BIOSTATISTICS COURSE

INTRODUCTION TO STATISTICS AND STUDY DESIGN

PAOLA REBORA paola.rebora@unimib.it

SCHOOL OF MEDICINE AND SURGERY

SCHOOL OF MEDICINE AND SURGERY

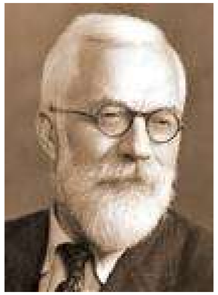
Why do you have a Biostatistics course in your degree plan?

From: *An Introduction to Medical Statistics,* Martin Bland 2015.

- .. a question many students of medicine ask as they struggle with statistics: is it worth it? As Altman (1982) has argued, bad statistics leads to bad research and bad research is unethical. Not only may it give misleading results, which can result in good therapies being abandoned and bad ones adopted, but it means that patients may have been exposed to potentially harmful new treatments for no good reason. Medicine is a rapidly changing field. In 10 years' time, many of the therapies currently prescribed and many of our ideas about the causes and prevention of disease will be obsolete. They will be replaced by new therapies and new theories, supported by research studies and data of the kind described in this book, and probably presenting many of the same problems in interpretation. The practitioner will be expected to decide for her- or himself what to prescribe or advise based on these studies. So a knowledge of medical statistics is one of the most useful things any doctor, nurse, dentist, or physiotherapist could acquire during her or his training.

Introduction

"[…] Statistics may be regarded as the study of populations, the study of methods of the reduction of data, the study of variation."



Sir Ronald Aylmer Fisher (1950)

The key aspect is the possibility to isolate **common behaviors and associations** between phenomena while controlling for or taking into account the different sources of **variability**

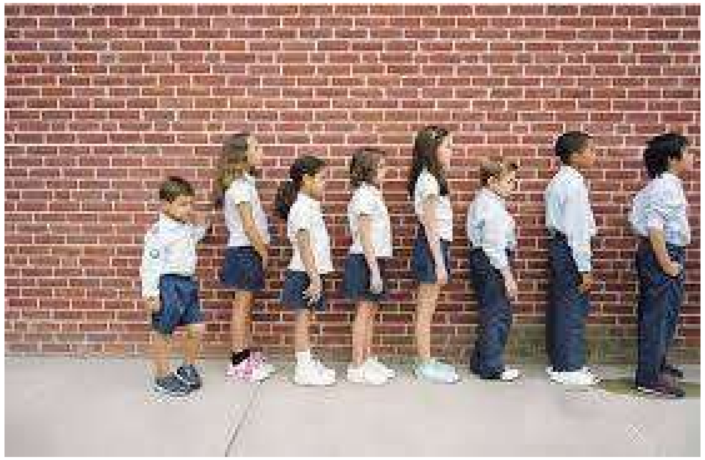
Statistical methodology is involved in all phases of scientific research:

- **Design** of the study/experiment
- Data collection and **elaboration**
- **Interpretation** of the results

Variability sources

Instrumental

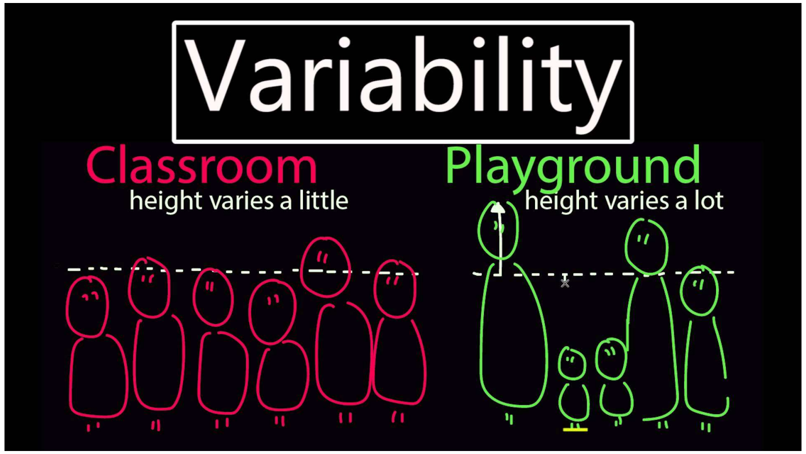
biological



Variability sources

Instrumental

biological



Instrumental variability can be completely controlled by acting on the measurement methods (optimizing the operating procedure, calibrating the instrument, training the staff).

Biological variability can only be partially limited, making the set of subjects analyzed more homogeneous.

Statistics

Statistical methodology is involved in all phases of scientific research:

1. **Design** of the study/experiment
2. Data collection and **elaboration**
3. **Interpretation** of the results

Planning/ Design of the study

- Primary question of interest
  - Quantifying information about a single group
  - Comparing multiple groups
- sample size
  - How many subject needed in total?
  - How many in each group to be compared?
- Selecting study participant
  - Randomly chosen
  - Selected from a pool of (interested) persons?
  - Take whoever shows up

  if group comparison: how to assign to groups

Data collection /data analysis

- How to best summarise information from raw data
- Dealing with variability (natural and instrumental/sampling related)
  - Important patterns in data obscured by variability
  - Distinguish real patterns from random variation
- Inference: using information from the single study coupled with info about variability to make statement about the larger population/process of interest

Presentation and Interpretation

- What summary measures will best convey the «main messagges» in the data on the question of interest
- How to convey uncertainty in estimates based on data
- What do the results mean?

Terminology: population and sample

Population (also called universe space):
Any collection of individuals in which we might be interested, where these individuals might be anything, and the number of individuals may be finite or infinite.

Sample:
a subset of the population, used to draw conclusions on the target population.

Terminology: population and sample

Population (also called universe space):
Any collection of individuals on which we might be interested, where these individuals might be anything, and the number of individuals may be finite or infinite.

Example:
-all people in Italy
-all people with diabetes
-all possible measures of blood pressure on a patient

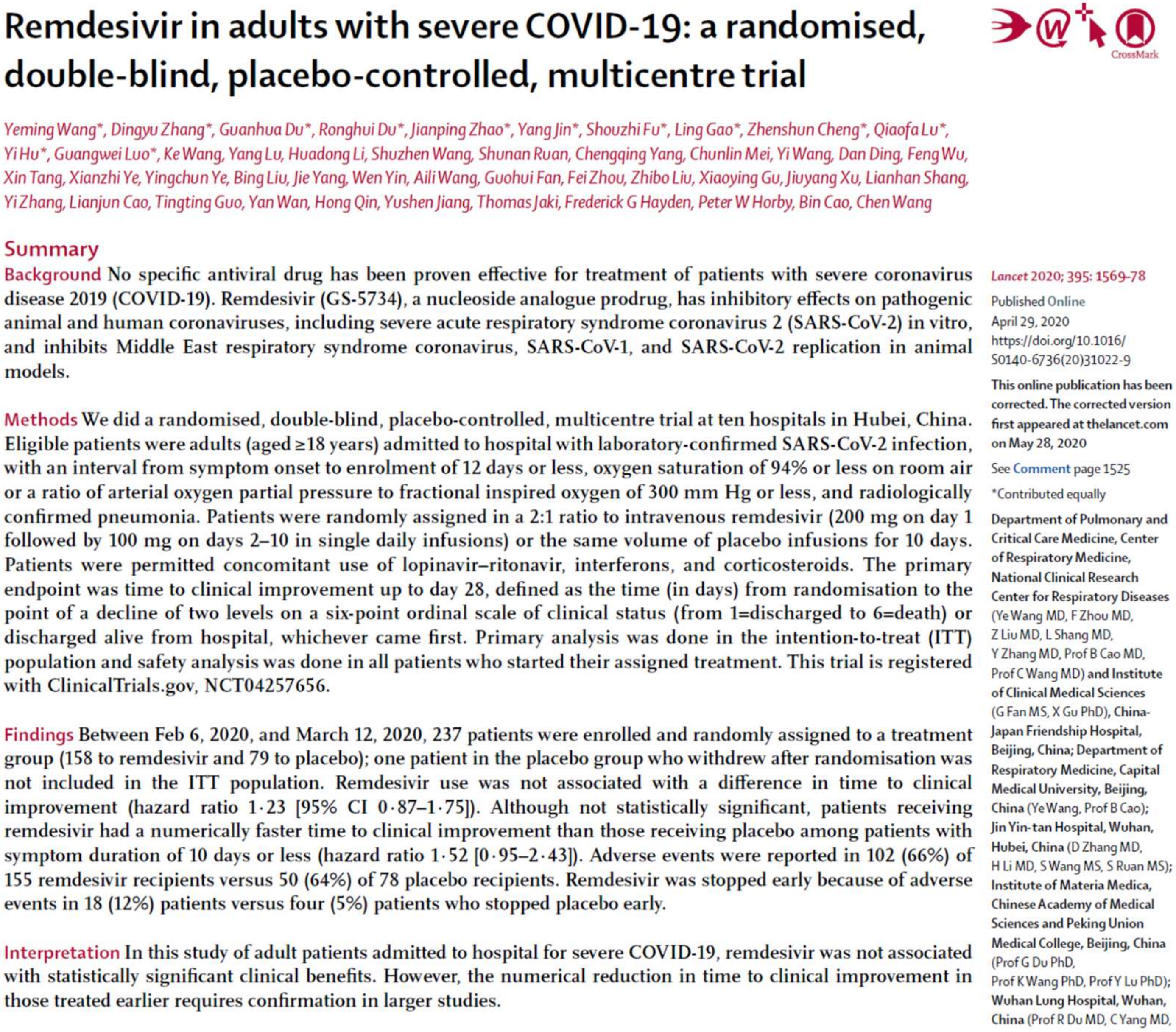
Sample:
a subset of the population, used to draw conclusions on the target population.

Terminology: population and sample

Target Population:
the population to whom we wish to generalise our findings

Study population:
the population from which we sampled, also called the "study base"

Sample:
a subset of the population, used to draw conclusions on the target population.



Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial

Yeming Wang*, Dingyu Zhang*, Guanhua Du*, Ronghui Du*, Jianping Zhao*, Yang Jin*, Shouzhi Fu*, Ling Gao*, Zhenshun Cheng*, Qiaofa Lu*, Yi Hu*, Guangwei Luo*, Ke Wang, Yang Lu, Huadong Li, Shuzhen Wang, Shunan Ruan, Chengqing Yang, Chunlin Mei, Yi Wang, Dan Ding, Feng Wu, Xin Tang, Xianzhi Ye, Yingchun Ye, Bing Liu, Jie Yang, Wen Yin, Aili Wang, Guohui Fan, Fei Zhou, Zhibo Liu, Xiaoying Gu, Jiuyang Xu, Lianhan Shang, Yi Zhang, Lianjun Cao, Tingting Guo, Yan Wan, Hong Qin, Yushen Jiang, Thomas Jaki, Frederick G Hayden, Peter W Horby, Bin Cao, Chen Wang

Summary

Background No specific antiviral drug has been proven effective for treatment of patients with severe coronavirus disease 2019 (COVID-19). Remdesivir (GS-5734), a nucleoside analogue prodrug, has inhibitory effects on pathogenic animal and human coronaviruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in vitro, and inhibits Middle East respiratory syndrome coronavirus, SARS-CoV-1, and SARS-CoV-2 replication in animal models.

Methods We did a randomised, double-blind, placebo-controlled, multicentre trial at ten hospitals in Hubei, China. Eligible patients were adults (aged ≥18 years) admitted to hospital with laboratory-confirmed SARS-CoV-2 infection, with an interval from symptom onset to enrolment of 12 days or less, oxygen saturation of 94% or less on room air or a ratio of arterial oxygen partial pressure to fractional inspired oxygen of 300 mm Hg or less, and radiologically confirmed pneumonia. Patients were randomly assigned in a 2:1 ratio to intravenous remdesivir (200 mg on day 1 followed by 100 mg on days 2–10 in single daily infusions) or the same volume of placebo infusions for 10 days. Patients were permitted concomitant use of lopinavir–ritonavir, interferons, and corticosteroids. The primary endpoint was time to clinical improvement up to day 28, defined as the time (in days) from randomisation to the point of a decline of two levels on a six-point ordinal scale of clinical status (from 1=discharged to 6=death) or discharged alive from hospital, whichever came first. Primary analysis was done in the intention-to-treat (ITT) population and safety analysis was done in all patients who started their assigned treatment. This trial is registered with ClinicalTrials.gov, NCT04257656.

Findings Between Feb 6, 2020, and March 12, 2020, 237 patients were enrolled and randomly assigned to a treatment group (158 to remdesivir and 79 to placebo); one patient in the placebo group who withdrew after randomisation was not included in the ITT population. Remdesivir use was not associated with a difference in time to clinical improvement (hazard ratio 1·23 [95% CI 0·87–1·75]). Although not statistically significant, patients receiving remdesivir had a numerically faster time to clinical improvement than those receiving placebo among patients with symptom duration of 10 days or less (hazard ratio 1·52 [0·95–2·43]). Adverse events were reported in 102 (66%) of 155 remdesivir recipients versus 50 (64%) of 78 placebo recipients. Remdesivir was stopped early because of adverse events in 18 (12%) patients versus four (5%) patients who stopped placebo early.

Interpretation In this study of adult patients admitted to hospital for severe COVID-19, remdesivir was not associated with statistically significant clinical benefits. However, the numerical reduction in time to clinical improvement in those treated earlier requires confirmation in larger studies.

Lancet 2020; 395: 1569–78

Published Online
April 29, 2020
https://doi.org/10.1016/S0140-6736(20)31022-9

This online publication has been corrected. The corrected version first appeared at thelancet.com on May 28, 2020

See Comment page 1525

*Contributed equally

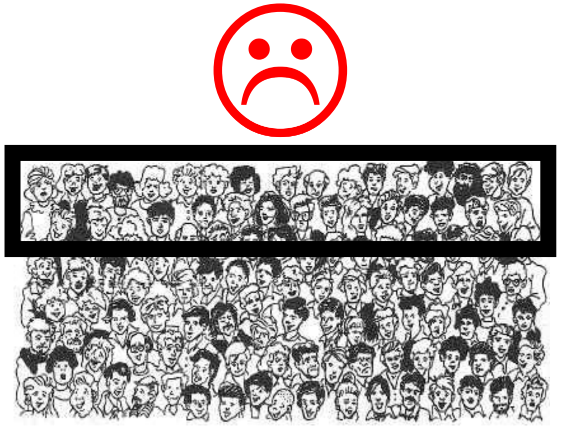
Department of Pulmonary and Critical Care Medicine, Center of Respiratory Medicine, National Clinical Research Center for Respiratory Diseases (Ye Wang MD, F Zhou MD, Z Liu MD, L Shang MD, Y Zhang MD, Prof B Cao MD, Prof C Wang MD) and Institute of Clinical Medical Sciences (G Fan MS, X Gu PhD), China-Japan Friendship Hospital, Beijing, China; Department of Respiratory Medicine, Capital Medical University, Beijing, China (Ye Wang, Prof B Cao); Jin Yin-tan Hospital, Wuhan, Hubei, China (D Zhang MD, H Li MD, S Wang MS, S Ruan MS); Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China (Prof G Du PhD, Prof K Wang PhD, Prof Y Lu PhD); Wuhan Lung Hospital, Wuhan, China (Prof R Du MD, C Yang MD,

Sample

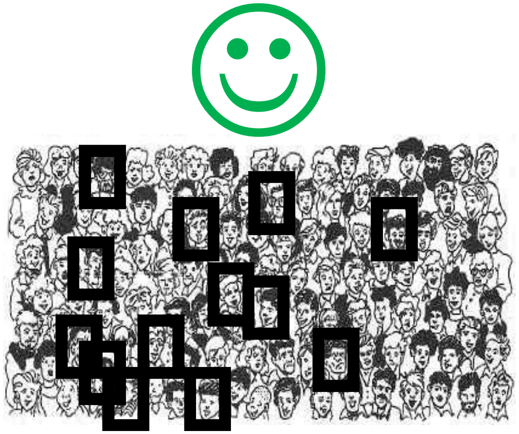
A subset of the study population, used to draw conclusions on the target population.

It has to be a probabilistic sample, possibly a random sample of the population

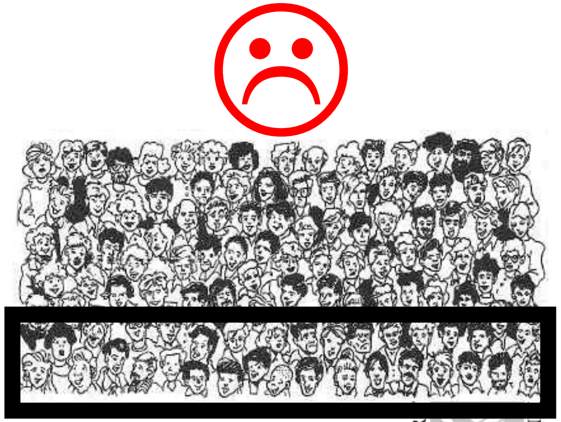
A subset of the population.



last row



random sample



1st row

Sample

A **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.

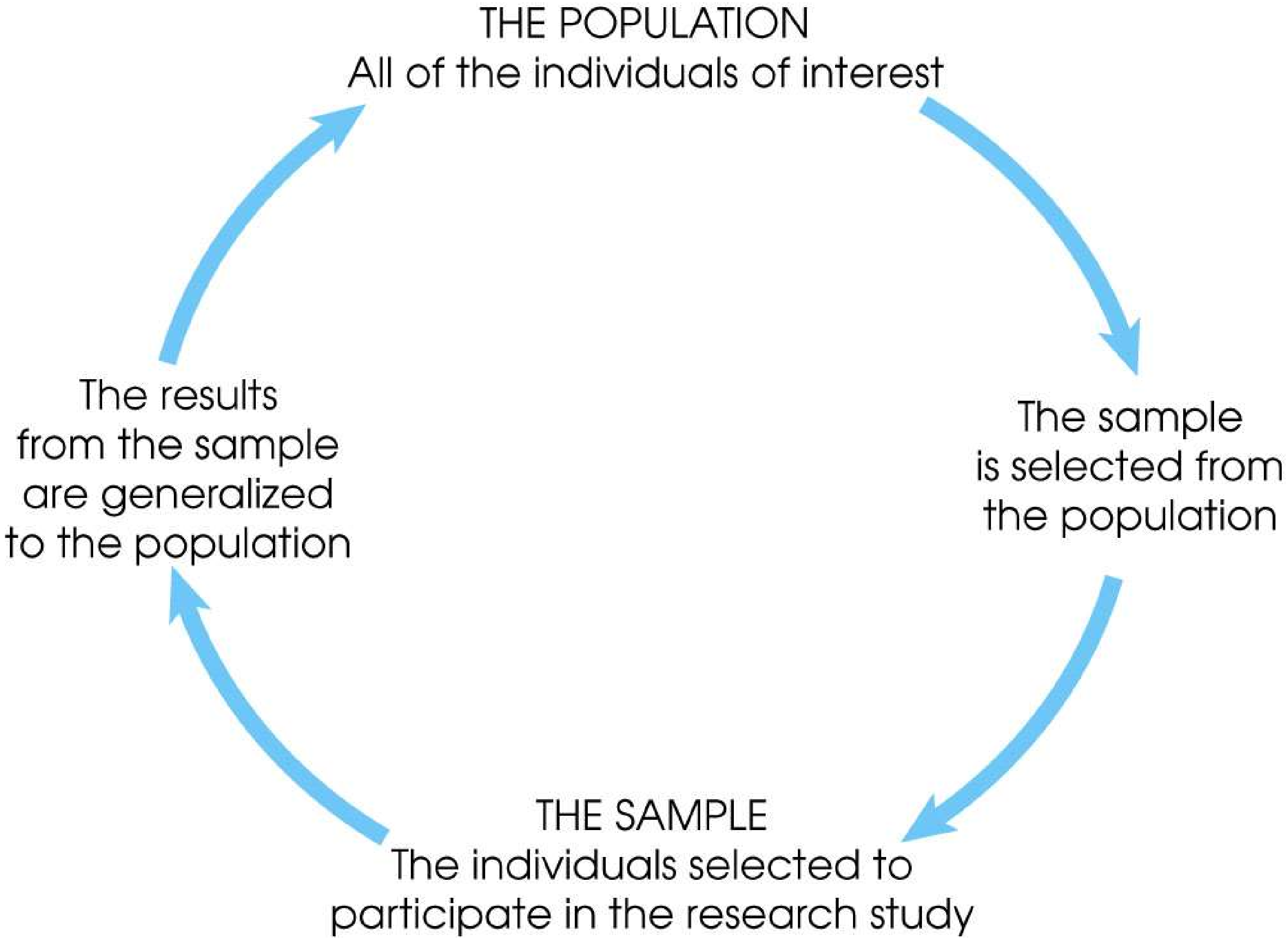
Represent the population

↓

Not verifiable condition to generalise results of the study to target population!

Example. A small number of people accurately reflect the members of an entire population. In a classroom of 30 students, in which half the students are male and half are female, a representative sample might include six students: three males and three females.

The **relationship** between a population and a sample:



THE POPULATION
All of the individuals of interest

The sample
is selected from
the population

THE SAMPLE
The individuals selected to
participate in the research study

The results
from the sample
are generalized
to the population

Study design - Key Concept

- If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.
- Method used to collect sample data influences the quality of the statistical analysis.
- Of particular importance is *simple random sample*.

What is Design?

Miettinen* (1982):

"a vision of the end product of a study on one hand
and
a scheme for carrying out a study on the other"

In biostatistics:

"end product"

a measure of occurrence or risk of an outcome (disease)

"schemes" for sampling and analysis

** Miettinen O. Design options in epidemiologic research: an update Scan. J Work and Environ Heath. 1982.*

Example

- You are involved in a project to find out if snus causes gastric ulcer.
  - A questionnaire is sent out to 300 randomly chosen subjects.
  - 200 subjects respond:

		Ulcer		
		Yes	No	
Snus	Yes	2	28	2/30≈0.07
	No	17	153	17/170=0.1

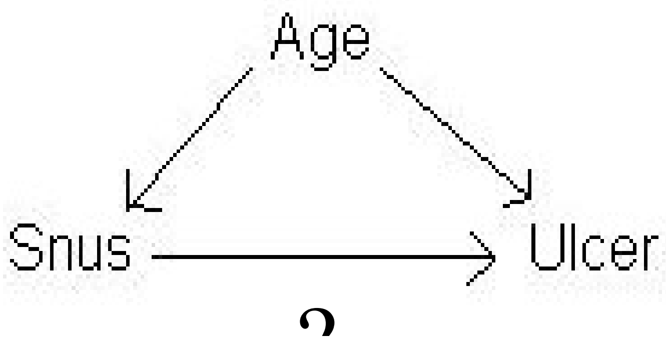
Can we safely conclude that snus prevents ulcer?

One possible explanation

- It is a wide spread hypothesis that snus causes ulcer.
- Snus users who develop ulcer may therefore feel somewhat guilty, and may therefore be reluctant to participate in the study
- Hence, result may be (partly) explained by an underrepresentation of snus users with ulcer among the responders.
- This is a case of **selection bias**.

Another possible explanation

- Because of age-trends, young people use snus more often than old people.
- For biological reasons, young people have a smaller risks for ulcer than old people.
- Hence, result may be (partly) explained by snus-users being in “better shape” than non-users.
- This is a case of **confounding**, and age is called a **confounder**



Yet another explanation

- It is a wide spread hypothesis among physicians that snus causes **and aggravates** ulcer.
- Snus users who suffers from ulcer may therefore be advised by their physicians to quit.
- Hence, results may be (partly) explained by a tendency among people with ulcer to quit using snus.
- This is a case of **reverse causation**.

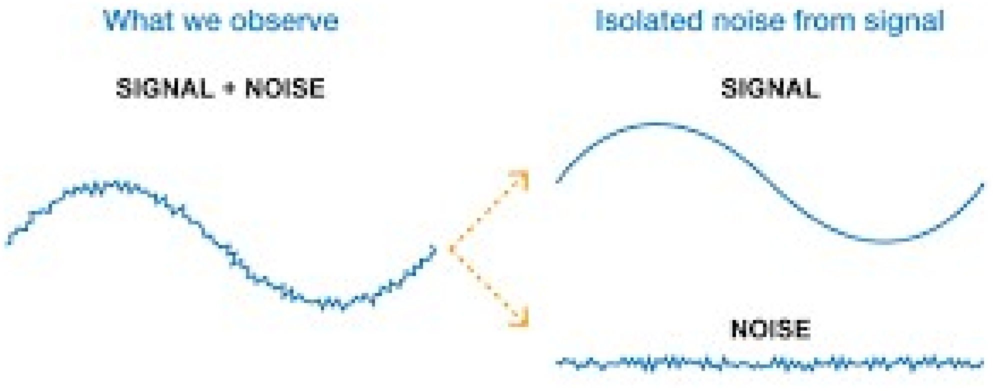
Biomedical research

Goal: to investigate a relationship between patient characteristics/treatments (exposure factor) and a health condition (outcome) through studies

EXPOSURE → OUTCOME

The relationship we are interested in is that of CAUSE and EFFECT.

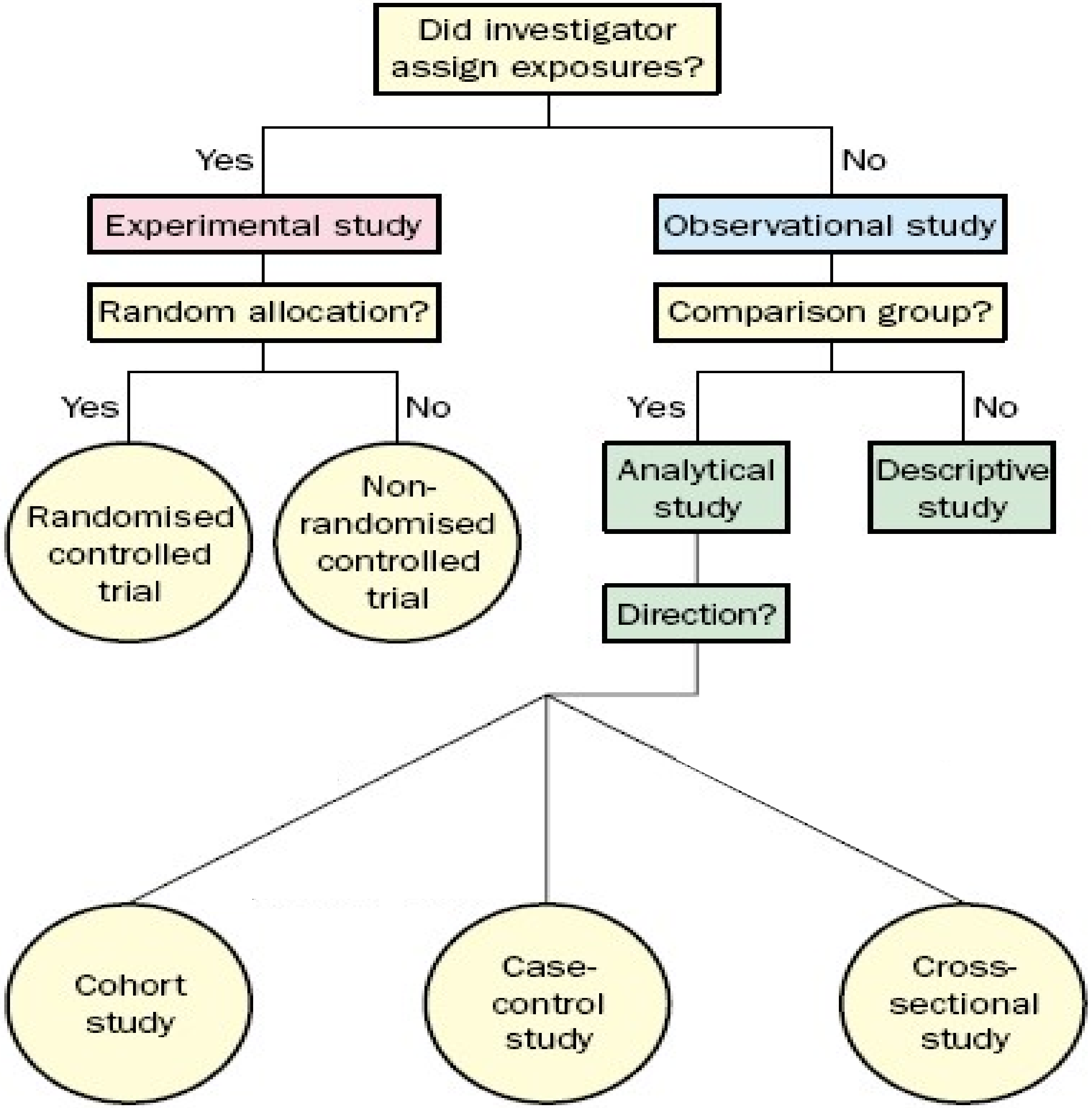
We have to distinguishing between signal and background noise



Distinctive features of a clinical or biomedical study

- The arguments, methods and conclusions are based on comparisons
- The conclusions are extended from the particular of the sample to the general of the population (inference) on the basis of a statistical-probabilistic model
- Everything is planned in detail and documented before the start of the study
- The conclusions are based on the comparison between "homogeneous" groups

Clinical research has two large “kingdoms”: Experimental vs observational studies



Did investigator assign exposures?

Yes

Experimental study

Random allocation?

Yes

Randomised controlled trial

No

Non-randomised controlled trial

No

Observational study

Comparison group?

Yes

Analytical study

Direction?

Cohort study

Case-control study

Cross-sectional study

No

Descriptive study

D.A. Grimes, K.F. Schulz, An overview of clinical research: the lay of the land. Lancet 2002; 359: 57-61

Experiment

- **Experiment**

  apply some treatment and then observe its effects on the subjects; (subjects in experiments are called experimental units)

Observational Study

- **Observational study**

  observing and measuring specific characteristics without attempting to modify the subjects being studied

Experimental studies

The design of experimental studies has two main objectives:

- **Delete** (or make negligible) the **bias** in the estimates and in the assessment of treatment effects (nonsampling error)

  ➢ This affects the **accuracy** of the results

- **Reduce** (or keep under control) the effect of the **sampling error** (random variability)

  ➢ This affects the **precision** of the results

Errors

No matter how well you plan and execute the sample collection process, there is likely to be some error in the results.

- Nonsampling error

  sample data incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)

- Sampling error

  the difference between a sample result and the true population result; such an error results from chance sample fluctuations

Experimental studies: How to avoid bias

The main strategies to avoid systematic (e.g. nonsampling) errors are:

a) Inclusion of a **control group**:
b) Randomization: random allocation of subjects to treatments
c) Blinding: subjects (and researchers) are not aware of which treatment was assigned to whom
d) Data is analyzed with the intention-to-treat principle:

How to avoid bias:

a) Inclusion of a control group

- Subjects not receiving the experimental treatment
  - ➢ Without a control group is not possible to completely ascribe the observed effects to the treatment

How to avoid bias:
b) Randomization

- Is used when subjects are assigned to different groups through a process of random selection. The logic is to use chance as a way to create two groups that are similar.

  - Prognostic factors (known and unknown) are randomly divided between arms
  - Eliminates systematic errors in assigning treatments to patients (Informative and unaware)
  - It is the most ethically acceptable way to assign patients to the compared treatments
  - Guarantees the validity of statistical tests

    ⇒ **Avoids selection bias and confounding**

How to avoid bias:

c) Blinding

- is a technique in which the subject doesn’t know whether he or she is receiving a treatment or a placebo. Blinding allows us to determine whether the treatment effect is significantly different from a placebo effect, which occurs when an untreated subject reports improvement in symptoms

➢ This avoids conscious or unconscious beliefs to influence the results of the experiment (e.g. placebo effect).

Exposure is adminstred blindly: the patient is anaware of the exposure

How to avoid bias:
Double Blind

Randomised controlled trial

Blinding occurs at two levels:

(1) *Exposure is adminstred blindly*: **The subject doesn’t know whether he or she is receiving the treatment or a placebo**

(2) *Outcome is observed blindly*: **The experimenter/assessor does not know whether he or she is administering the treatment or placebo**

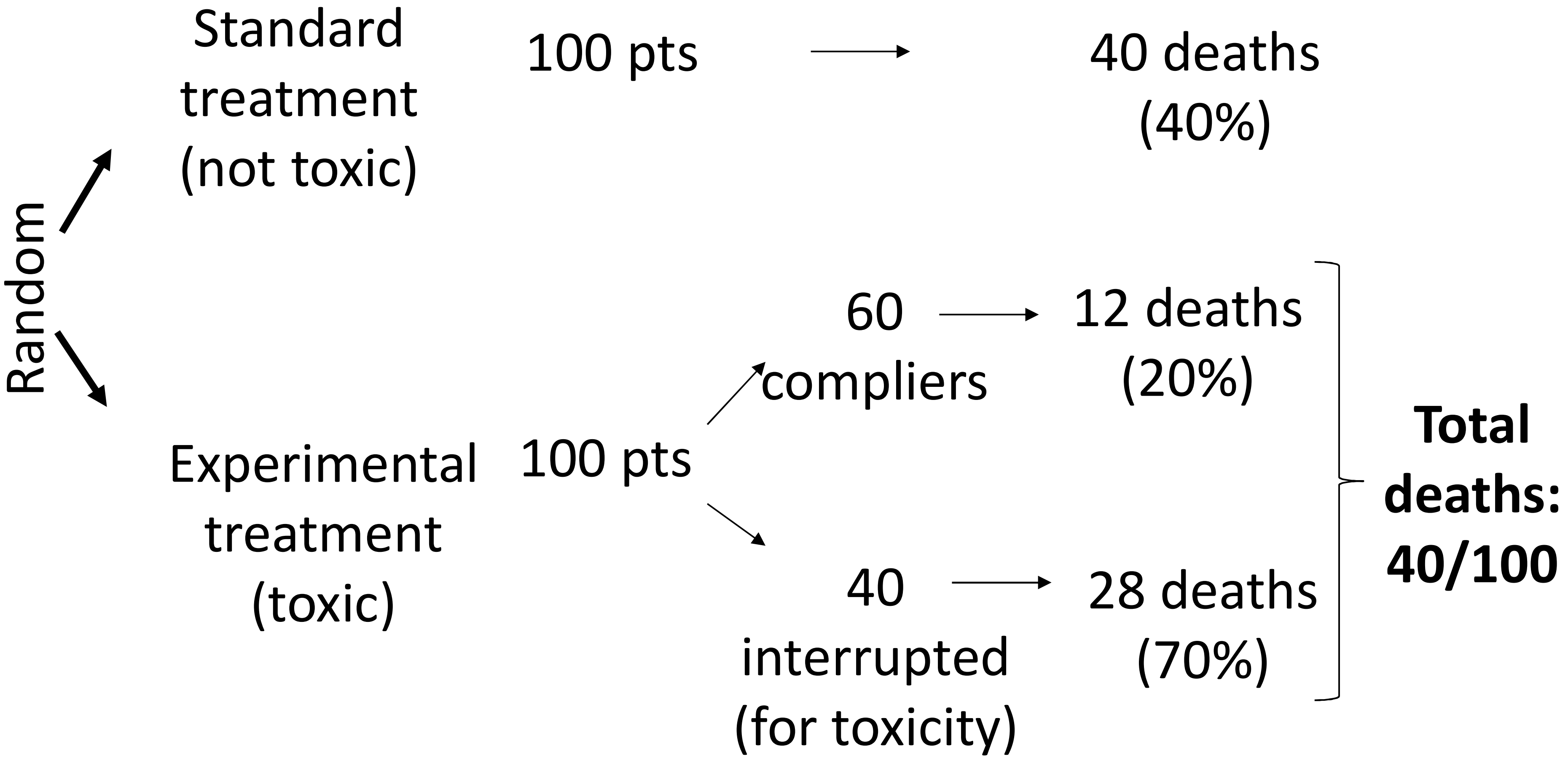
How to avoid bias:
d) intention-to-treat principle

- Data is analyzed with the **intention-to-treat principle**:

  – exposure status is that of the randomization even if exposure changes for some reason

Intention-to-treat vs per-protocol analysis

Randomised controlled trial



Random

Standard treatment (not toxic)

100 pts

40 deaths (40%)

Experimental treatment (toxic)

100 pts

60 compliers

12 deaths (20%)

40 interrupted (for toxicity)

28 deaths (70%)

Total deaths: 40/100

How to reduce the sampling error

The main strategies to control the random variability are:

- **Replication**
- **Balance**
- Use **blocks**

How to reduce the sampling error:
Replication

Is the repetition of an experiment on more than one subject. Samples should be large enough so that the erratic behavior that is characteristic of very small samples will not disguise the true effects of different treatments. It is used effectively when there are enough subjects to recognize the differences from different treatments.

- The bigger is the sample, the smaller becomes the uncertainty due to sampling in estimating the response

- Use a sample size that is large enough to let us see the true nature of any effects, and obtain the sample using an appropriate method, such as one based on randomness.

How to reduce the sampling error: balance

- **Balance:** the sample size of the treatment groups is the same

  ➢ The standard error of the estimates depends on the quantity $(1/n_1+1/n_2)$ which is minimum when $n_1=n_2$

How to reduce the sampling error: Use blocks

A block is a group of subjects that are similar, but blocks differ in ways that might affect the outcome of the experiment

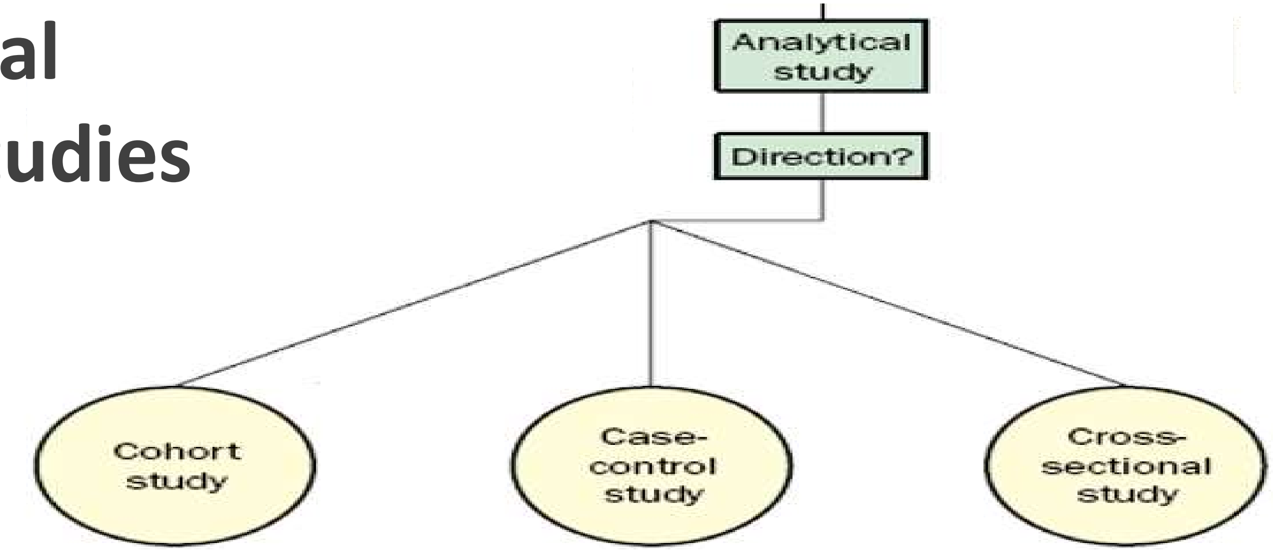
- Use **blocks:** repeat the randomization of treatments within each block (groups of units with same experimental conditions, ex. center)
  - ➢ This removes (or reduces) variability due to the experimental conditions and not to the treatment

Summary

Three very important considerations in the design of experiments are the following:

1. Use *randomization* to assign subjects to different groups
2. Use replication by repeating the experiment on enough subjects so that effects of treatment or other factors can be clearly seen.
3. *Control the effects of variables* by using such techniques as blinding and a completely randomized experimental design

Observational Analytical Studies



Assessment of Exposure	present future	past present	present
Assessment of Outcome	present future	past present	present
Selection of the sample	Random From Exposure	From the Outcome	Random

Classification of epidemiological studies

Pearce* (2012):

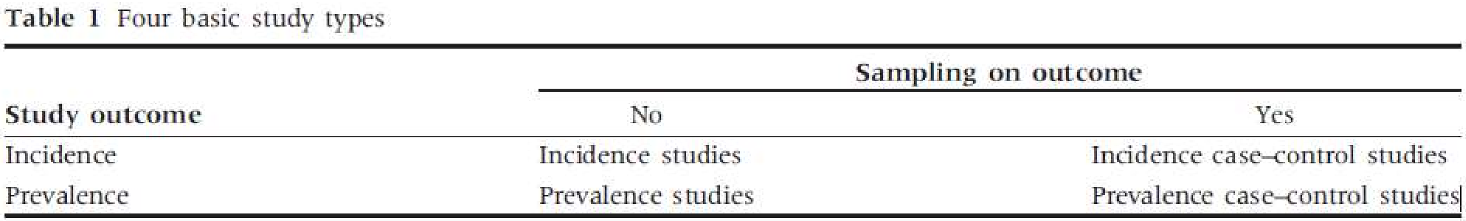


Table 1 Four basic study types

	Sampling on outcome	
Study outcome	No	Yes
Incidence	Incidence studies	Incidence case–control studies
Prevalence	Prevalence studies	Prevalence case–control studies

* *Pearce N. Classification of epidemiological study designs. Int. Jour Epi. 2012*

Study outcome & Measures of disease occurrence

Prevalence

The number of cases of the disease in a population at a specific time divided by the number of members in the population

Measures:

- Prevalence: The proportion of people having the disease at a specified time, π (t)
- Prevalence Odds: $\frac{\pi(t)}{1-\pi(t)} = \frac{No.of\ cases}{No.of\ non\text{-}cases}$ **at time t.**

Incidence

The number of new cases of the disease in a population in a specified time.

Measures:

- Rate number of events occurring *per unit of time* **r(t)**
- Cumulative incidence: The proportion of people who get the disease during the follow up period Π.

Another way of presenting proportions (Odds)*

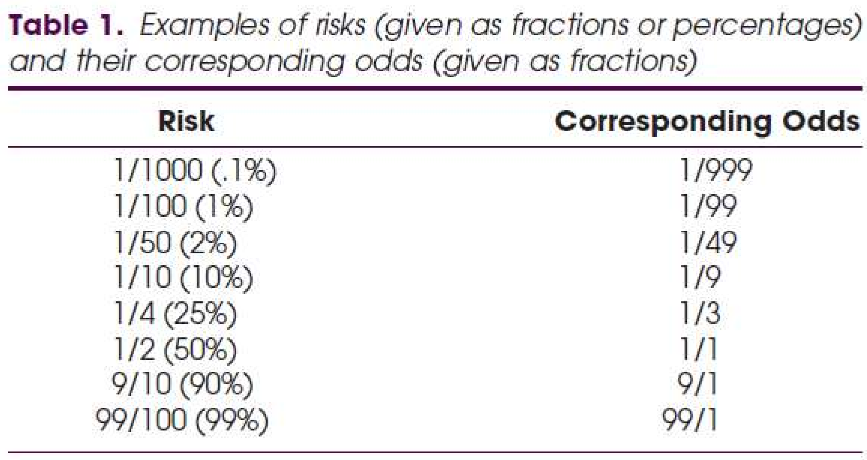


Table 1. *Examples of risks (given as fractions or percentages) and their corresponding odds (given as fractions)*

Risk	Corresponding Odds
1/1000 (.1%)	1/999
1/100 (1%)	1/99
1/50 (2%)	1/49
1/10 (10%)	1/9
1/4 (25%)	1/3
1/2 (50%)	1/1
9/10 (90%)	9/1
99/100 (99%)	99/1

* From Sainani, Physical Med and Rehab 2011, *Understanding odds ratios.*

Quiz: If you describe your ability at some game as “4 wins in every 5 attempts”, is this an odds of 4/5, ¼, 1/5 or 4?

Cross-sectional studies (Surveys)

- Like taking a "snapshot" of the population
- ascertain outcome (Y) and exposure (X)

Cohort studies

- Enrol a well-defined group of individuals at a given time ("time" can be age/date/other start)
- "follow" the experience of those individuals over time
- Like taking a "video"

Case-control studies

- Start by identifying cases (Y=1) and reference/controls (Y=0).

Terminology: Random Variable & data

Random Variable: Characteristic observable in a target population. It may vary from subject to subject.
Data: Numbers or modalities expressed by a random variable (r.v.)

Example:
RANDOM VARIABLE (denoted with UPPER CASE):

X : SEX

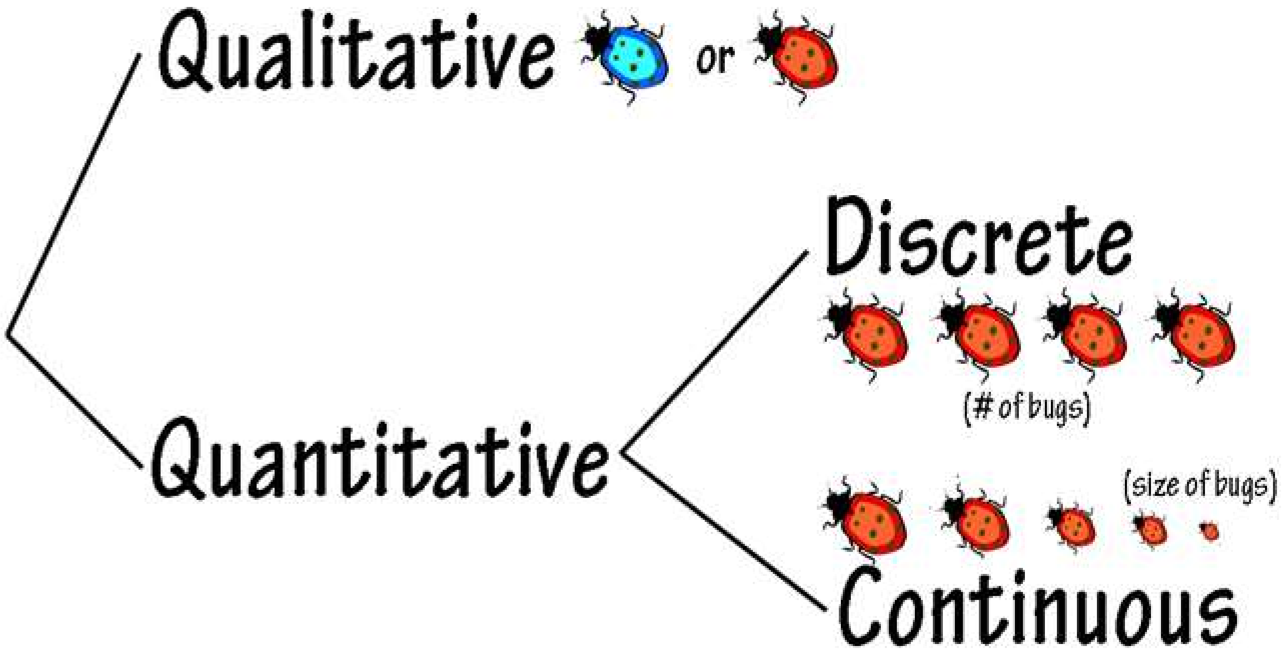
Data (denoted with lower case):

x_1=Female x_2=Male x_3=Female

RANDOM VARIABLE: **Y : AGE**

Data: y_1=1 y_2=4 y_3=12

Terminology: Random Variables classification



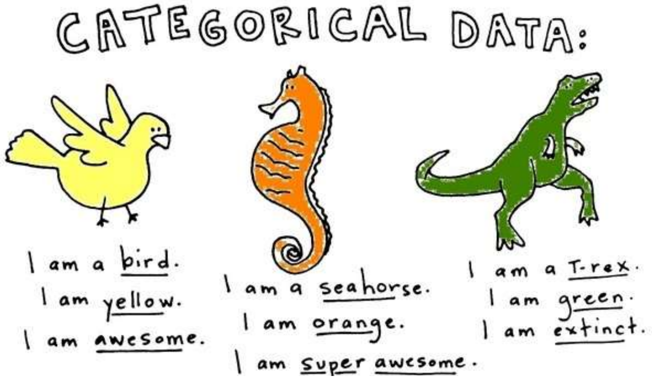
<u>Terminology:</u> Random Variables classification

Qualitative Variable – Non numerical characteristic, attribute

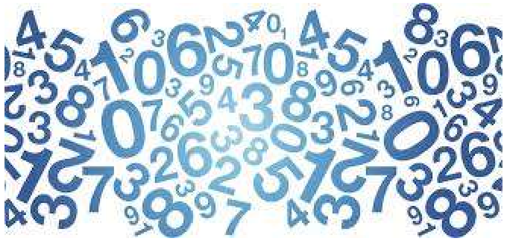
Nominal – Words without a logical natural ordering (e.g. marital status, ethnicity)
Ordinal – Words with logical natural ordering (e.g. perceived pain, education)

Some nominal variables could become ordinal (e.g. patient diseases can be ranked according to a severity of the disease), thus the boundary between the two definitions becomes flexible.

Qualitative variables are also called categorical variables.



Terminology: Random Variables classification



Quantitative Variable – numerical characteristic

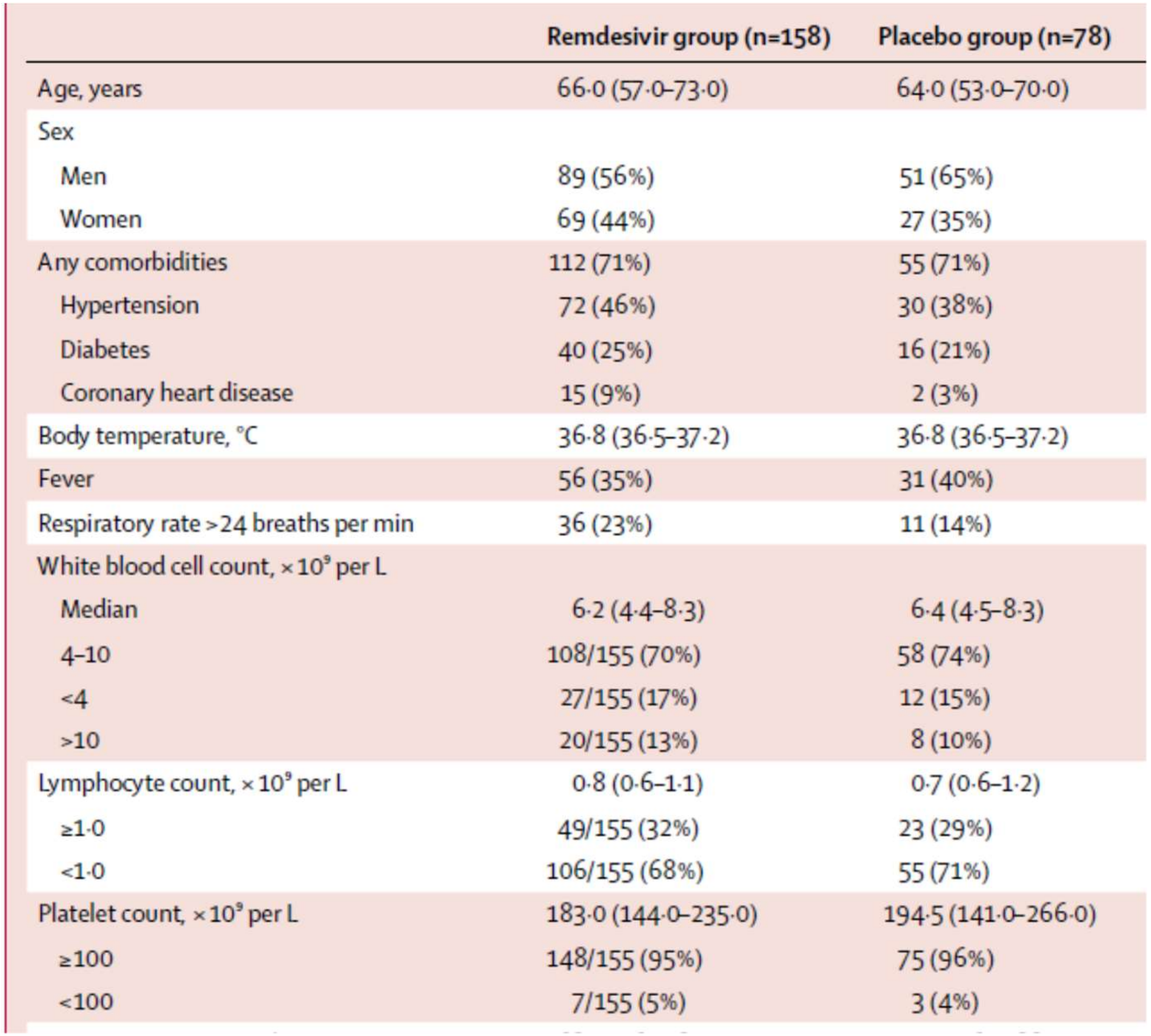
Discrete – natural numbers obtained by counts (e.g. number of sons or daughters)

Continuous – continuous numbers obtained by measurements (e.g. weight, income, mechanical strength)

Any continuous number is rounded to a unit measurement (e.g. weight in kilograms), thus the boundary between the two definitions becomes flexible.

Discrete variables usually assumes a limited number of values, whereas continuous variables assumes a relatively big number of values even under this rounding operation.

Terminology: Variables classification



	Remdesivir group (n=158)	Placebo group (n=78)
Age, years	66·0 (57·0–73·0)	64·0 (53·0–70·0)
Sex		
Men	89 (56%)	51 (65%)
Women	69 (44%)	27 (35%)
Any comorbidities	112 (71%)	55 (71%)
Hypertension	72 (46%)	30 (38%)
Diabetes	40 (25%)	16 (21%)
Coronary heart disease	15 (9%)	2 (3%)
Body temperature, °C	36·8 (36·5–37·2)	36·8 (36·5–37·2)
Fever	56 (35%)	31 (40%)
Respiratory rate >24 breaths per min	36 (23%)	11 (14%)
White blood cell count, × 10^9 per L		
Median	6·2 (4·4–8·3)	6·4 (4·5–8·3)
4–10	108/155 (70%)	58 (74%)
<4	27/155 (17%)	12 (15%)
>10	20/155 (13%)	8 (10%)
Lymphocyte count, × 10^9 per L	0·8 (0·6–1·1)	0·7 (0·6–1·2)
≥1·0	49/155 (32%)	23 (29%)
<1·0	106/155 (68%)	55 (71%)
Platelet count, × 10^9 per L	183·0 (144·0–235·0)	194·5 (141·0–266·0)
≥100	148/155 (95%)	75 (96%)
<100	7/155 (5%)	3 (4%)

Which variable in table 1 is Quantitative-continuous?

1. **Sex (Males or Female)**
2. **Hypertension (yes or no)**
3. **Blood pressure (mmHg)**
4. **COPD (yes or no)**
5. **Respiratory support (Oxygen Mask, Noninvasive ventilation, Invasive mechanical ventialtion)**
6. **PEEP**
7. **No. of patients with PEEP measurement**

TEAM-BASED LEARNING 06/10 on descriptive statistics :

1. Read chapter “4. Summarising data” from the book “An Introduction to Medical Statistics, Martin Bland 2015.” (pdf in the course webpage)

Contents of the chapter:

4.1 Types of data
4.2 Frequency distributions
4.3 Histograms and other frequency graphs
4.4 Shapes of frequency distribution
4.5 Medians and quantiles
4.6 The mean
4.7 Variance, range, and interquartile range
4.8 Standard deviation