Collecting Sample Data / 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*.

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

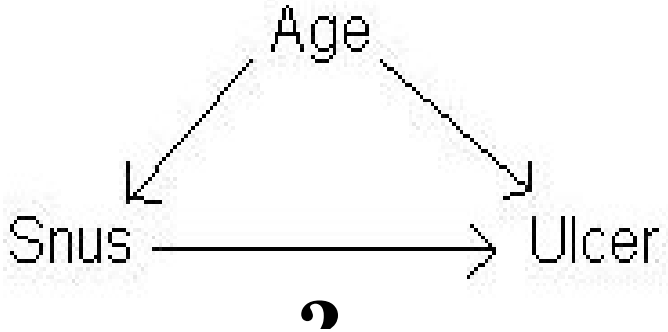
Can we safely conclude that snus prevents ulcer?

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**.

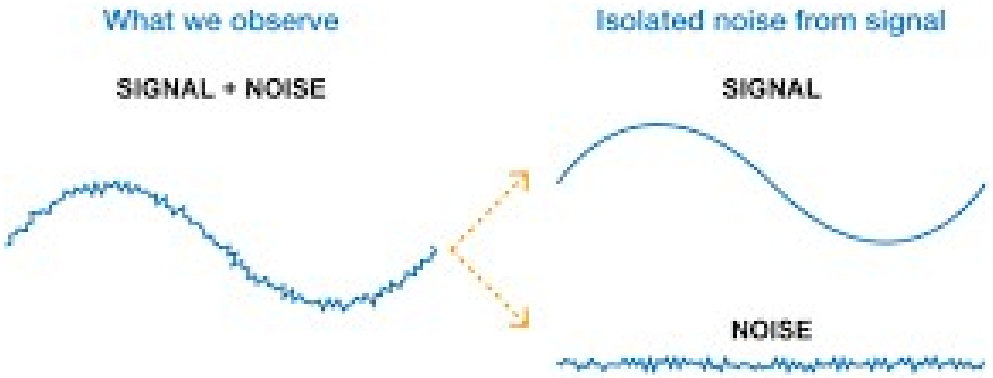
Biomedical research

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.

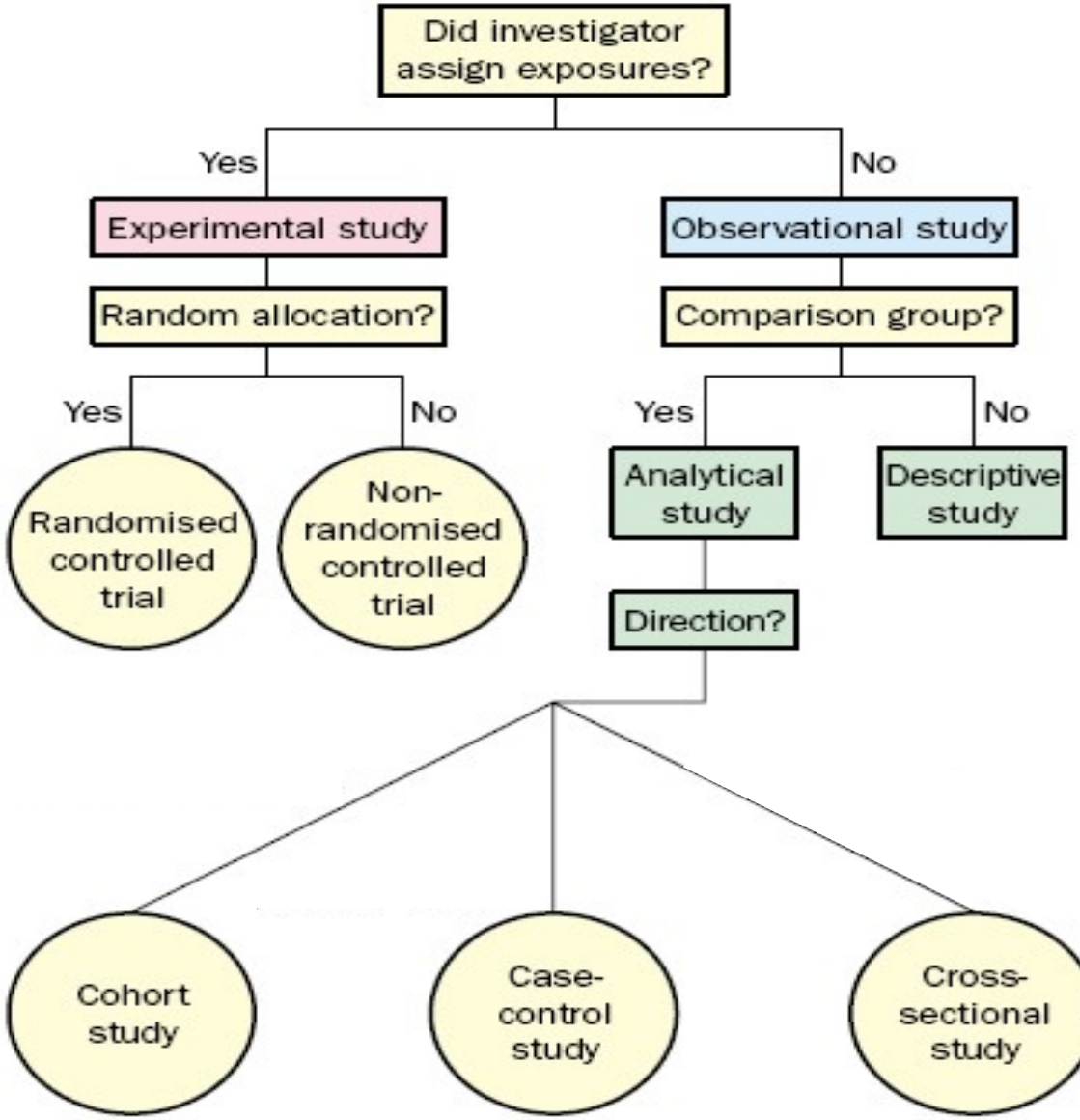
We have to distinguishing between signal and background noise



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



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

Observational Study

Observational study

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

Experiment

Experiment

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

Beyond the Basics of Collecting Data

Different types of observational studies and experiment design

Types of Studies - observational

- **Cross sectional study**

  data are observed, measured, and collected at one point in time

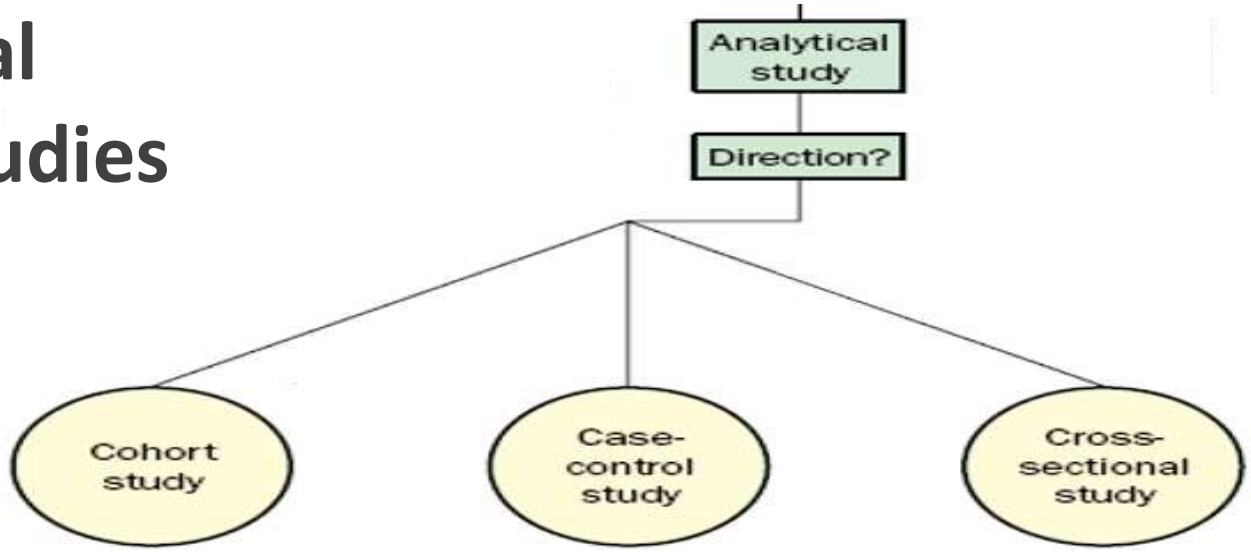
- **Retrospective study (ex. case-control)**

  data are collected from the past by going back in time (examine records, interviews, …)

- **Prospective (ex. longitudinal or cohort) study**

  data are collected in the future from groups sharing common factors (called cohorts)

Observational Analytical Studies



	Cohort study	Case-control study	Cross-sectional study
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
Note	Prospective	Retrospective	Contemporary

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

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**

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

    $\Rightarrow$ 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

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

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 $n_1=n_2$**

$$(p_E\text{-}p_{NE}) \sim N\left(0; \sqrt{\pi(1-\pi)\left(\frac{1}{n_E}+\frac{1}{n_{NE}}\right)}\right)$$

$$\bar{Y}_E - \bar{Y}_{NE} \sim N\left(0; \sigma\sqrt{\frac{1}{n_E}+\frac{1}{n_{NE}}}\right)$$

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**