Biophysics I.
for dentistry students

Lecture 2nd:

Biostatistics II.

2019. September 16.

Dániel Veres

Repetition:Population and Sample

(Problematics of inferential statistics)

Population

RANDOMNESS!

Sample

The size of the **population** usually does not allow the examination of all of its elements.

Therefore, only a subset of the population is examined. That is what we call a **sample**.

UNCERTAINTY!
(P and E)

Characteristics of the sample can be used to draw conclusions on the population.

We carry out measurements on the sample elements, then this data set (which is also called **sample**) will be characterized by graphs and numbers

Estimation

Population		Sample
Real value	Estimation	Estimate
Probability		Relative frequency
Expected value (mean of population)		Mean of the sample
Theoretical variance		(Empirical) sample variance
Difference of 2 expected value		Difference between 2 sample mean

Estimation

Population		Sample
Real value	Estimation	Estimate
Probability		Relative frequency
Expected value (mean of population)		Mean of the sample Median of the sample?
Theoretical variance		(Empirical) sample variance – which one?

$$v1 = \frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{n-1} \qquad v1 = \frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{n}$$

...

Error

Population

- Real value (estimand)

Estimation

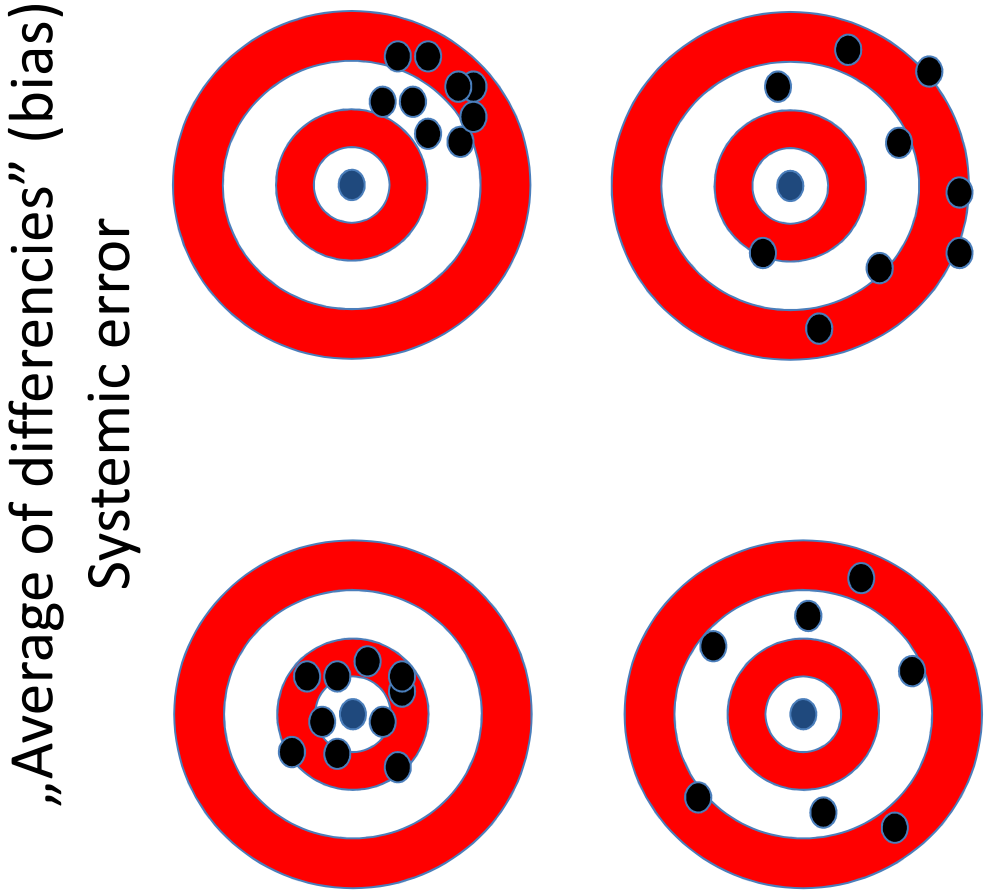
Sample

- Estimates

Error

Imagine more random samples – more estimates

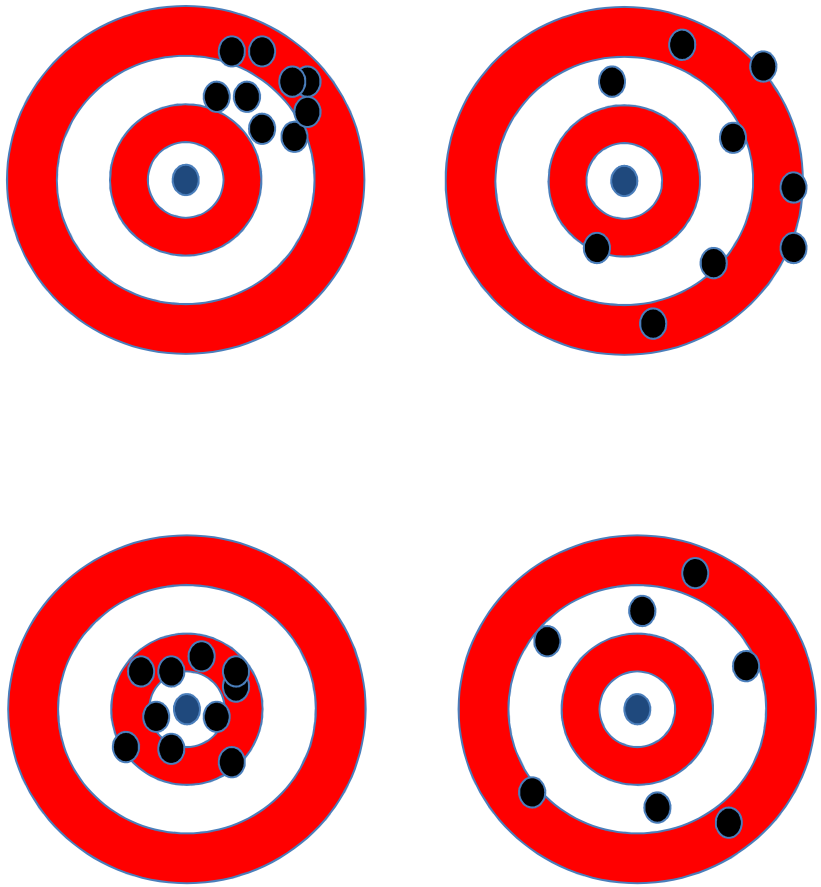
Error – 2 dimension



1. Variability of estimates
2. Difference between the real value and the „center” of the estimates

Error – 2 dimension

„Average of differencies” (bias)
Systemic error



„Variability” (standard deviation);
Random error

Good estimation, if:

Unbiased:
the expected value of the estimates is equal with the estimand

Effective:
the variability is small

(Konsistent:
the variability of estimates decreases with increased sample size)

Notes

- The mean is an unbiased, consistent and most effective estimate for estimating the expected value
- In the estimation of variance, the division by n-1 is unbiased, while division by n will be biased

Confidenc intervals

But we have only 1 estimate and we don’t know the real value!

– Could we measure the error?

YES, the random error (sampling error) could be estimated based on the sample! It is called the standard error of the estimate.

HOW?

Confidence intervals

But we have only 1 estimate and we don't know the real value!

– Could we measure the error?

The variability (random error, sampling error) could be estimated based on the sample!

Example:

The standard error of the mean (shortly standard error, SEM)

Reminder: CLT: Central limit theorem (on sampling): for given conditions, sampling with large sample size (n) the distribution of the sample means is normal with:

$$\mathrm{Var}_{\mathrm{normal}} = \frac{\mathrm{Var}_{(\mathrm{original})}}{n}$$

Confidence intervals

1. Therefore the standard error of the mean (the variability of estimating the mean): is the square root of the nth part of the sample variance.
2. We can construct a range, that contains the mean with a given probability: called confidence interval of the mean.
   the given probability: called confidence level

In the case of the mean (because of normal estimation – see CLT) eg.:
The limits of the 95% confidence interval of the mean are:

$$\sim \bar{x} \pm 2 * SEM$$

We can construct confidence intervals for other estimats too!
It shows the value of the estimate, its error and its confidence interval

Hipothesis tests

Sampling („random”) error with an other method

Aim of hypothesis tests: **Statistical answer on YES/NO question**

HOW to prove a statement?

Direct proof: prove for all cases that the statement is true.

eg: sum of 1,2…n : (n+1)*(n/2) – proving with induction

Indirect proof: assume the opposite statement and I prove that is false.

Indirect proof

We have a box containing 100 marbles. Each of them are either red or white.

Case #1: hypothesis (H): all of them are white.
Experiment: We randomly take a marble out of the box.
Our observation: It is red.
Conclusion: The probability of our observation *given* our hypothesis is 0: Our hypothesis is for 100% sure wrong.

Case #2: hypothesis (*H*): 99 are white and one is red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: Our hypothesis is for *almost* 100% sure wrong: The probability of our observation *given* our hypothesis is $0.01^5 = 10^{-10}$: practically impossible.

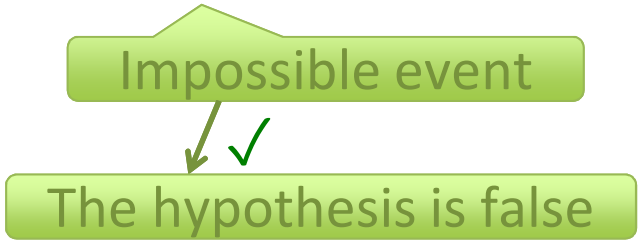
Case #3: hypothesis (H): 50 are white and 50 are red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: Now we are not sure what to do: The probability of our observation given our hypothesis is $0.5^5 = 0.03125$: low but not that unlikely...

Case #4: hypothesis (*H*): all of them are red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: The probability of our observation *given* our hypothesis is $1^5 = 1$.
Are we sure what to do now?

Indirect proof

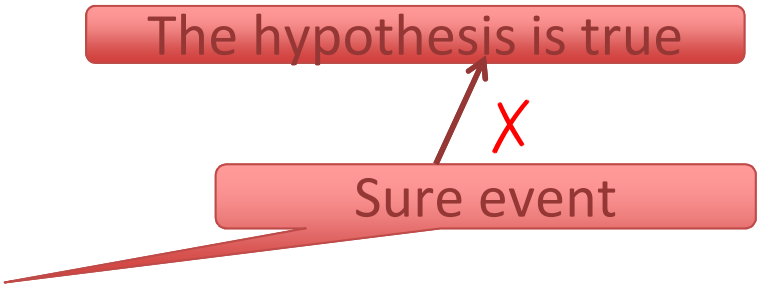
We have a box containing 100 marbles. Each of them are either red or white.

Case #1: hypothesis (H): all of them are white.
Experiment: We randomly take a marble out of the box.
Our observation: It is red.
Conclusion: The probability of our observation *given* our hypothesis is 0: Our hypothesis is for 100% sure wrong.



Falsification works

Verification does not work



Case #4: hypothesis (*H*): all of them are red.
Experiment: We randomly take a marble out of the box and put it back; we do this 5 times.
Our observation: All of them are red.
Conclusion: The probability of our observation *given* our hypothesis is $1^5 = 1$. Are we sure what to do now?

Indirect proof

What about #2 and #3?

Mathematical Logic:

We have a hypothesis (*H*).

If *H* is true, *E* event cannot occur.

E occurs.

So *H* is not true.

As we saw it before, **a hypothesis can only be rejected**.

Statistical Logic:

We have a hypothesis (*H*).

If *H* is true, *F* event is very unlikely to occur.

F occurs.

So we reject *H*. But we are not 100% sure if *H* is not true.

In this case a hypothesis cannot even be rejected with 100% certainty.

WE HAVE TO ESTIMATE THIS PROBABILITY

What kind of questions can we test?

Y/N ...must be a yes/no (a.k.a. dichotomous or polar) question. **5W1H**

- Is the 5-years survival rate (i.e. probability) for myeloma 50%?
- Does the total blood cholesterol level of Cushing's syndrome patients differ from the general 200 mg/dL population mean? ✓

- What is the 5-years survival rate for myeloma?
- What is the expected value of total cholesterol level in Cushing's syndrome patients? ✗

...must refer to a set of observations, not to individual cases.
(And the question is aimed at a population, not a sample.)

- Is the 5-years survival rate for myeloma 50%? ✓

- Will this myeloma patient survive for 5 years? ✗

...must have at least one unambiguous answer.

- Is the 5-years survival rate for myeloma 50%? ✓

- Is the 5-years survival rate for myeloma less than 50%? ✗

What kind of answers can we test?

We have two answers for our question:

The null hypothesis (H_0)

- **Unambiguous**: can be realized in only one way. It contains some form of =.
  The 5-years survival rate
  for myeloma is 50%.

- Represents the current well-established, generally **accepted scientific knowledge**,
The total blood cholesterol level of Cushing's syndrome patients is same as the population mean
or something that is the **most trivial** with the least assumptions (Occam's razor).
The probability of landing on heads in a coin tossing experiment is 50%.
- It is **not** necessarily the negative answer to the question.

The alternative hypothesis (H_1)

- Typically can be realized in more than one way.
  The 5-years survival rate
  for myeloma is not 50%.
  (can be a little more, a lot less etc.)

- Represent a **new statement** challenging the current scientific consensus,
The total blood cholesterol level of Cushing's syndrome patients differs from the population mean
or a set of all the **not-so-trivial** answers needing more or special assumptions.
The probability of landing on heads in a coin tossing experiment is other than 50%.
- It is typically **complementary to H_0** (i.e., its negation).
  H_1 = not H_0

Sampling („random”) error

Aim of hypothesis tests: **Statistical answer on YES/NO question**

Starting point: create a specific statistical question and answers:

H_0: null hypothesis – „random” error only

H_a (or H_1): alternative hypothesis – not H_0

Decision is based on: role of „randomness” if H_0 true (sampling error)

A sample is that could contradict H_0

Mintavételből származó („véletlen”) hiba

Hipotézisvizsgálat **célja**: **eldöntendő kérdésre statisztikai választ** adjon

Kiindulópont: az eldöntendő kérdés statisztikai átfogalmazása, majd :

H_0: nullhipotézis – „véletlen” hiba

H_a (vagy H_1): alternatív hipotézis (ellenhipotézis) – nem H_0

Döntés alapja: a „véletlen” szerepe a H_0 esetén (mintavételi hiba)

A minta az, amely esetleg megcáfolja H_0-t

		A populációban (a valóságban) a null hipotézis:	
		Igaz	Hamis
A döntés: a null hipotézist:	Megtartom (Nem vetem el)	Helyes döntés	Hiba (**másod fajú**) (β) (álnegatív eredmény)
	Elvetem	Hiba (**első fajú**) (α) (álpozitív eredmény)	Helyes döntés (**erő**) (1-β)

An example – first steps

Situation: We play a board game with dice – we do not win...
This is a „wrong" dice?

What is the question??!! (and what is it about):

All of the sides has the same probability? (for this dice)
The probability of six-throw is different from 1/6, even bigger?
– let's use this question

Null hipothesis ??!!:

The probability of rolling 6 is 1/6 or less – hmm (*multiple hypothesis*)
Let's use the worst according to alternative hypothesis
(since if only one element of the multiple hypothesis can "conform" to the null hypothesis, there is no evidence of rejection in the full range of hypothesis):
H_0: The probability of rolling 6 is 1/6.
H_a: greater than 1/6

An example – next steps

What is the question : The probability of six-throw is different from 1/6, even bigger?

Null hypothesis: H_0: The probability of rolling 6 is 1/6.

How much evidence do we need for saying it is differ?

– Significance level??!!:

Given by authority… Used in the literature--- **STARTING VALUE**, BUT smaller/larger, more/less frequent side effects?, cheaper/ more expensive?...

here: we play this game every weekend, it's not expensive to change the die

– let's use 10% instead of 5% (smaller evidence enough to reject H_0)

Collecting evidence – the sample ??!! (how, how much… ask your statistician):

results: 6 times 6 out of 24 rollings

I really answer the question I have created?

Do I need to modify the question? (eg. the population of interest)

An example – additional steps

What is the question : The probability of six-throw is different from 1/6, even bigger?

Null hypothesis: H_0: The probability of rolling 6 is 1/6..

Significance level:10%

Sample: 6 times 6 out of 24 rollings.

Is the difference important at all?– relevant ??!!:

6 times 6 out of 24 rollings, it is 6/24 = 1/4 probability that **1,5** times higher than 1/6 – YES it is

How much evidence – p-value ??!!:

How to calculate it?, (unbiased, effective, consistent…)

Important aspects: the question, type of variables (measuring scale), measurement conditions

We need for calculation:

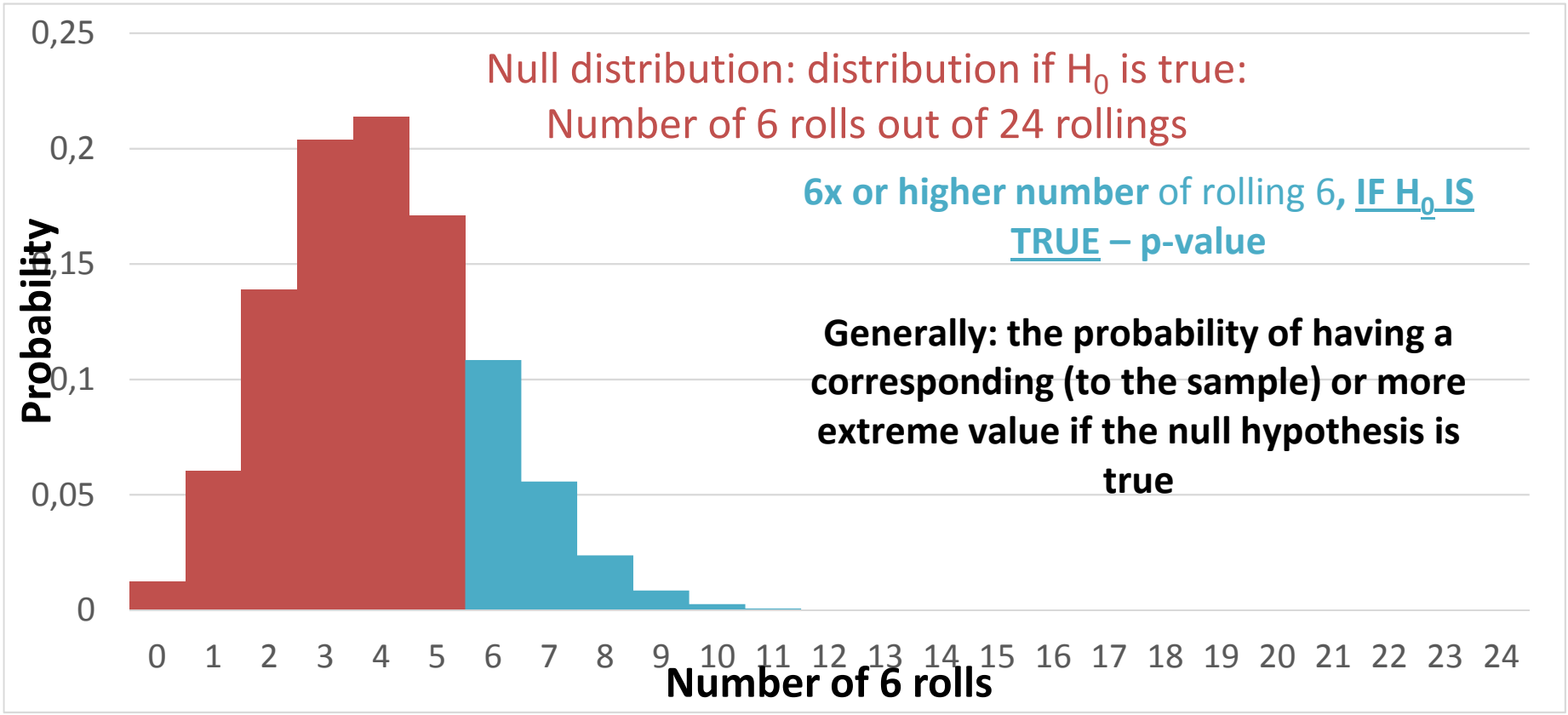
Null distribution – if null hypothesis is true, what is the probability of all theoretical samples we could have?

The „position" of our sample in this distribution:

test statistics (or statistics, eg: frequency, probability , t-value , p-value…)

Calculation in the „background”

Using binomial test!



An example - last steps

What is the question : The probability of six-throw is different from 1/6, even bigger?

Null hypothesis: H_0: The probability of rolling 6 is 1/6..

Significance level:10%

Sample: 6 times 6 out of 24 rollings.

Is the difference important at all?– relevant : 1/4 probability, that **1,5** times higher than 1/6 – YES it is

How much evidence? – p-value: 0,1995

Decision: there is not enough evidence for reject H_o– accept H_0

		In population (in reality) the null hypothesis is:	
		True	False
Decision on null hypothesis:	Accepting (Not rejecting)	Good decision	Error (**type II**) (β) (false negative result)
	Rejecting	Error (**type I**) (α) (false positive result)	Good decision (power) (1-β)

One sample Student t-test

What I'm curious about

Expected value of the sample is equal with a known population mean

Type of variable

1 numerical and continuous

Assumption

Independent observations

distribution of means is normal:

normally distributed sample or large sample size (CLT)

Notes: Calculation:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Do NOT test normality with other hypothesis test (it increases Type I. error, „multiplicity")! Use previous knowledge on the variable and make graphs.

Paired Student t-test

What I'm curious about

Two expected values in two groups are equals – in paired groups

Type of variable

1 numerical and continuous, 1 binary („groups")

Assumptions

Independent observations in the groups, paired groups

the distribution of the difference of means is normal

normally distributed differences or large sample size

Notes:

paired test usually has higher power

suitable to compare other location parameters (quantiles)

difference of the means = mean of the differences

2 sample Student t-test

What I'm curious about

Two expected values in two groups are equals

Type of variable

1 numerical and continuous, 1 binary („groups")

Assumptions

Independent observations between and within groups

distribution of means is normal in each group:

distribution is normal in each group or large sample size

distribution of standard deviations are the same

Notes:

suitable to compare other location parameters (quantiles)

if we don't know the variances do not test (multiplicity) – use Welch test instead!

Welch test

What I'm curious about

Two expected values in two groups are equals

Type of variable

1 numerical and continuous, 1 binary („groups")

Assumptions

Independent observations between and within groups

distribution of means is normal in each group:

distribution is normal in each group or large sample size

Notes:

suitable to compare other location parameters (quantiles)

not sensitive for different variances (robust for variance differences)

Chi-square test for independence

What I'm curious about

Two variables are depends on each other

Type of variable

2 categorical variable

Assumptions

Independent observations

None of the „expected" frequencies smaller than 1 and maximum 20% smaller than 5.

„Correlation” t-test (Pearson linear regression)

What I’m curious about

Two variables are (linearly) depends on each other

Type of variable

2 numerical variable (X and Y)

Assumptions

Independent observations for pairs

linear relation assumed

x values measured with no error

y-s have a normal distribution at each x

y-s have same variance at each x

Notes

2 estimates: slope and intercept, slope is important and tested

Relevant, but not significant…

Reasons:

small power:

small sample size (limitation: money, ethical issues)*

large variability

less powerfull statistical test

we could not measure it accurately

violated assumptions for the test

we were unlucky (sampling error)

other errors

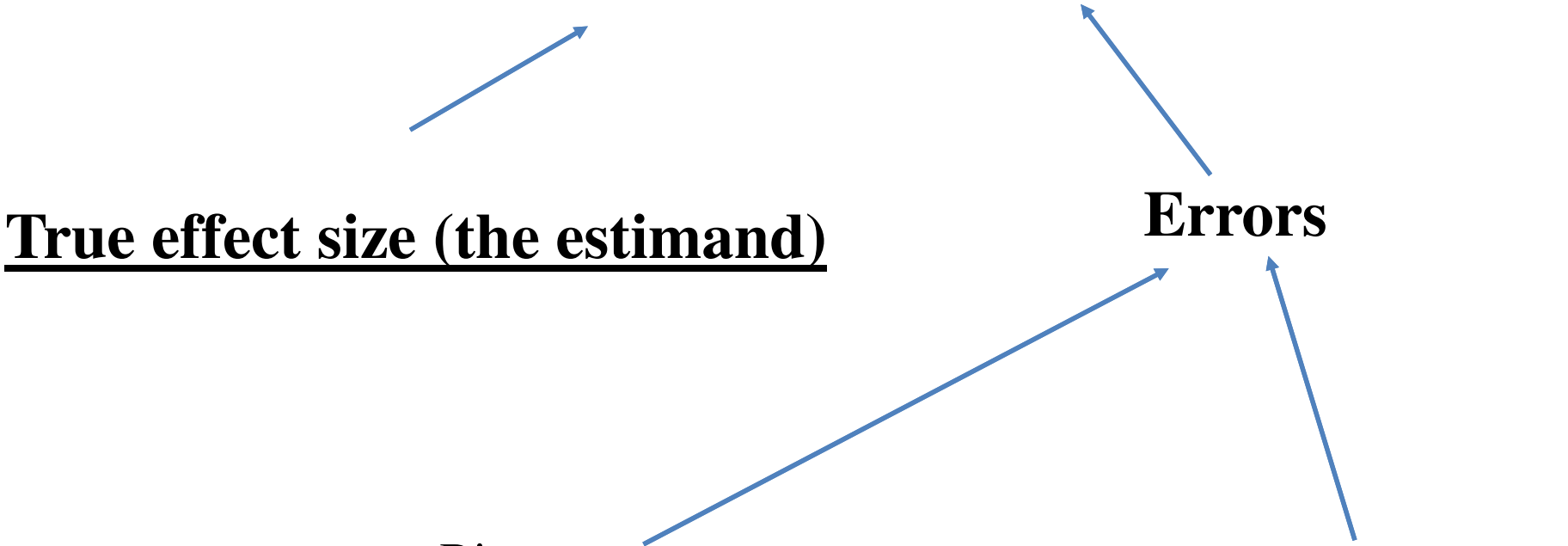
Plan ahead!!

-*Ask yor statisticians...

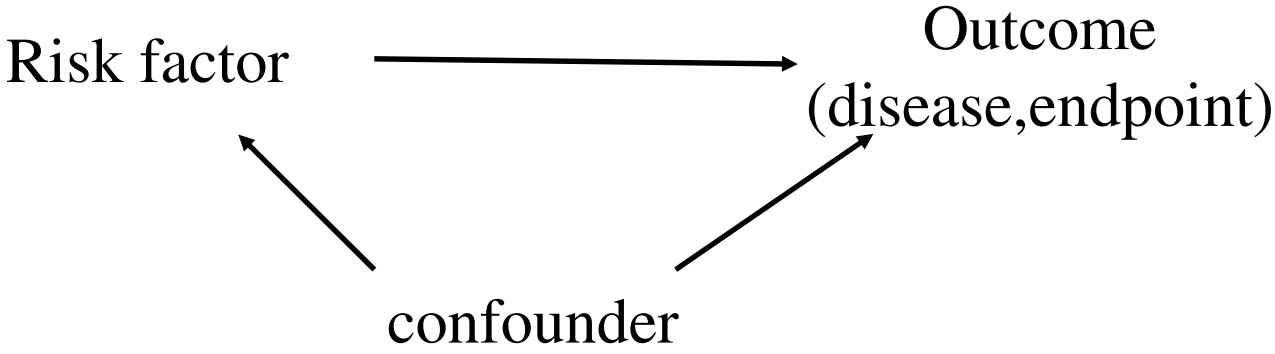
(☺ eg: https://www.youtube.com/watch?v=PbODigCZqL8)

Other errors

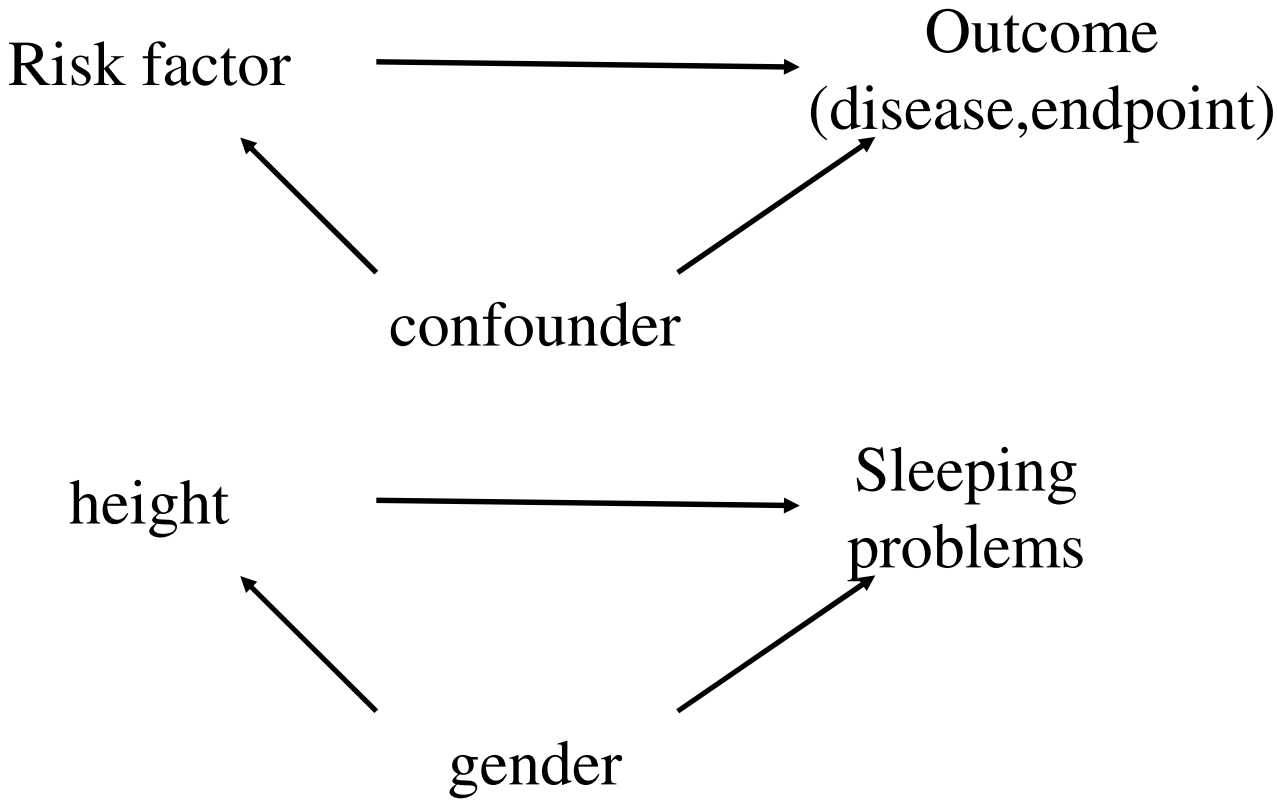
Effect size based on the sample (the estimate)



Confounding bias



Confounding bias



Most common confounders: gender, age – always think about them!

Selection bias, Information bias

Selection bias:
There is a difference between the selected and not selected individuals, or difference between assignment to groups (erroneous selection with respect to an outcome influencing parameter)

tipical: age, gender different in the groups
different population
different follow-up time

Information bias:
erroneous data collection about or from subjects (which affects the outcome)

tipical: recall bias
more carefull monitoring for diseased, young

Test Questions

- What does systestemic and random error means?
- What does unbiased estimation mean?
- What does effective estimation mean?
- What does consistent estimation mean?
- What does confidence interval mean?
- What is the aim of a hypotheisis test?
- Give the criteria of a good question in a hypotheisis test.
- Give the criteria of a good null hypothesis in a hypotheisis test.
- What does significance level means?
- What does relevant mean?
- Define first type error.
- Define second type error.
- Define the p-value in a hypothesis tests.
- What does confounding bias mean?
- What does selection and information bias mean?