 1	Pt 2	Pt 3	Probability	Nr of successes
F	F	F	$(1-p)^3$	0
S	F	F	$p \cdot (1-p)^2$	1
F	S	F	$p \cdot (1-p)^2$	1
F	F	S	$p \cdot (1-p)^2$	1
S	S	F	$p^2 \cdot (1-p)$	2
S	F	S	$p^2 \cdot (1-p)$	2
F	S	S	$p^2 \cdot (1-p)$	2
S	S	S	p^3	3

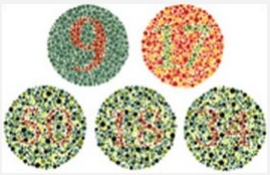
$X \sim B(n = 3, p) \rightarrow P(X = 2) = 3 \cdot p^2 \cdot (1-p)$

.108

In larger samples, we can use a function in R to calculate the probability of a specific number of successes $P(X = k)$ as well as the cumulative probability $P(X \leq k)$. We give another example.

Example: Color blindness

- Frequency of color blindness in Caucasian American male population is about 8%
- We take random sample of size 25 from this population



Probability that exactly 3 individuals in sample are color blind?

- Use R: `dbinom(3,size=25,prob=0.08)` ("d" in dbinom refers to "*d*ensity/probability mass" function $P(X = k)$)
- Result: $P(X = 3) = 0.188$

Probability that 2 or fewer in the sample are color blind?

- Use R: `pbinom(2,size=25,prob=0.08)` ("p" in pbinom refers to the cumulative *p*robability $P(X \leq k)$)
- Result: $P(X \leq 2) = 0.68$

.109

Estimation of π

- If $X \sim B(n,p)$, then $\text{mean}(X) = np$ and $\text{sd}(X) = \sqrt{np(1-p)}$
- Estimation of population probability π: compute frequency $\hat{p} = X/n$
- Note: frequency is a mean. Let

$$X_i = \begin{cases} 1 & \text{if "success"} \\ 0 & \text{if "fail"} \end{cases}$$

  then $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = X/n$, frequency of "success"
- SE=$\text{sd}(\hat{p}) = \frac{1}{n}\text{sd}(X) = \frac{1}{n}\sqrt{n\pi(1-\pi)} = \sqrt{\frac{\pi(1-\pi)}{n}}$
  Decreases with increasing n

.110

13 Testing a hypothesis

Hypothesis testing plays a very important role in answering inferential and causal study questions. The main quantities in (frequentist) hypothesis testing are the p-value and the confidence interval. We first give an example in which we want to infer upon the probability of some binary outcome.

13.1 Single proportion

Example

- A new treatment for cancer
- Current treatments achieve a tumour response probability of 0.5
- In a clinical study of the new treatment with 20 patients, 15 showed a tumour response

Does this "prove" that the new treatment is better than the current treatments?

.111

The standard approach of hypothesis testing is to assume that the old situation holds, which is called the null hypothesis. Our purpose is usually to defeat the null hypothesis in favour of some alternative hypothesis. For this, we quantify the probability of our observation, or an even more extreme value, under the null hypothesis. This is called the p-value. If this probability is very low, then we reject the null hypothesis and conclude that the alternative hypothesis holds.

Formalisation of the problem

Probability model for the data

- The number of tumour responses $X \sim B(n = 20, \pi)$
  - Population success probability π
  - Variation based on $n = 20$ individuals in sample follows binomial distribution

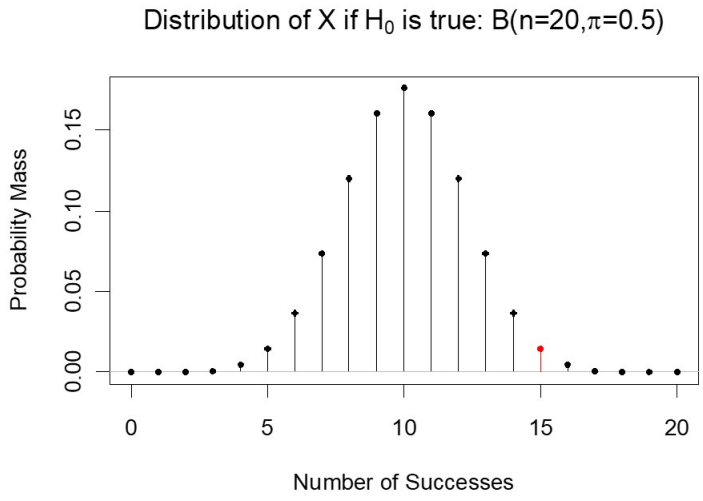
Null hypothesis H_0

- $H_0 : \pi = 0.5$ (no effect)
  - $\hat{p} = 15/20$ successes occurred by chance. How likely is this?

Alternative hypothesis H_A

- H_A : True $\pi > 0.5$ (one-sided) - "better"
- H_A : True $\pi \neq 0.5$ (two-sided) - "different"

.112



.113

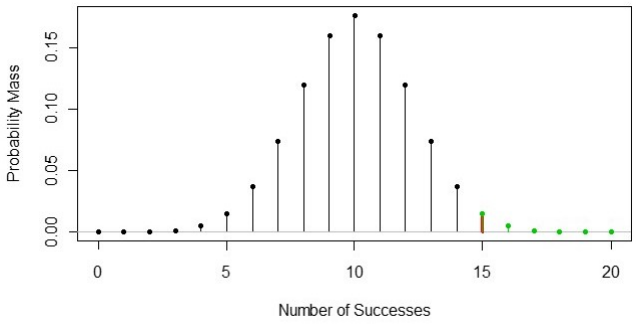
In most situations the null hypothesis formulates a specific value, and we formulate as alternative hypothesis any value different from the null hypothesis. This value may be lower and higher than the value as specified under the null hypothesis. Sometimes we specify the alternative hypothesis in one specific direction. In that case the null hypothesis can equivalently be formulated as any value in the other direction. Hence, if we choose $H_A : \pi > 0.5$, then we could as well choose

$H_0 : \pi \leq 0.5$. This does not make any difference with respect to the p-value, because we calculate the *maximum* value of the probability of the observed value under the null hypothesis, which is attained if $\pi = 0.5$.

P-value for one-sided alternative $H_A : \pi > 0.5$

- How likely are 15/20 successes or more extreme ("**better**") if null hypothesis is true?
- If $\pi = 0.5$:

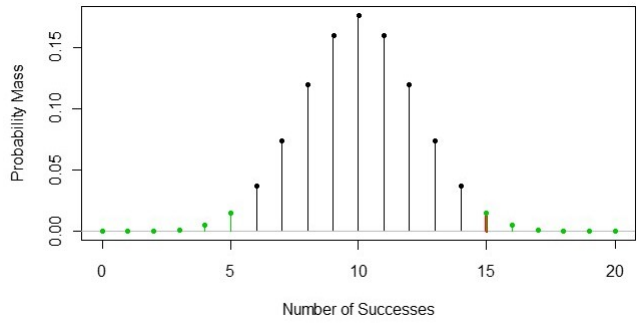
$$\begin{aligned} \textit{p-value} &= P(X \geq 15) \\ &= P(X = 15) + P(X = 16) + ... + P(X = 20) \\ &= 0.014 + 0.004 + ... \\ &= 0.0207 \end{aligned}$$



.114

P-value for two-sided alternative $H_A : \pi \neq 0.5$

- If the null hypothesis is true (which on average gives 10 successes), how likely to have deviation of 5 or more extreme
- "More extreme" : $X \leq 5$ or $X \geq 15$
- If $\pi = 0.5$: p-value = $P(X \leq 5) + P(X \geq 15) = 0.041$



.115

We see that the p-value is larger under the two-sided alternative hypothesis. However, whether the p-value is small depends on the possible range of probabilities. Under the one-sided alternative hypothesis we only quantify the probability in one direction; this probability is typically always smaller than 0.5. Under the two-sided alternative, it is often close to twice the probability under the one sided alternative. Hence, under the one-sided alternative we are typically twice as strict with respect to the decision when to reject the null hypothesis.

P-value for the example with R

```
> prop.test(x=15,n=20,p=0.5,alternative="two.sided")

        1-sample proportions test with continuity correction

data:  15 out of 20, null probability 0.5
X-squared = 4.05, df = 1, p-value = 0.04417
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5058845 0.9040674
sample estimates:
   p
0.75
```

- For one-sided tests, use alternative="greater" or alternative="less"
- Reject null hypothesis if p-value is below a certain *significance level* α

.116

One- and two-sided tests

- With symmetric distribution:
  - *p-value (two-sided test)* = $2 \times$ *p-value(one-sided test)*
- Two-sided alternatives are generally preferred
  - We cannot a priori exclude harm of a new treatment
  - One-sided p-value always smaller than 0.5
- Hence
  - Either two-sided tests at significance level α
  - Or one-sided tests at significance level $\alpha/2$
  - Often, $\alpha = 0.05$ is chosen

.117

More generally, a hypothesis formulates a relation between exposure and outcome that is supposed to exist in a population. Typically this relation is assumed to be causal. It starts by assuming that there is no relation, the null hypothesis, and we hope to reject the null hypothesis based on our data. We reject the null hypothesis if the value of our test statistic is very unlikely to occur if the null hypothesis were true.

Hypothesis testing, basic structure

Effect exposure **E** on disease **D**

$$\mathbf{E} \longrightarrow \mathbf{D} \quad \text{in population}$$

1. Null hypothesis H_0: no effect (often "$\beta = 0$")
2. Alternative hypothesis H_1: there is an effect (often "$\beta \neq 0$")
3. Calculate value of test statistic *TEST* based on sample data
4. Calculate p-value: probability that *TEST* exceeds some value if H_0 were true
5. If p-value small (often: $< 5\%$): reject H_0
   unlikely that observed difference is due to chance
6. Otherwise: do not reject H_0
   Does not imply that H_0 is true (power):
   No proof of effect $\neq$ *proof of no effect*

.118

13.2 Compare proportions between two subgroups

2x2 table; example group and outcome

	Disease present	Disease absent	total
male	*32*	17	*49*
female	*118*	127	*245*
total	150	144	294

- Percentage diseased in the male group:

$$\frac{32}{49} = 65\%$$

- Percentage diseased in the female group:

$$\frac{118}{245} = 48\%$$

- Does disease probability differ by gender?

.119

Hypothesis test

- H_0: probability of disease the **same** in both groups
- H_A: probability of disease **differs** per group
- Can the difference be due to chance, i.e. $H_0 : \pi_F = \pi_M$ holds?
- How do we quantify difference from equal probability?
- "Equal disease probability" is same as saying that group and outcome are unrelated/independent: knowing the group does not help in learning the outcome

.120

Chi-squared test for independence

- For each cell: compare **observed number** (O) with **expected number** (E) under null hypothesis of independence
- Large discrepancy between O and E is an indication that probabilities in both groups differ

.121

Expected count under null hypothesis

	Disease present	Disease absent	total
male	32	17	*49*
female	118	127	245
total	*150*	144	*294*

- Proportion male: *49/294 = 0.17*
- Proportion diseased: *150/294 = 0.51*
- Expected number diseased and male under H_0:

$$n \times \hat{P}(\text{male}) \times \hat{P}(\text{Diseased}) = 294 \times (49/294) \times (150/294) = 25$$

.122

Chi-squared statistic calculation (I)

	O	E
Disease present + male	32	25
Disease present + female	118	125
Disease absent + male	17	24
Disease absent + female	127	120
Total	294	294

.123

Chi-squared statistic calculation (II)

	O	E	**O-E**
Dis. present + male	32	25	7
Dis. present + female	118	125	-7
Dis. absent + male	17	24	-7
Dis. absent + female	127	120	7
Total	294	294	0

Always equals 0

.124

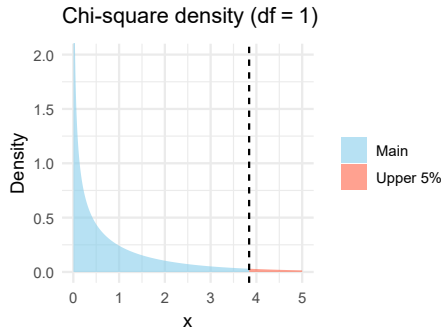
Chi-squared statistic calculation (III)

	O	E	O-E	$(O-E)^2/E$
Dis. present + male	32	25	7	1.960
Dis. present + female	118	125	-7	0.392
Dis. absent + male	17	24	-7	2.042
Dis. absent + female	127	120	7	0.408
Total	294	294	0	$X^2=4.802$
				P = 0.028

Under null hypothesis, statistic X^2 has a chi-squared distribution with one degree of freedom

.125

Upper percentile of χ^2-distribution



- Starts at zero
- 95-percentile at 3.84

.126

Chi-squared test of independence in R

```
> chisq.test(matrix(c(32,118,17,127),nrow=2),correct=FALSE)          # Variant 1

        Pearson's Chi-squared test
X-squared = 4.802, df = 1, p-value = 0.02843

> prop.test(x=c(32,118),n=c(49,245),correct=FALSE)     # Variant 2, with CI for difference

        2-sample test for equality of proportions without continuity correction
X-squared = 4.802, df = 1, p-value = 0.02843
alternative hypothesis: two.sided
95 percent confidence interval for the difference:
 0.024 to 0.32
sample estimates:
   prop 1    prop 2
     0.65      0.48
```

- Note: *correct=TRUE* gives "Yates' continuity correction" (default in R)

.127

Interpreting the chi-squared test of independence

- After identifying the association, you may want to know the strength of this association
- The strength of the association is measured by the difference in proportion, relative risk (RR), or odds ratio (OR) (see Thursday class on logistic regression)

.128

13.3 Categorical variables - more than 2 groups

Chi-squared test for independence

- The chi-squared statistic can be used for testing association in any two-way contingency table
- Variables do not need to correspond to group and outcome
- For a table with c column and r rows, the distribution of the statistics (under null hypothesis) is a chi-squared distribution with $(r-1)\times(c-1)$ degrees of freedom.

.129

2x3 table

	Body Mass Index (BMI)			
	Low	Normal	Obese	total
male	a	b	c	a+b+c
female	d	e	f	d+e+f
total	a+d	b+e	c+f	

.130

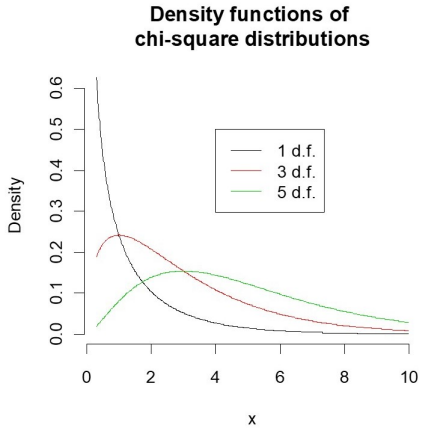
Example with more than two subgroups

TLR2 T597C SNP bacterial genotype frequencies in patients with TBM caused by Beijing genotype isolates compared to chord blood controls (Caws *et al.*; PLoS Pathog 2008)

	Genotype			
Lineage/group	TT	TC	CC	Row Total
Chord blood controls	205 0.544	154 0.408	18 0.048	377
TBM Beijing	24 0.400	25 0.417	11 0.183	60
Column Total	229	179	29	437

.131

Density function of chi-squared distributions



.132

Example with more than two groups in R

```
> chisq.test(matrix(c(205,24,154,25,18,11),ncol=3),
corr=FALSE)

        Pearson's Chi-squared test

data:  matrix(c(205,24,154,25,18,11),ncol=3)
X-squared = 16.3897, df = 2, p-value = 0.0002761
```

"There is clear evidence for an association between TLR2 T597C genotype and TBM caused by Beijing genotype isolates (p=0.0003)."

.133

13.4 Alternative tests for specific settings

There are a couple of variants of the chi-squared test for independence that are the preferred choice in specific settings. The Fisher's exact test is preferred if al least one of the combinations is very unlikely under the null hypothesis. This formally translates into: use Fisher's exact test if the expected number in at least one of the cells in the $c \times r$ table is smaller than 1 (in the past this was set at "smaller than 5"). This is more likely to happen with small sample size. If the alternative hypothesis assumes a trend in the relation between the two variables, then the Cochran-Armitage test is preferred. If the test is not about a difference in probability between two or more subgroups, but a difference between two measurements *from the same individual*, then the Mcnemar test is recommended.

Fisher's exact test

- Used in small samples: expected number under null hypothesis is ≤ 1 in at least one cell
  - note: many programs (including R) give a warning if chi-squared test is used and expected frequency under null hypothesis is ≤ 5 in at least one cell

.134

Chi-square trend test (Cochran-Armitage)

- One or both variables are seen as ordinal
  - example: MRC grade at diagnosis (values I, II and III) in HIV-negative and HIV-positive TBM patients

	I	II	III	total
HIV−	298	342	149	702
HIV+	196	251	73	520

  - H_0 : MRC grade at diagnosis does not depend on HIV status
    H_1 : there is a systematic trend, i.e. HIV infection makes higher MRC grade at diagnosis more (or less) likely

.135

Mcnemar test for paired dichotomous outcomes

- Two correlated observations per individual
  - example: disease status per individual before and after treatment

	mild after	severe after	total
mild before	100	50	150
severe before	200	100	300
total	300	150	450

.136

Part IV
Study Designs: RCTs; Sample Size Calculation

You have learned the principles of statistical inference and the use and interpretation of regression models. In the last two sections we explain how these models can be used as tools to answer a research question and we give some suggestions how to report the answers in an article.

It is informative to read the reviewer guidelines from scientific journals. Here we give them from PLOS neglected tropical diseases, which are split into **Methods**, **Results** and **Conclusions**. We start with the **Methods**.

Criteria for reviewers (PLOS neglected tropical diseases)

Methods:

1. Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
2. Is the study design appropriate to address the stated objectives?
3. Is the population clearly described and appropriate for the hypothesis being tested?
4. Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
5. Were correct statistical analysis used to support conclusions?
6. Are there concerns about ethical or regulatory requirements being met?

.137

14 Study questions and study designs

Every study starts with formulating the research questions that align with the study objectives. In Section 10 we described the main types of research questions.

Note that the PLOS neglected tropical diseases guidelines focus on formulating a hypothesis. This is of primary importance for causal study questions that formulate a hypothesized causal mechanism. Inferential study questions often serve to generate a better understanding of the mechanisms by considering potential risk factors for the outcome. Some people therefore call them exploratory in nature. Inferential studies often do report p-values, which only make sense if there is a corresponding hypothesis. However, the hypotheses that are formulated do not explicitly refer to causal mechansism. Only after the results have been obtained, one tries to interpret the results and suggests possible mechanisms.

Predictive study questions concern the relation between variables and outcome at the individual level. The aim is to come up with a model that gives individual predictions that are close to the actual outcome and that helps discriminate low and high risk patients. P-values are tightly connected to testing of a hypothesis about the average relation at the population level. Other measures are required to evaluate performance of an individual prediction model. Although variables with small p-value are likely to be good predictors as well,

this is not always the case: studies based on large samples may find small but statistically significant effects. On the other hand, in small studies non-significant variables may have a strong relation with the outcome and including them in the model may improve predictive performance.

What is your question?

- Study question: descriptive, exploratory, *inferential, predictive, causal*
- *Causal*
  - effect of intervention (treatment) or exposure
  - understanding disease etiology and biological mechanisms

  Tool: hypothesis tests. Decision criteria: p-value, confidence interval
- *Inferential*
  - find most important risk factors for the outcome, often based on p-values
  - no explicit causal hypotheses; *association* instead of effect
  - suggest possible mechanisms based on the results
  - no strong conclusions can be drawn; findings need to be validated in separate studies
- *Predictive: prognosis/diagnosis*

.138

Predictive questions

- Obtain *accurate* predictions for the *individual* case
- "Accurate" via measures of predictive value instead of p-values
  - calibration: on average accurate (unbiased) *"average Tmax in HCMC on March is* 34 °C*"*
  - discrimination: distinguish between low and high risk cases (AUC, Brier score) *"this March 31, Tmax in HCMC will be* 32 °C*"*
- Proper validation of predictive value important
  Calibration and discrimination with new data may be much worse, due to overfitting or being different population
- Use statistical regression models or machine learning techniques

.139

14.1 Study designs

After the objectives of the study and the study questions have been formulated, the next step is to choose the study design. The main characteristics to consider when deciding on study design in relation to the study question are given in slide 140.

In medical and epidemiological research, the most common type of experimental study design is the randomized controlled trial, which is the gold standard to quantify the effect of an intervention.

Types of studies designs

Observational Researcher collects data without influencing course of events	⟷	**Experimental** Researcher influences course of events and studies effect of the intervention
Prospective Data collected forwards in time from the start of the study	⟷	**Retrospective** Data refer to past events and is acquired from existing sources
Longitudinal Study changes over time, observations taken on more than one occasion	⟷	**Cross-sectional** Observations taken only once

.140

Experimental: trial

- *Prospective* study comparing the causal effect of a novel *intervention* against a *control* intervention
- *Intervention*
  - therapy/drug
  - medical device, decision support tool
  - behavioral intervention/training
- *Control*
  - placebo
  - standard of care
- *Prospective*
  - all study procedures (including study hypotheses and planned analyses) defined before the start of the study
  - participants followed from a well-defined "baseline" point (intervention)

.141

Randomized Controlled Trial (RCT)

- Random assignment of participants to intervention or control
- Control other factors
  - equality of patient care (except for randomized intervention)
  - equality of endpoint assessment
- Blinding (if possible)
  - patients
  - investigator/physician
  - both (double-blind)
- Special issues of experiments in humans
  - safety
  - compliance
  - ethics

.142

Not all causal study questions can be answered via an experimental study design. Some can only be answered via observational studies.

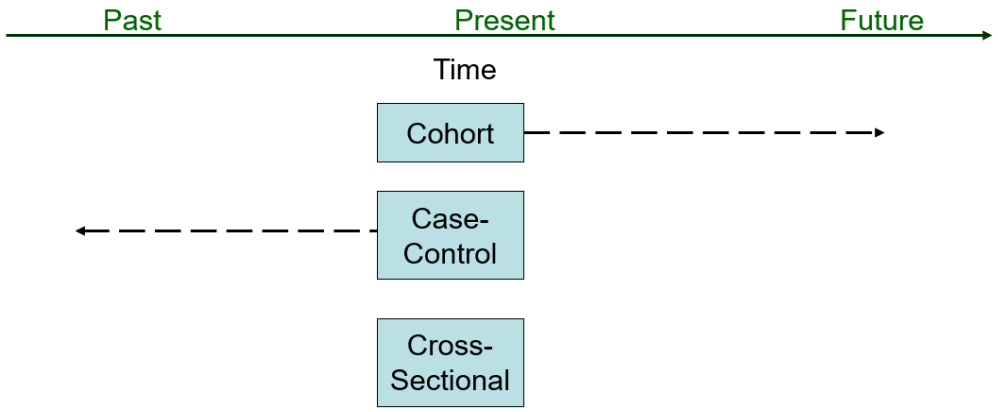
Observational studies

- To answer (causal) questions if exposure or intervention cannot be randomized
- Examples:
  - relation between passive smoking and lung cancer
  - relation between radiation exposure and miscarriage
  - relation between alcohol drinking and suicide
- "The aim of the observational study should be to arrive at the same conclusions that would have been obtained by an experimental trial" Gray-Donald and Kramer (1988)

.143

Observational study designs

- Main study designs:
  - cohort study (longitudinal)
  - case-control study
- Cross-sectional studies
  - usually inferential
  - we don't know what came first, reverse causation possible



.144

Bias in causal observational studies

- Lack of randomization → assessment of causal factors in observational studies may be biased: *confounding*
- Smoking and lung cancer: an observed association could be due to
  - smoking causes lung cancer (causation)
  - cancer (or a pre-cancerous condition) is a factor inducing cigarette smoking (reverse causation)
  - both smoking and lung cancer are caused by a specific genotype (confounding)
- The design of observational studies needs careful epidemiological and causal considerations

.145

Statistical methods to control confounding

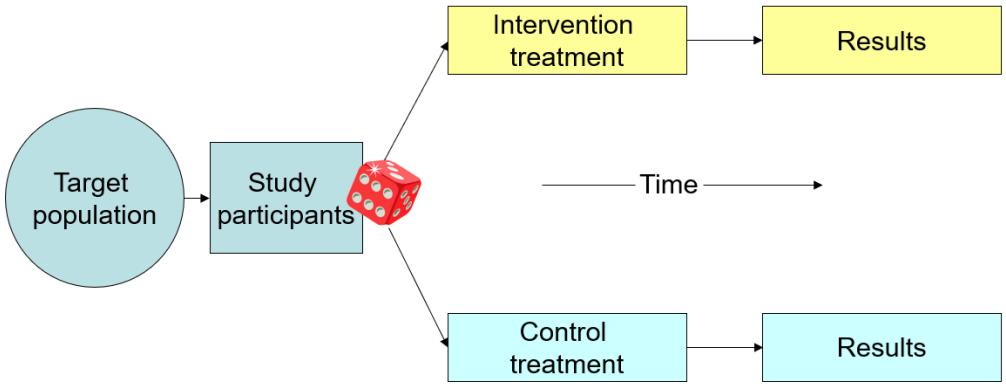
- Avoid confounding by design → randomization (RCT)
- Adjust for confounding in the analysis or the design
  - stratification
  - regression
  - matching

  May require use of advanced statistical methods
- Confounders can only be adjusted for in the analysis if the information on the confounder has been collected, i.e. we cannot adjust for "unmeasured confounding"

.146

15 Randomized controlled trial

Structure of an RCT



.147

Randomization

- Chance decides about the assignment of the patients to their treatment group
- Advantages of randomization
  - (on average) produces treatment groups which are comparable with respect to known and unknown risk factors
  - avoids investigator bias in the allocation of patients
  - avoids bias due to confounding factors
- Randomization is the only known method with the above characteristics

.148

The RCT is the most important study design for testing the efficacy of an intervention. Given its paramount importance for clinical practice, specific guidelines for RCTs have been formulated that help establishing reliability of results. A recent one is the ICH E9(R1) addendum on estimands and sensitivity analysis that was issued by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) in 2019.

The estimand concept in the ICH E9 addendum precises the description of treatment effect. It involves five components ("attributes") that reflect the clinical question to be answered: 1. treatment conditions; 2. target population 3. variable of interest (outcome, endpoint) 4. population-level summary 5. the role of intercurrent events (e.g. noncompliance and loss to follow-up). They need to be specified before the trial starts.

To illustrate the estimand framework, we use a large trial on the efficacy and harm of dexamethasone in HIV-infected patients with tuberculous meningitis (ACT-HIV).

Adjunctive Dexamethasone for Tuberculous Meningitis
in HIV-Positive Adults

- Objective: is dexamethasone a safe and effective addition to anti-TBM treatment
- Treatment: 6 (MRC I)/8 (MRC II or III) weeks of adjunctive dexamethasone (N=263) versus placebo (N=257)
- Endpoint: overall survival within 12-months from randomisation
- *Intercurrent events:*
  - switch to open-label dexamethasone (n=138; 26.5%)
  - <7 days randomised study drug for reason other than death (n=22; 4.2%)
  - <30 days of anti-tuberculosis drugs for reason other than death (n=7; 1.3%)

.149

ICH E9(R1) addendum on estimands

Precise definition of the treatment effect reflecting the clinical question posed by a given clinical trial objective

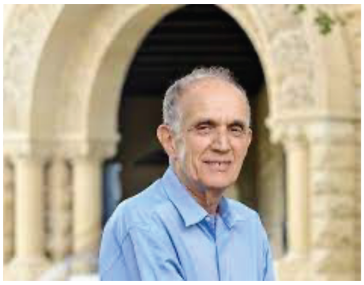
Precise specification of five attributes:

1. treatment conditions *dexamethasone versus placebo*
2. *target population (e.g. eligibility criteria)*
3. *variable (outcome, endpoint)*
4. population-level summary (ratio, difference) *hazard ratio:* $0.85(0.66, 1.10); P = 0.22$
5. *other intercurrent events, occurring after treatment initiation*
   - treatment switch or discontinuation, additional medication, rescue medication
   - terminal events (death, leg amputation for diabetic foot ulcers)

.150

Intercurrent events can pose major problems in interpretation of the results. In the ACT-HIV trial, a quarter of the participants switched to open label corticosteroid treatment during follow-up, which breaks the initial randomization.

Why haven't research questions been defined sufficiently precisely?



Efron is the second winner of International prize in Statistics Foundation

"There could not be worse experimental animals on earth than human beings;

- they complain,
- they go on vacations,
- **they take things they are not supposed to take**,
- they lead incredibly **complicated lives**, and, sometimes,
- they **do not take their medicine**."

(Efron, 1998)

Intercurrent events may break randomization

.151

In order to take account of the impact of intercurrent events, traditionally a distinction has been made between intention-to-treat, per-protocol and as-treated analyses. The main purpose of the ICH E9(R1) addendum is to reconsider and update these guidelines. For this reason, a distinction is made between several types of analysis.

Main types of analysis

- Intention-to-treat (ITT)
  - includes all randomized patients
  - participant is analyzed according to the randomized treatment group (regardless of the actual treatment given during the study)
  - usually main analysis

  Describes efficacy of an intervention in "real life" (where patients stop treatments, switch to other treatments etc)
- Per protocol (PP)
  - includes all patients who "followed" the protocol (no major violation of inclusion/exclusion criteria, treated according to the randomized treatment arm, completed follow-up)
  - often used, but incorrectly: randomization is lost if they are excluded based on what happens during follow-up
  - proper correction for loss of randomization may require use of advanced statistical methods

.152

In the ACT-HIV trial, the switch to open-label dexamethasone was not taken into account, not even in the PP analysis. For the individuals in the dexamethasone arm, treatment continued as allocated. However, for those in the placebo group, a switch to dexamethasone may have improved their disease course if dexamethasone were beneficial. This would dilute the observed effect size because the patients who were doing worse were more likely to switch from placebo to dexamethasone.

ACT HIV trial: analyses

- Intention-to-treat population: dexamethasone versus placebo, including possible switch to open-label dexamethasone
  - switch to open-label dexamethasone if after clinical and neuroradiological review the attending physician believes a patient's neurological deterioration is due to tuberculoma
  - mentioned as limitation in discussion
- Per-protocol population: exclude
  - negative confirmatory HIV test (n=2; 0.4%)
  - received >6 days of TB drug before enrollment (n=1; 0.2%)
  - <7 days randomised study drug for reason other than death (n=22; 4.2%)
  - <30 days of anti-tuberculosis drugs for reason other than death (n=7; 1.3%)
- No correction for switch to open-label dexamethasone. Reviewer on incomplete PP-analysis: "Use an approach that appropriately accounts for making comparison between non-randomized groups." *Not easy!*

.153

15.1 Some practical advice for study design

(RCT) protocol development

- Describe study design
- Define target population and treatment groups
- Formulate clear study hypotheses and study outcomes
  - define primary outcome and hypothesis
  - secondary outcomes (not too many)
  - additional exploratory analyses
  - avoid fishing expeditions
  - preference for objective "hard" outcomes
- Ensure key outcomes are accurately collected in the Case Report Form
- Clearly describe planned key analyses, with effect measure
- Give sample size justification

.154

During the study

- Ensure high data quality
- Accurate capture of key outcomes
- Minimize the amount of missing data
- Write a separate statistical analysis plan (as early as possible)
  - Detailed description of all planned statistical analyses
  - Tests and analysis methods used
  - Derivation rules for complex outcomes, based on the case report form
  - How to deal with missing data etc.
- Clean the data prior to the analysis (consistency checks etc.)

.155

16 Sample Size Calculation

Importance

- Question: How many samples/patients do I need to include in my study to show an effect with reasonable certainty?
- Critical issue in the planning of an RCT (also seen as important in other studies)
  - ethical aspects (for studies of an intervention)
    - study large → too many patients exposed to the risk of the intervention
    - study too small → study has not enough power to detect clinically important differences
  - economical aspect (any study)
    - resources and time are wasted
- Based on assumptions with respect to true value of parameter → often a fairly wild guess

.156

Based on primary outcome analysis

- Variable measured during study to answer the primary study question.
  - must be clearly defined upfront (in the study protocol)
- Example
  - ACT-HIV: death within 12 months

.157

16.1 Estimating a single proportion

Confidence interval for a proportion

- Given an observed proportion $\hat{p} = x/n$, an approximate 95% CI for the population proportion π is given by

$$\hat{p} \pm 1.96 \times \text{SE} = \hat{p} \pm 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$$

- The half-width ("precision") of the confidence interval is

$$1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$$

.158

Recipe for sample size calculation

- Make a realistic assumption about how large the population proportion π might be (if nothing is known, chose $\pi = 0.5$ which is conservative)
- *Choose a target precision*
- Solve the formula

$$\text{precision} = 1.96 \times \sqrt{\pi(1-\pi)/n}$$

for the sample size n:

$$n = 3.84 \times \pi(1-\pi)/\text{precision}$$

.159

Example

- Setting: HIV-positive adults hospitalized at HTD with tuberculous meningitis (TBM)
- Goal: determine the 9-month mortality in this population
  - mortality is assumed to be around 0.5 (50%)
  - target precision is 0.1
  - $n = 3.84 \times \frac{0.5 \times 0.5}{0.1^2} = 96$
- Need to enroll at least 96 patients

.160

16.2 Comparing two independent groups (of equal size)

Steps in hypothesis testing

Effect exposure **E** on disease **D**

$$\mathbf{E} \longrightarrow \mathbf{D} \quad \text{in population}$$

1. Define null hypothesis H_0 (usually "no effect, $\beta = 0$") and alternative hypothesis H_1 ("$\beta \neq 0$") for the population
2. Plan study and collect data
3. Calculate summary statistics
4. Calculate test test statistic *TEST* based on sample data
5. Calculate p-value: probability that *TEST* exceeds some value if H_0 were true
6. *Draw conclusion on whether to reject H_0*

.161

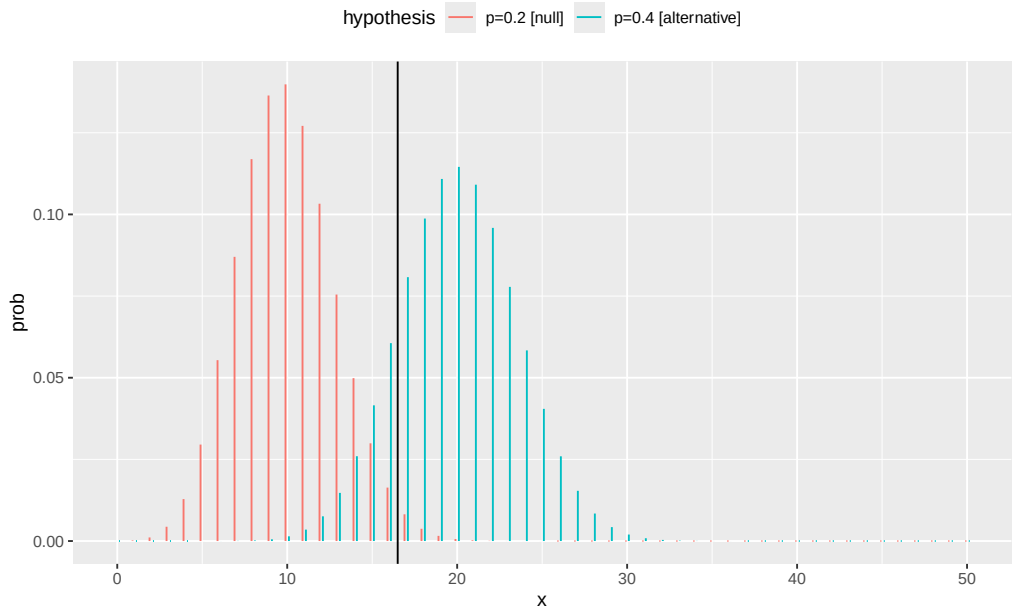
Decisions

Decision	H_0 true (no effect)	H_0 false (effect)
H_0 not rejected	undecided	*type II error*
H_0 rejected	*type I error*	correct

- *TYPE I error: reject H_0 while it is true*
  - maximum probability type I error determined in advance via significance level α (typically $\alpha = 5\%$ or $\alpha = 1\%$)
  - p-value $\leq \alpha$: reject H_0, observed difference unlikely due to chance
  - p-value $> \alpha$: do not reject H_0. Does not imply that H_0 is true
- *TYPE II: do not reject H_0 while it is not true*
  - probability of type II error depends on effect size and sample size
  - choose appropriate sample size based on chosen power (typically: power = 80% or power = 90%) and hypothesized magnitude of effect size
  - power: 1 − probability of type II error

.162

Power visualized



.163

Sample size calculation (numeric outcome)

- Assume (hypothesize)
  - difference in population means between the groups: δ
  - standard deviation of outcome around the group means: σ

  $\rightarrow$ standardized effect size: $\Delta = \frac{\delta}{\sigma}$
- Choose
  - significance level α
  - power
- Simplified formula
  - $\alpha = 5\%$, power 80%: sample size per group is $n \approx \frac{16}{\Delta^2}$
  - $\alpha = 5\%$, power 90%: sample size per group is $n \approx \frac{21}{\Delta^2}$

.164

Example RCT

- Test of a blood pressure lowering drug versus placebo
- Primary endpoint: Lowering of blood pressure one month after randomization (= date of intake of the drug)
- Assumptions:
  - Data in both arms are approximately normally distributed with known $\sigma = 10$mm Hg
  - We want to have sufficient power to detect a 5 mm Hg larger reduction in the intervention arm compared to placebo ($\delta = 5$)
  - $\alpha = 5\%$, 90% power
  - How many patients required?
- Formula gives $\Delta = 0.5 \rightarrow n = 84$ per group. In total, 168 patients need to be randomized

.165

Sample size calculation (binary outcome)

- Assume (hypothesize)
  - Probability of the outcome in the control group: π_1
  - Probability of the outcome in the intervention group: π_2
- Choose significance level α and power
- n per group and $\alpha = 5\%$:

π_1	π_2	n(80%power)	n(90%power)
0.05	0.1	435	582
0.1	0.2	199	266
0.2	0.4	82	*109*
0.3	0.6	42	56
0.05	0.15	141	188
0.2	0.3	294	392
0.3	0.4	356	477
0.4	0.5	388	519
0.05	0.25	49	65
0.1	0.3	62	82
0.3	0.5	93	124
0.4	0.6	97	130

.166

Example (RCT)

- Comparison of a novel chemotherapy compared to the standard of care in cancer
- Primary endpoint: Proportion of patients showing a complete tumor response
- Assumptions: $\pi_1 = 0.2$, $\pi_2 = 0.4$
  Choose $\alpha = 5\%$, power 90%
- Table gives $n = 109$ per group $\rightarrow N = 218$ in total

.167

Typical sample size justification in a study protocol

- The study is powered for the primary endpoint, i.e. the proportion of patients showing a complete tumor response in each study arm
- Based on earlier studies, about 20% of patients achieve a complete tumor response in the control arm
- An increase in the primary endpoint by 20% (from 20% to 40%) due to the intervention was judged as both realistic and clinically relevant
- 218 patients (109 per arm) are required to detect such an increase with 90% power at the two-sided 5% significance level
- To account for some potential loss-to-follow-up and protocol violations, a total of 240 patients (120 per arm) will be randomized.

.168

Discussion

- Determination of sample size critical issue in study planning and essential part of protocol (including justification of the assumptions on which the calculation is based)
- However, true effect size is uncertain at best (from earlier studies), or completely unknown in worst case
- Clinical and feasibility considerations also affect the final sample size of a study
- Sample size formulae exist for many situations/tests and there exists specialized software
  - Basic calculations in R (`power.t.test, power.prop.test`)
  - R task view on power and sample size calculations
  - `https://github.com/vubiostat/ps`
  - via simulation (see practical)

.169

Part V
Analysis and Reporting

In the previous part we showed the PLOS NTD criteria for reviewers with respect to the methods used. We now show the criteria with respect to presentation of results and whether the conclusions drawn from the results make sense. We discuss some of them, mostly those that refer to proper statistical analysis and reporting of results.

Criteria for reviewers (PLOS neglected tropical diseases)

Results:

- *Does the analysis presented match the analysis plan?*
- *Are the results clearly and completely presented?*
- *Are the figures (Tables, Images) of sufficient quality for clarity?*

Conclusions:

- *Are the conclusions supported by the data presented?*
- Are the limitations of analysis clearly described?
- Do the authors discuss how these data can be helpful to advance our understanding of the topic under study?
- Is public health relevance addressed?

.170

17 Which variables to include in our regression model

The decision which variables to include in your model depends on the type of study question, in combination with your knowledge of the subject matter.

Which variables to include?

- *Use expert knowledge*
- Inferential: any variable that is possibly related to the outcome
- Predictive: any variable that helps better predicting/diagnosing the outcome
  - does not need to have a direct causal relation with outcome
  - proxy variable that is easy and cheap to measure may be preferred
- Causal
  - observational design: correct for confounders to reduce, or ideally eliminate, bias
  - randomized design: no additional variables need to be included to reduce bias.
    However, power can be increased by including independent baseline variables that are strongly related to the outcome
    - may explain part of the variation in outcome
    - there may be random imbalance (especially in small trials)

.171

In the Covid-19 Recovery trial they corrected for age group because of random imbalance.

Dexamethasone in Hospitalized Patients with Covid-19

Recovery Trial; imbalance in important predictor of outcome.

Through the play of chance in the unstratified randomization, the mean age was 1.1 years older among patients in the dexamethasone group than among those in the usual care group. To account for this imbalance in an important prognostic factor, estimates of rate ratios were adjusted for the baseline age in three categories (<70 years, 70 to 79 years, and ≥ 80 years).

.172

17.1 Variable reduction

We already discussed stepwise regression when we introduced multivariable linear regression. He we repeat the main chatacteristics and its use, taking into account the type of study question.

Stepwise regression

- Forward
  1. Fit regression models for each variable separately. Select the one with smallest p-value
  2. Fit regression models with two variables: the one selected in 1. and one-by-one each of the other ones. Select the one with smallest p-value amongst the other ones
  3. Repeat this until all variables not in the model are not significant when added
- Backward
  - Start with model including all variables
  - Eliminate one-by-one based on largest p-value until all p-values smaller than some value
- Combination: include all variables from univariate analyses with p-value < 0.1 (or some other number) in multivariable model and perform backward regression until all $p < 0.05$

.173

Stepwise regression and study question

- Inferential
  - often parsimonious model chosen that only includes statistically significant variables
  - ok to obtain first idea of relationships, but results may be hard to interpret
  - example: risk factors for STI's in sex workers in China
- Predictive: ok, but validation very important
- Causal: not recommended
  - Choice should not be determined by p-value, but by veriable being confounder or having strong relationship with outcome

The following table (vd Hoek *et al.*, 2001) shows results from an inferential analysis that used a combination of forward and backward stepwise regression. In the forward stepwise regression, variables with a p-value < 0.10 in the univariable[7] analysis were included in a multivariable model. All others (the blanks in the table) were not analysed further. Next, backward stepwise regression was performed. Variables that became insignificant in multivariable analysis were presented as "NS" with the direction of the effect.

Such a presentation of results is parsimonious and concise, but not necessarily close to reality. For example, ORs for a variable like age are only shown for Chlamydia, but that does not imply that the prevalence of the other STI's does not depend on age. Significance also depends on the prevalence itself, and syphilis and HIV are much less common than Chlamydia. It may even be that their OR's are larger, but not significant due to lack of power.

Table 2. Multivariate analyses of the relationship between general and sexual characteristics, HIV/sexually transmitted disease knowledge and self-efficacy and sexually transmitted diseases among 966 sex workers at Guangzhou, China, March 1998–October 1999.

	HIV OR (95% CI)	Gonorrhoea OR (95% CI)	Trichomoniasis OR (95% CI)	Chlamydia OR (95% CI)	Syphilis[a] OR (95% CI)
General					
Age (years)					
< 21				2.0 (1.2–3.4)	
21–22				1.2 (0.7–2.2)	
23–25				1.4 (0.8–2.5)	NS (–)
26–30				1.0 (0.6–1.7)	
> 30				1	
Not always lived in Guangzhou	0.2 (0.0–0.8)			NS (+)	NS (–)
Injected drugs (since 1990)	8.0 (2.1–30.3)		2.5 (1.3–5.0)	0.2 (0.1–0.4)	NS (+)
Recruitment clients on street/via pimps	NS (+)	NS (+)	NS (+)		2.6 (1.7–4.0)
No regular salary	NS (+)	2.3 (1.3–4.1)	6.1 (3.5–10.4)		
STD check-up (past 12 months)		0.4 (0.2–0.9)		NS (–)	NS (–)
Sexual activity					
Number of clients (per week)					
< 6	1				1
6–7	0.1 (0.0–0.7)	NS (+)	NS (+)		1.7 (1.0–3.0)
> 7	0.2 (0.0–1.2)				2.2 (1.3–3.7)
Duration of prostitution (years)					
< 1					1
1				NS (–)	1.5 (0.9–2.7)
3					2.5 (1.4–4.6)
> 3					1.8 (0.9–3.4)
Steady partner (past 12 months)					
No steady partner					1
1 steady partner			NS (–)		2.1 (1.4–3.2)
> 1 steady partner					1.1 (0.6–2.3)
Knowledge					
Knowledge about AIDS		NS (–)	NS (–)	NS (–)	0.9 (0.8–1.0)
Knowledge about condom use		NS (–)	0.9 (0.7–1.0)	0.8 (0.8–0.9)	NS (–)
Condom use during vaginal sex (past 2 months)					
Always		1		1	1
Frequently		3.1 (3.2–8.3)		2.1 (1.0–4.9)	1.5 (0.9–2.6)
Rarely		8.6 (3.2–23.3)	NS (+)	2.2 (0.7–2.2)	2.5 (1.3–4.6)
Never		9.6 (3.0–30.4)		2.7 (1.5–7.3)	3.8 (1.7–8.7)
Diagnostic evidence of:					
HIV			5.0 (1.4–17.0)		5.7 (1.6–20.7)
Gonorrhoea			2.3 (1.3–4.1)	2.9 (1.7–4.7)	
Trichomoniasis	11.2 (2.9–42.7)	1.8 (1.0–3.3)		2.6 (1.6–4.1)	NS (+)
Chlamydia		3.0 (1.8–5.0)	2.8 (1.7–4.4)		

CI, Confidence interval; OR, odds ratio; STD, sexually transmitted diseases.
In building multivariate models, all univariate predictors (with *P* value < 0.10) were included in a stepwise backward procedure. All overall *P* values included in the multivariate models and presented in the table are $\leqslant 0.05$.
A blank cell indicates that the variable was not included in the multivariate model ($P < 0.10$ in univariate analyses).
NS indicates that the variable after inclusion in the multivariate model was no longer statistically significant, although in univariate analyses the model *P* value was < 0.10. The (+) or (–) behind the 'NS' indicates the direction of the risk estimate in univariate analyses.
[a] *Treponema pallidum* haemagglutination assay positive.

[7] I use the word univariable, but more often the term univariate is used.

Drawbacks stepwise regression in inferential research

- Excluded covariables may in reality have effect (power): "absence of proof" is not "proof of absence"
  Small effects would be included if sample were much larger
- Repeated testing, such that risk of spurious significant results increases
- Based on methods that were intended to be used to test *prespecified* hypotheses
- Stepwise methods are often a complicated equivalent to throwing darts blindfolded (the final model is more due to random chance than anything else)
- See also What are some of the problems with stepwise regression?

.176

17.2 Multiple testing and fishing expeditions

Example: Aspirin trial (Lancet 1988; 2: 349–60)

- A RCT in patients with a cardiac infarction showed a highly significant survival benefit of aspirin therapy vs. placebo
- Prior to publication, the editors of the Lancet asked for 40 additional subgroup analyses
- The authors agreed under the condition that they can chose any additional characteristic themselves

.177

Example: Aspirin trial (II)

- The authors chose the zodiac sign
- Gemini and libra showed a slightly higher mortality in the aspirin arm compared to placebo
- For all other zodiac signs, aspirin was highly significantly superior
- "All these subgroup analyses should, perhaps, be taken less as evidence about who benefits than as evidence that such analyses are potentially misleading"

.178

What's the problem?

- Assume we perform significance tests of N independent null hypotheses which are all true
- Probability that a specific null hypothesis is falsely rejected: 0.05
- Probability that at least one of the null hypotheses is falsely rejected: $1 - P(\text{none rejected}) = 1 - 0.95^N$

N	5	10	20	50
$1 - 0.95^N$	23%	40%	64%	92%

- Similar problems occur for general (possibly dependent) multiple significance tests

.179

Instances of multiple testing

- Multiple testing often occurs
  - Subgroup analyses
  - Several endpoints
  - Pairwise comparisons of > 2 subgroups
  - Interim analyses
  - Data dredging
- What to do to cope with multiple testing
  - Pre-define primary study hypothesis and all important secondary hypotheses
  - Avoid unplanned fishing expeditions for significant results
  - Separate confirmatory and exploratory/inferential analyses
    - You cannot generate a hypothesis and statistically "prove" it based on the same data
  - May use statistical adjustments for multiple testing
    - Most simple (but very conservative): Bonferroni adjustment
    - Slightly better: Holm correction

.180

18 How to include variables

18.1 Stratified analysis

Instead of including an interaction term, often a stratified analysis is performed. A stratified analysis is performed if the effect of covariables is quantified separately for each value of some categorical variable such as gender. The main reason to perform a stratified analysis is simplicity; the parameters are easy to interpret. However, exactly the same results are obtained by fitting one regression model that includes gender as extra covariable and that allows the effect of all other covariables to be modified by gender via interaction terms. Using one single model with interaction terms has several advantages. We can *test* whether the effect of the covariables differs by gender. And if we want to *assume* the effect of a covariable to be the same for both genders, we can remove the interaction term from the model. This increases power if the assumption is correct. An example is in the next slide.

Example stratified analysis

HIV Prevalence and Associated Risk Factors among Individuals Aged 13-34 Years in Rural Western Kenya

Table 2. Age-group adjusted risk factors associated with HIV infection by gender among sexually active participants, N = 1202.

	Females					Males				
	N	Weighted HIV prevalence (%)	Age group aOR[a]	95% CI[b]	*P-value*	N	Weighted HIV prevalence (%)	Age group aOR	95% CI	*P-value*
Total	619	27.7	4.2	[2.7, 6.6]	<0.001	583	14.1	11.8	[6.4, 21.6]	<0.001
Age (years), overall median = 21	median = 22					median = 20				
IQR[c] = [18,28]	IQR = [18,28]					IQR = [17,26]				
Demographic characteristics										
Education										
Some primary school	316	30.1	*ref*[d]	–	NS[e]	230	9.5	*ref*	–	NS
Completed primary school	186	27.6	0.8	[0.5, 1.0]		190	17.6	1.1	[0.6, 2.0]	
Beyond primary school	117	22.2	0.6	[0.5, 1.2]		163	16.4	1.1	[0.7, 2.4]	
Occupation										
Has cash income[f]	213	40.7	2.1	[1.4, 3.1]	<0.001	215	22.3	1.2	[0.7, 2.0]	NS
Religion										
Muslim groups/other	117	36.2	*ref*	–	NS	109	11.2	*ref*	–	NS
Protestant groups	374	25.9	0.8	[0.4, 1.4]		321	16.4	1.6	[0.9, 3.0]	
Catholic groups	128	25.2	0.8	[0.5, 1.2]		153	11.1	1.0	[0.5, 2.1]	
Marital status										
Never married	224	8.6	*ref*		<0.001	369	5.6	*ref*	–	<0.05
Currently married	328	29.4	3.4	[1.5, 7.7]		189	32.4	2.4	[1.1,5.3]	
Divorced/Separated	6	36.2	5.0	[0.8, 21.0]		16	31.3	1.7	[0.5, 5.9]	
Widow/Widower	61	77.8	28.5	[10.6, 76.5]		9	49.7	5.2	[1.4, 19.7]	

.181

Why do a stratified analysis?

Because it is easier.

But,

same and more can be obtained by including stratum variable (e.g. gender) in one overall model:

- If relationship of variable with outome differs by gender:
  - Include interaction between gender and variable
  - Advantage: we can test and obtain p-value for interaction term
- If relationship of variable with outome does not differ by gender:
  - we can restrict parameter to be equal for both genders
  - advantage: more power

.182

18.2 Dichotomania

In many analyses, numeric variables are cetogorized into subgroups. This is almost always a bad idea.

Dichotomania

http://www.senns.uk/Geep.htm

Dichotomania is an obsessive compulsive disorder to which medical advisors in particular are prone... Show a medical advisor some continuous measurements and he or she immediately wonders. "Hmm, how can I make these clinically meaningful? Where can I cut them in two? What ludicrous side conditions can I impose on this?"

Dichotomization of continuous information tends to encourage dichotomous thinking, this limits the research questions we can ask and the conclusions we're able to draw. There is a literary canon decrying this practice written by statisticians (unfortunately it seems it is only read by statisticians).

.183

Problems Caused by Categorizing Continuous Variables

- It assumes the relationship between the predictor and the response
  - is flat within intervals
  - abruptly changes as interval boundaries are crossed
- Researchers seldom agree on the choice of cutpoint. Some may compare blood pressure > 140 with ≤ 140, while others compare > 120 with ≤ 120.
- A patient does not report to her physician "my blood pressure exceeds 140" but rather reports 142 mmHg. The risk of stroke for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.
- Loss of power and precision

.184

19 The role of p-values

When and how to use p-values

- P-value linked to hypothesis about population characteristic
  - Do not use p-value in baseline table
- Significance level dichotomy is not a gold standard for statistical inference
  - Draw conclusions based on several considerations
- Report all p-values in detail, unless they are very small (≥ 0.01 or ≥ 0.001)

.185

Usually the baseline table summarizes variables by subgroup. Sometimes the difference is quantified via a p-value. This practice is to be avoided if the differences are specific for this study and cannot be extrapolated to some larger population.

Do not use p-values in baseline table

Table 1. Demographic Data and Postoperative Rescue Medications

	Group D (n = 30)	Group DP (n = 30)	P value
Sex (M/F)*	16/14	13/17	0.438
Age (yr)†	8.1 ± 3.4	10.0 ± 3.9	0.052
Wt (kg)†	27.3 ± 11.5	34.7 ± 15.9	0.042
Ht (cm)†	127.5 ± 21.4	133.7 ± 20.5	0.257
Op. duration (min)†	119.3 ± 19.3	113.0 ± 26.9	0.299
Anes. duration (min)†	161.3 ± 23.3	165.1 ± 26.5	0.554
Recovery time (min)†	14.3 ± 8.4	12.1 ± 10.5	0.382
Postoperative analgesics/antiemetic			
Fentanyl consumption (μg/kg)†	10.7 ± 2.6	11.1 ± 2.0	0.534
Rescue analgesic needed (no. of patients)‡	14 (46.7%)	8 (26.7%)	0.180
Rescue antiemetic needed (no. of patients)‡	14 (46.7%)	13 (43.3%)	1.0

Data are mean ± SDM unless addressed. Statistical analyses were performed using *Chi-square test, †Student t-test, or ‡Fisher's exact test. Group D and DP represent dexamethasone only treated- and dexamethasone and propofol treated- patients, respectively.

- http://dx.doi.org/10.4097/kjae.2013.64.2.127
- *What is the null hypothesis?*
  There is no larger population that this table refers to
- P-values make no sense; just a description of the study population

.186

Too often, conclusions are completely guided by the p-value. If the p-value is below 0.05[8], we reject the null hypothesis and conclude that there is an effect.

[8] or equivalently if the 95% confidence interval does not cover the value of the effect measure that relates to "no effect"

We know that sometimes (at most 1 out of every 20 studies, because $1/20 = 0.05$), we conclude that there is an effect while in reality such an effect is non-existent (type I error). If the p-value is above 0.05, we tend to say that there is no effect. However, it may be that there is an effect, but it is too small to be detected by the small or moderate sample size of our study (the study is underpowered; type II error).

Instead of focusing on the type I error (which is equal to one minus the sensitivity of a test) and the type II error (which is equal to one minus the specificity of the test), it is much more important to focus on how many of the positive results (i.e. H_0 rejected) are false (which is one minus the positive predictive value). Even if a test has a fairly high sensitivity and specificity, positive predictive value can be low if the prevalence of the outcome is low. We translate this to the field of hypothesis testing. If we look at all statistically significant results at $\alpha = 5\%$ and make an effort to quantify how many reflect true effects, which is the PPV, conclusions are much worse than you might expect.

SENS, SPEC, PPV, NPV

- SENS: probability to test positive given disease
- SPEC: probability to test negative given disease free
- PPV: probability to have disease given positive test
- NPV: probability not to have disease given negative test

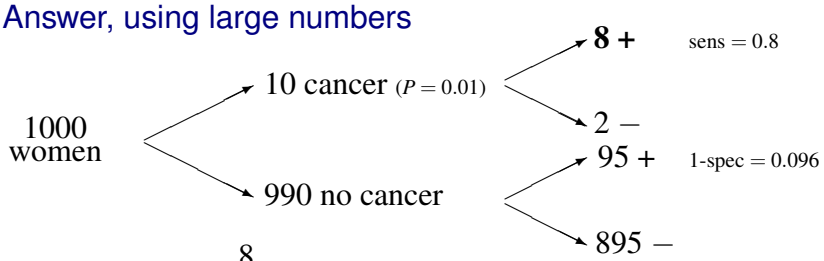
.187

Positive predictive value

- Woman aged 40 diagnosed with breast cancer by mammography
  Rare in women aged 40: $P(\text{cancer}) = 1\%$
- Mammography is not perfect *(+ or −: result from mammography)*
  - 20% false negative: $P(+|\text{cancer}) = 80\%$ (sensitivity)
  - 9.6% false positive: $P(-|\text{no cancer}) = 90.4\%$ (specificity)
- What is the probability that she has breast cancer?
  She needs to know $P(\text{cancer}|+)$

.188

Answer, using large numbers



$$P(\text{cancer}|+) = \frac{8}{95+8} = 0.078$$

$8+95 = 103$ test positive; 8 of these have cancer *Positive predictive value (PPV)* $= 7.8\%$

.189

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

Example: Suppose we perform 1000 comparable experiments. Assume H_0 holds in 90% of them (no effect). Assume level of significance 5% and power 50%

Result in experiment	H_0 true (no effect)	H_0 false (effect)	Total
H_0 not rejected	855	50	905
H_0 rejected	45	50	95
Total	900	100	1000

In only 50 out of 95, rejection of null hypothesis is correct!!
Do not test randomly, first think about biological plausibility
(statistical program is not a p-value generator)

.190

In this table, we made certain choices and assumptions with respect to significance level (5%), how many hypotheses refer to existing effects (10%) and the power of the studies (50%). The significance level is chosen by the investigator, but for some reason in most medical research it is set at 5%. We do not know how many hypotheses reflect real effects and the power of a study. In slide 191 we investigate how the percentage of false positive results depends on these choices and assumptions.

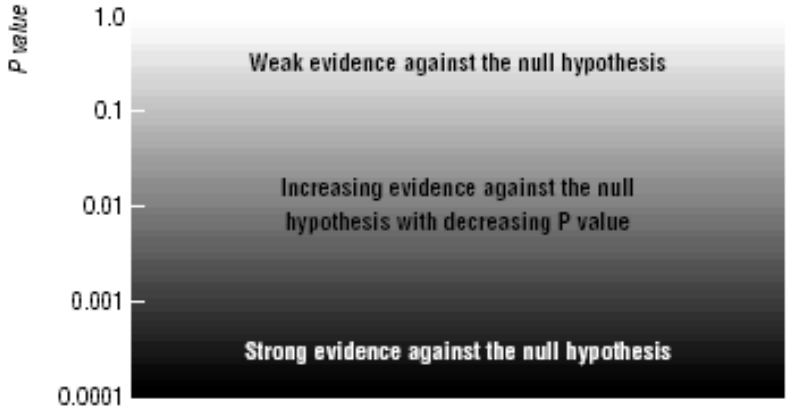
If the significance level is lowered to 0.001, the probability of a false positive result is only 1.8%. Another important observation is that things improve if we perform better studies, i.e. if more hypotheses reflect true effects and if the power of the studies increases. For example, at a significance level of 5% (first column), if 80% of the ideas is correct and if the power of the studies increases to 80%, only 1.5% of the significant effects are false. Increasing the power of studies may be difficult due to financial or logistic contraints. But thinking about relevant and plausible hypotheses, and not just performing an analysis because we have the data, is something we can imply by using the power of our brains. Testing too many hypotheses comes at a cost; statistics is not a magic free lunch box from which we can obtain significant results until we are satisfied.

Power of study (proportion (%) of time we reject null hypothesis if it is false)	Percentage of "significant" results that are false positives		
	P=0.05	P=0.01	P=0.001
80% of ideas correct (null hypothesis false)			
20	5.9	1.2	0.10
50	2.4	0.5	0.05
80	1.5	0.3	0.03
50% of ideas correct (null hypothesis false)			
20	20.0	4.8	0.50
50	9.1	2.0	0.20
80	5.9	1.2	0.10
10% of ideas correct (null hypothesis false)			
20	69.2	31.0	4.30
50	47.4*	15.3	1.80
80	36.0	10.1	1.10
1% of ideas correct (null hypothesis false)			
20	96.1	83.2	33.10
50	90.8	66.4	16.50
80	86.1	55.3	11.00

*Corresponds to assumptions in table 2.

.191

$p = 0.05$ is to large extent an arbitrary choice



Suggested interpretation of P values from published medical research

.192

Statement American Statistical Association

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an

.193

Some ASA statements

- The widespread use of "statistical significance" (generally interpreted as $p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.
- Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.
- Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

.194

19.1 Examples

We already saw an example of the use of "NS" instead of the actual p-value in slide 175. Here is another example. Also, this paper mixes a baseline table with

outcome analysis.

Risk factors for anal carcinoma caused by HPV

- HPV: human papilloma virus
- Give p-value instead of "NS" (not significant)
- Also, we don't want to quantify how age differs by outcome, we want to know how age influences outcome (reverse causal order)

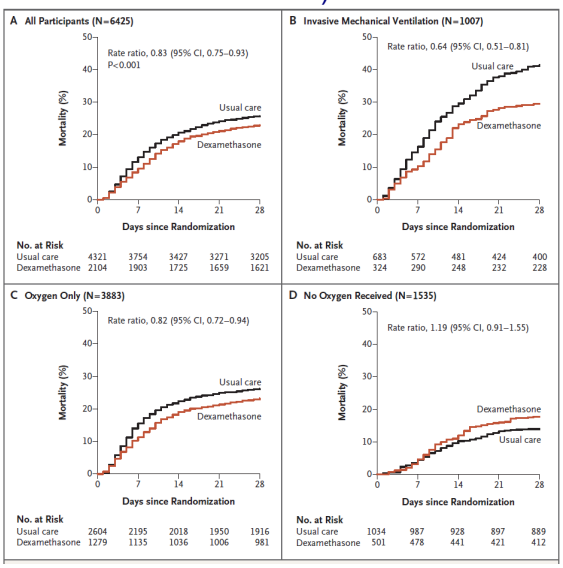
Table 1. Characteristics of patients according to high-grade dysplasia and cancer outcomes.

	With high-grade dysplasia and cancer		
	Yes	No	*P*
n	38	161	
At baseline			
Age (years) (mean ± SD)	37 ± 10	36.9 ± 11	NS
Male (%)	94	82	NS
CD4 T-cell count ($\times 10^6$ cells/ml serum) (mean ± SD)	362 ± 319	455 ± 340	NS
Langerhans' cells/mm^2 mucosa (mean ± SD)	16 ± 13	27.4 ± 24	0.01
HIV positive (%)	84	16	0.007
Oncogenic human papillomavirus subtype (%)	33	1.6	0.009
Epstein–Barr virus (%)	12	3.3	0.05
Herpes simplex virus (%)	39	12.5	0.007
Anal co-infection (%)	51	18	0.0002
During follow up			
CD4 T-cell count ($\times 10^6$ cells/ml serum)[a] (mean ± SD)	334 ± 270	360 ± 312	NS
Anal infections (mean ± SD)	1.8 ± 0.7	1.3 ± 0.8	< 0.01
Relapses (mean ± SD)	2.4 ± 0.6	1.9 ± 0.8	0.05
Langerhans' cells/mm^2 mucosa[a,b] (mean ± SD)	7 ± 6	19 ± 9	0.001
Oncogenic human papillomavirus subtype (%)	29	2.2	< 0.001

.195

In the well-known RECOVERY trial they looked at the effect of dexamethasone in three severity groups. Note that the hazard ratio of 1.19 towards harm "no oxygen" group is almost as large as the beneficial effect of 0.82 in the "oxygen only" group (1/0.82=1.22). However, they largely ignore this in the presentation of their results, probably because it had a p-value > 0.05. Note that the number of patients differed by severity group, which makes it harder to obtain a low p-value in the "no oxygen" group and we certainly cannot exclude the possibility that dexamethasone is harmful in that group.

Recovery (dexamethasone in Covid-19) trial



In patients hospitalized with Covid-19, the use of dexamethasone resulted in lower 28-day mortality among those who were receiving either invasive mechanical ventilation or oxygen alone at randomization but not among those receiving no respiratory support.

.196

Another example is from an RCT that compared azithromycin and trimethoprim-sulfamethoxazole (SXT) in patients with undifferentiated febrile illness [9]. The hypothesis was formulated, but has an inconsistency that may be due to mixing up population characteristics and the hypothesis testing procedure.

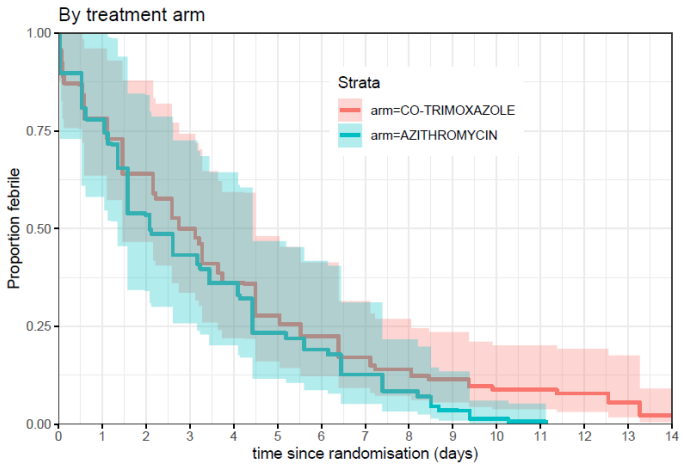
SXT versus azithromycin in patients with undifferentiated febrile illness

- *We hypothesized that azithromycin is superior to SXT for UFI treatment, but the drugs are non-inferior to each other for culture-positive enteric fever treatment*
  culture-positive: Salmonella Typhi or paratyphoid fever
- *What is the null hypothesis?*
  What is the population?
- There are two hypotheses, one for overall population of patients with UFI and the other for the culture-confirmed subgroup
- If superior for overall population, but non-inferior for culture-positive subgroup, then it must be superior for the culture-negative subgroup

.197

SXT versus azithromycin in patients with UFI: results

Despite similar fever clearance time in the two arms (primary outcome, P-value: 0.059), significantly fewer complications and relapses make azithromycin a better choice for empirical treatment of UFI in Nepal



.198

My suggestion

We found moderate evidence that seven days of azithromycin is more effective than 7 days co-trimoxazole for the treatment of all cause UFI in Nepal.

Comment by reviewer: *The claims about differences in fever clearance times for all-cause fever in the SXT group compared with the azithromycin group are overstated. These differences were not statistically significant. The language should be revised for a more honest rendering of the main findings.*

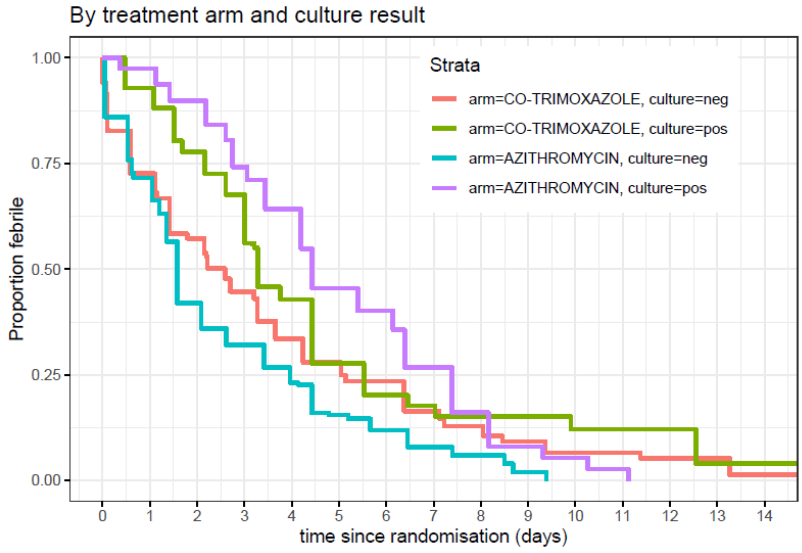
.199

[9] http://doi.org/10.1093/cid/ciaa1489

Note that indeed we found a clear indication of a shorter fever clearance time in the culture-negative group.

By culture

P-value: 0.024 in culture negative; 0.81 in culture positive



.200

20 Reporting

Reporting

- Report units for measurements (also in axis labels)
- Report N's (denominators) and the amount of missing data
- For effect estimates, report estimates and CI in addition to p-values
- Use reasonable precision for reporting, no "fake precision"
  - Usually, 2 decimals for p-values are enough, except if they are very small (e.g. $p \leq 0.0001$)
  - For %: usually, full numbers or max 2 decimals are enough
    - Good example: "17 (74%) of 23 patients had a clinical response"
    - Bad example: "17 (73.913%) of 23 patients…"

.201

Think about rescaling of numeric variable

- Parameter is measure per unit increase
- Choose informative unit. Not as in e.g.

Incidence and clearance of genital HPV infection in men

	Any HPV		Oncogenic HPV		Non-oncogenic HPV	
	Univariate	Multivariate*	Univariate	Multivariate*	Univariate	Multivariate†
Country						
USA	1·00	1·00	1·00	1·00	1·00	1·00
Brazil	1·07 (0·82–1·40)	0·93 (0·69–1·27)	0·89 (0·69–1·15)	0·81 (0·61–1·09)	1·56 (1·2–2·03)	2·04 (1·47–2·83)
Mexico	0·82 (0·63–1·08)	0·83 (0·63–1·10)	0·65 (0·49–0·86)	0·75 (0·56–1·00)	0·99 (0·75–1·31)	1·27 (0·90–1·78)
Age	1·00 (0·99–1·00)	0·99 (0·98–1·00)	0·99 (0·98–1·00)	0·99 (0·98–1·00)	1·00 (0·99–1·01)	0·99 (0·98–1·00)

- Better show age effect per ten years (or use more digits)

.202

Reporting (continued)

- Follow standard reporting guidelines
- Consort for RCTs (www.consort-statement.org)
- Strobe for observational studies (www.strobe-statement.org)
- STARD for diagnostic tests (www.stard-statement.org)
- More on https://www.equator-network.org

.203

21 Help!

Analysis and reporting

- Reserve enough time for a careful analysis
- Data quality checks, missing data, outliers
- Use descriptive and graphical analyses to understand the data
- Structured analysis strategy (following the analysis plan)
- Restrict fishing expeditions; be aware that they produce at best exploratory/preliminary evidence
- Report descriptive statistics first, then the primary analysis, then secondary analyses
- Report interesting exploratory analyses but clearly declare them as hypothesis-generating only

.204

When do you need a statistician?

- Involve a statistician as early as possible, i.e. during the design stage of the trial (not only at the analysis stage)!
- Involve a statistician for all studies that are large, have a complex design or will require a substantial amount of statistical analyses
- Complex data structures or analyses
- Longitudinal data, survival data, clustered data
- High-dimensional data
- Large amount of missing data
- Whenever you feel insecure about the correct design or analysis
- Prevent miscommunication

.205

Recommended books (entire course)

- Altman DG (1991). *Practical Statistics for Medical Research* Chapman & Hall/CRC
- Neale Batra *The Epidemiologist R Handbook*. https://epirhandbook.com/en/ Also in Vietnamese!
- Frank E. Harrell (2022). *Biostatistics for Biomedical Research* http://hbiostat.org/bbr/
- Katz MH (2006). *Study Design and Statistical Analysis* Cambridge University Press
- Kirkwood BR and Sterne JAC (2003). *Essential Medical Statistics* (2nd Edition). Blackwell Science.
- Vu J and Harrington D (2020). *Introductory Statistics for the Life and Biomedical Sciences* OpenIntro https://www.openintro.org/book/biostat/

.206