Sample Size Estimation for Exploratory Animal Research: A critical assessment

Servan Grüninger
MSc Biostatistics (UZH), MSc Computational Science & Engineering (EPFL)

PhD candidate in Epidemiology and Biostatistics
@SGruninger, www.servangrueninger.ch

Outline

1. Overview of design and analysis of experiments

2. A critical review of approaches to sample size calculations

3. Take home messages

Expectation management

My goal for today is to give you an insight into the rationale behind frequentist hypothesis testing and sample size calculations.

I will (hopefully) be able to

- explain the logic behind different approaches to sample size estimation;
- critically assess these approaches.

I will not be able to:

- help you with an immediate problem regarding sample size calculation (requires more than 45 minutes)

Warning: There will be theory!

1. Overview of design and analysis of experiments

1.A. What does it mean to design an experiment?

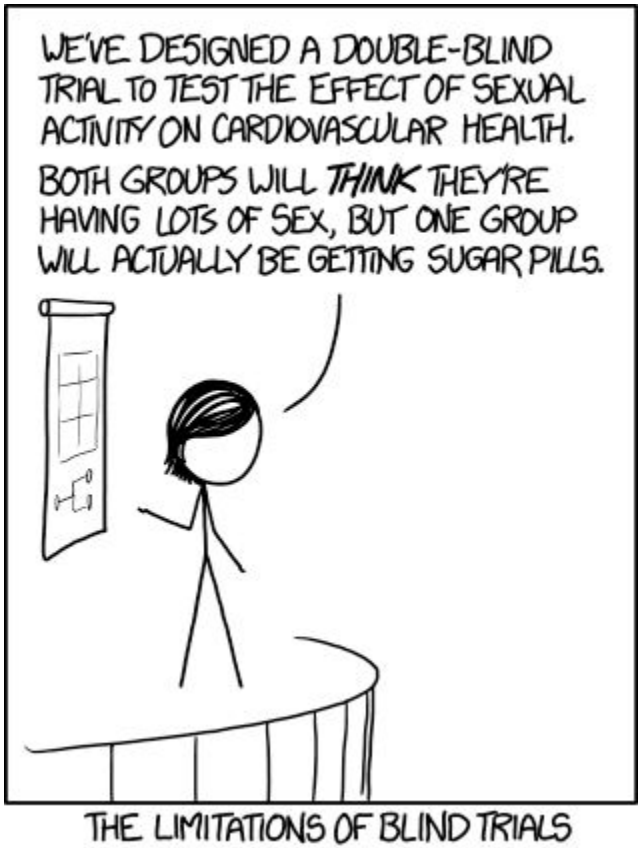
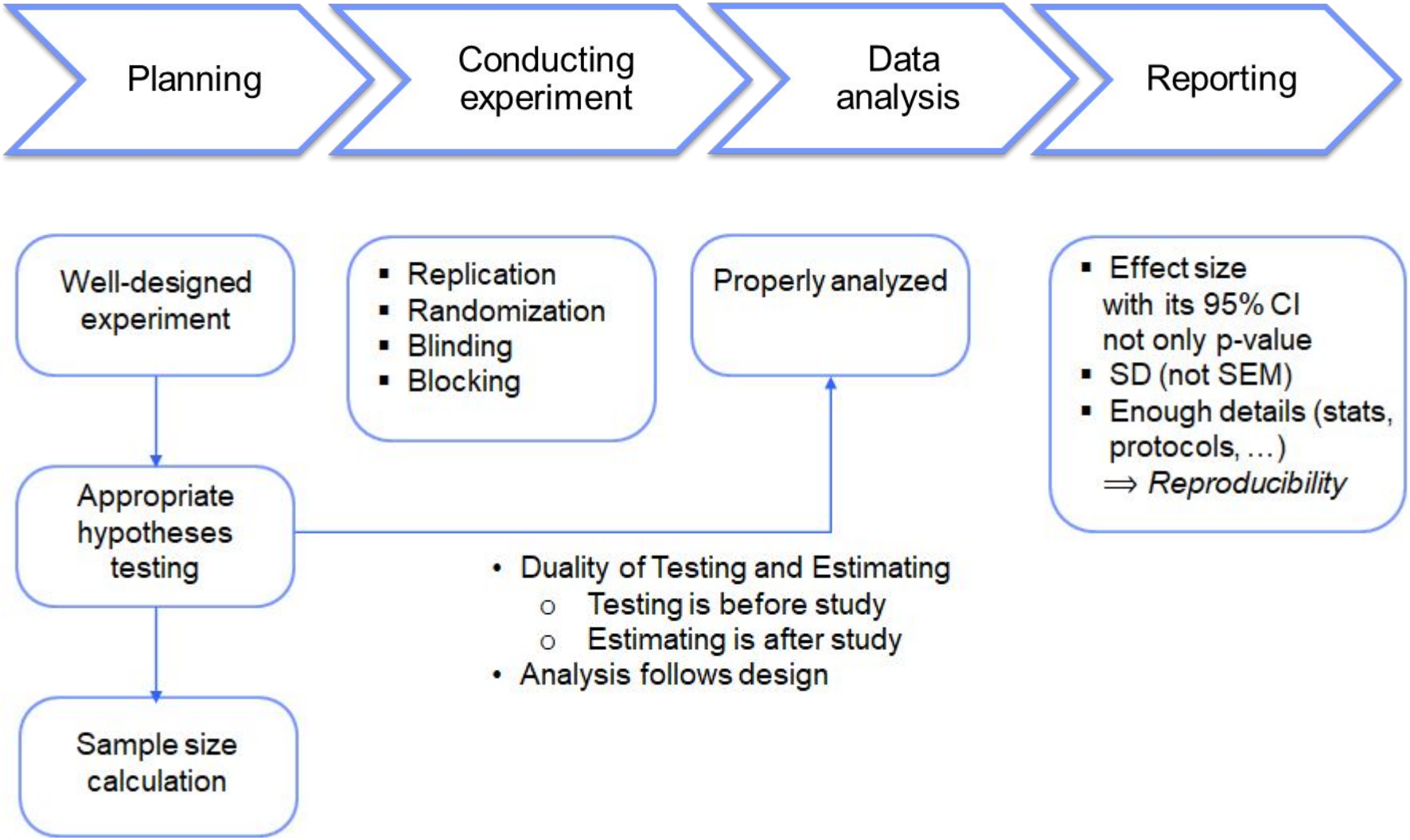


Image: xkcd comics (https://xkcd.com/1462/)

Details of the four stages of an experiment



Characteristics of well-defined experiment

Characteristics	How to do it
Clear objective	PICO-B method (Population, Intervention, Control, Outcome, Blocking)
Clear definition of experimental units	Think about the smallest unit to which you can apply a different treatment
Unbiased	Randomized, Blinding
High precision (low variability)	Replication, Blocking
Able to estimate uncertainty	Replication, Randomized
Wide range of applicability	Blocking (deliberate variation)
Simple	Protect against mistakes

Characteristics of well-defined experiment

Characteristics	How to do it
Clear objective	PICO-B method (Population, Intervention, Control, Outcome, Blocking)
Clear definition of experimental units	Think about the smallest unit to which you can apply a different treatment
Unbiased	Randomized, Blinding
High precision (low variability)	**Replication**, Blocking
Able to estimate uncertainty	**Replication**, Randomized
Wide range of applicability	Blocking (deliberate variation)
Simple	Protect against mistakes

The Design Informs the Sample Size!

If two researchers use

- the **same** species,
- the **same** tools,
- the **same** interventions,
- the **same** statistical test,

to answer the **same** scientific questions, they **might still need different sample sizes depending on the statistical design they chose**!

The fundamental experimental design equation

Fundamental experimental design equation:

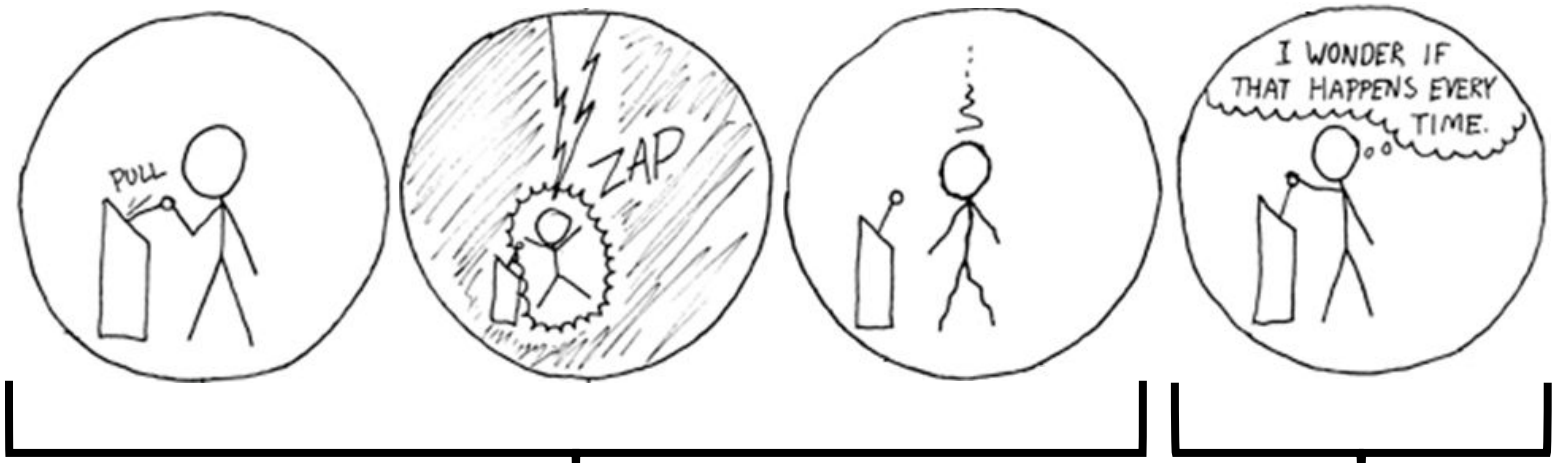
Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”)

Conceptually easy, but details depend on many factors:

- Exploratory or confirmatory experiment
- Knowledge about factors and effect sizes
- Available resources

1.B. Exploratory vs. confirmatory research

Why the difference matters



Explore

Confirm

Image: adapted from xkcd comics (https://xkcd.com/242/)

Exploratory vs. confirmatory research

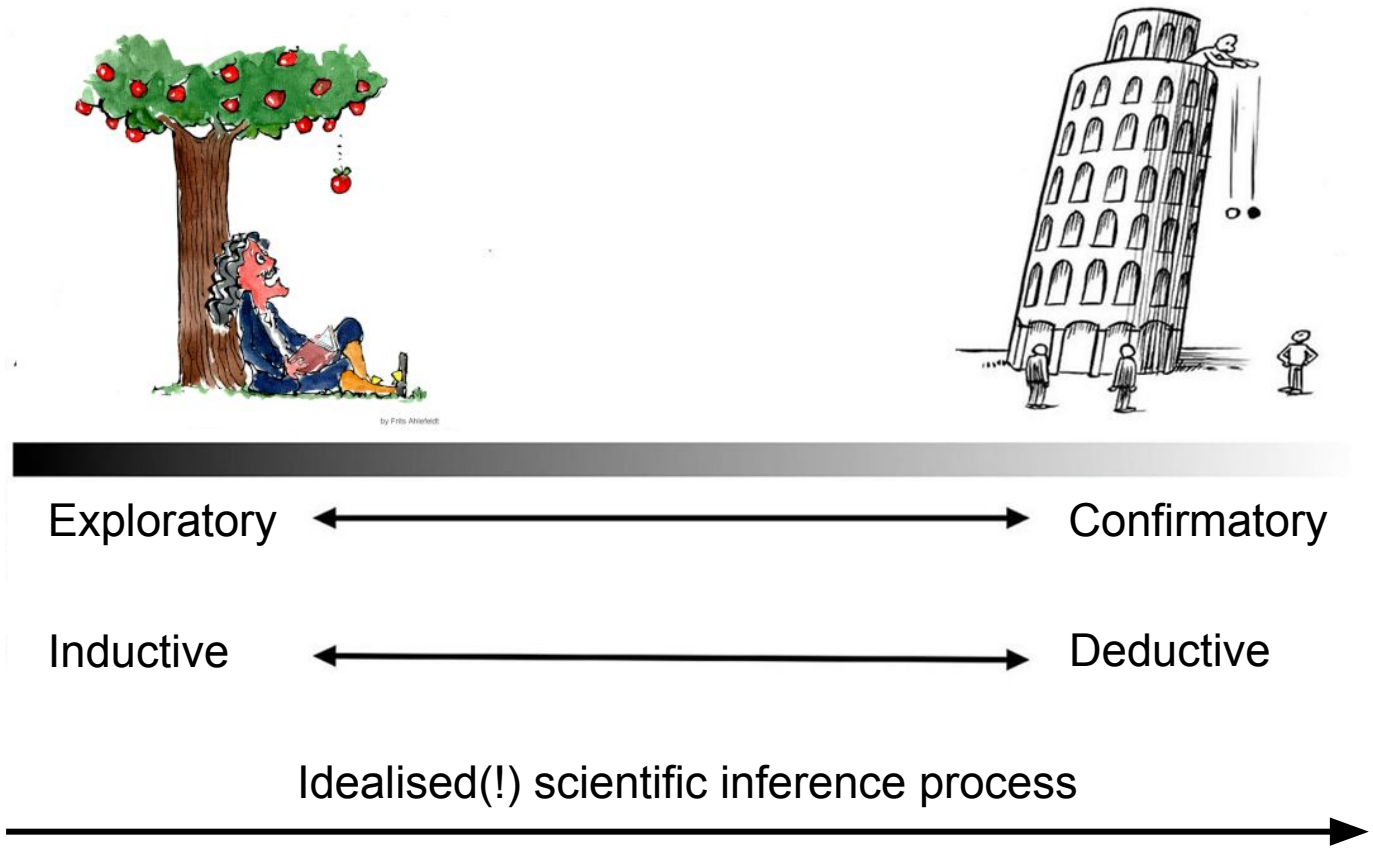


Image: adapted from Dirk-Jan Hoek and Frits Ahlefeldt
see also: Wagenmakers et al. 2012, An Agenda for Purely Confirmatory Research

2. A critical review of different approaches to sample size calculations

2. A critical review of approaches to sample size calculations

Many approaches to sample size estimation

- Power calculations
- The Resource Equation Method
- Simulation-based methods
- Previous experience
- KISS: Keep it simple, stupid
- The Fermi approximation

Not all of them are recommendable, not all of them are recommendable in the exploratory setting!

Power Calculations: Cookbook recipes for sample size estimations

Power Calculations: Cookbook recipes for sample sizes

In short, power calculation leverages the relationship of the following six elements that are part - in different forms - of every frequentist test:

- n — sample size: number of **true** replications in the experiment
- $z_{1-\alpha}$ — f(alpha level): rate at which one falsely rejects the null hypothesis **if it is true**
- $z_{1-\beta}$ — f(beta level): rate at which one falsely rejects the alternative hypothesis **if it is true**
- μ_0 — effect size under H_0: unstandardized size of the effect on the outcome of interest if the null hypothesis is true
- μ_1 — effect size under H_1: unstandardized size of the effect on the outcome of interest if the alternative hypothesis is true
- σ^2 — variability: variability of the outcome of interest

$$n = \sigma^2 \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\left(\mu_1 - \mu_0\right)^2}$$

An easy example

You want to know whether C57BL/6J mice on a certain diet are on average heavier than 30 grams at the age of 10 weeks.

In other words, you want to test the following set of hypotheses:

$$H_0 : \text{average weight of mice is } \leq 30g$$
$$H_1 : \text{average weight of mice is } > 30g$$

Now you want to know how many mice you need to test this statistically.

For testing, we need a test statistic

Very simple choice: the *z*-statistic (close relative of the famous *t*-statistic)

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

with

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X_i \sim \mathcal{N}(\mu, \sigma^2); \qquad \text{for all } i = 1, \ldots, n$$

$$\mu = \begin{cases} \mu_0 = \text{mean of } X_i \text{ if the null hypothesis is true} \\ \mu_1 = \text{mean of } X_i \text{ if the alternative hypothesis is true} \end{cases}$$

$$\sigma = \text{standard deviation of } X_i$$

$$\sigma/\sqrt{n} = \text{standard deviation of } \bar{X}$$

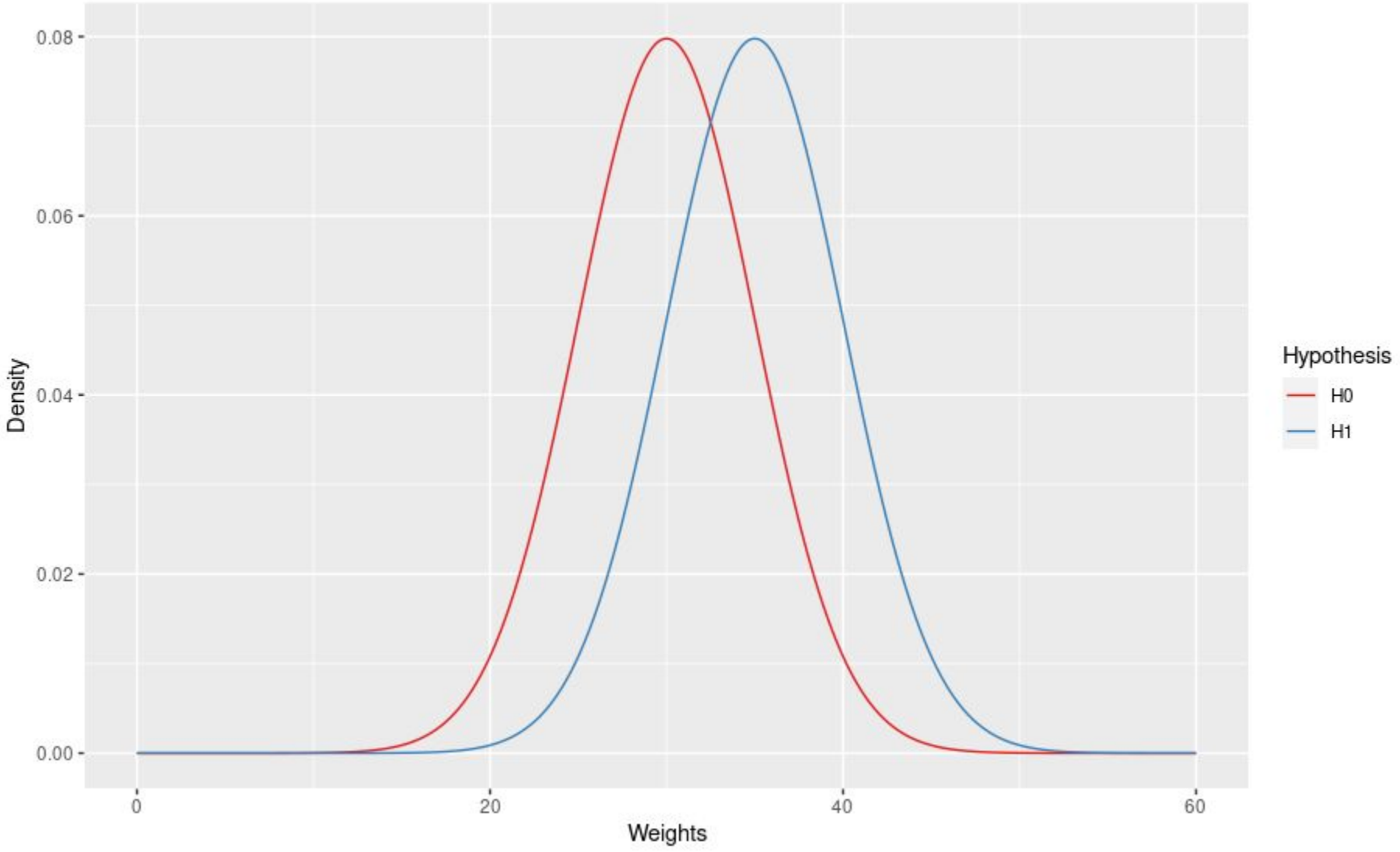
Quick return to the weight example

Let us first pretend that we know more than we actually do

$$\mu = \begin{cases} \mu_0 = 30g = \text{mean of } X_i \text{ if the null hypothesis is true} \\ \mu_1 = 35g = \text{mean of } X_i \text{ if the alternative hypothesis is true} \end{cases}$$

$$\sigma = 5 = \text{standard deviation of } X_i$$

Quick return to the weight example



Quick return to the weight example

Let us also pretend that we have already conducted the experiment (and have actual data which comes from the universe of H_1).

$$x_i = [24.26, 42.29, 39.27, 32.70, 35.34]$$

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{n} x_i = (24.26 + 42.29 + 39.27 + 32.70 + 35.34)/5$$

$$= 34.772$$

$$\sigma/\sqrt{n} = 5/\sqrt{5} = \sqrt{5}$$

Test statistic applied to the weight example

Taken together, we get:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{34.772 - 30}{5/\sqrt{5}} = 2.13$$

But what do we do with this? We need to compare this with something.

This is where the ominous alpha-level aka the Type-I-error rate comes into play.

A test statistic is a random variable!

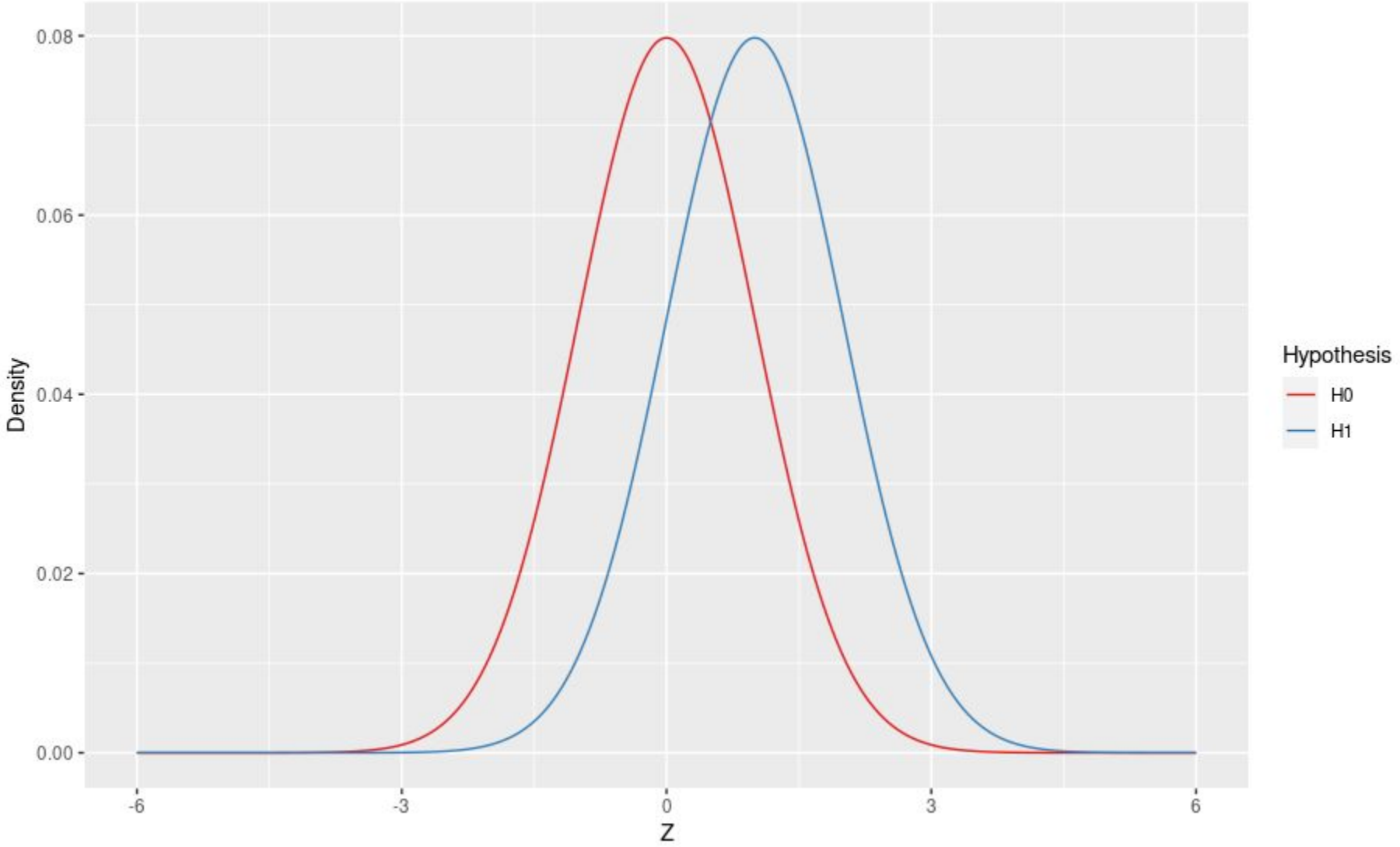
The ***Z***-statistic is a random variable

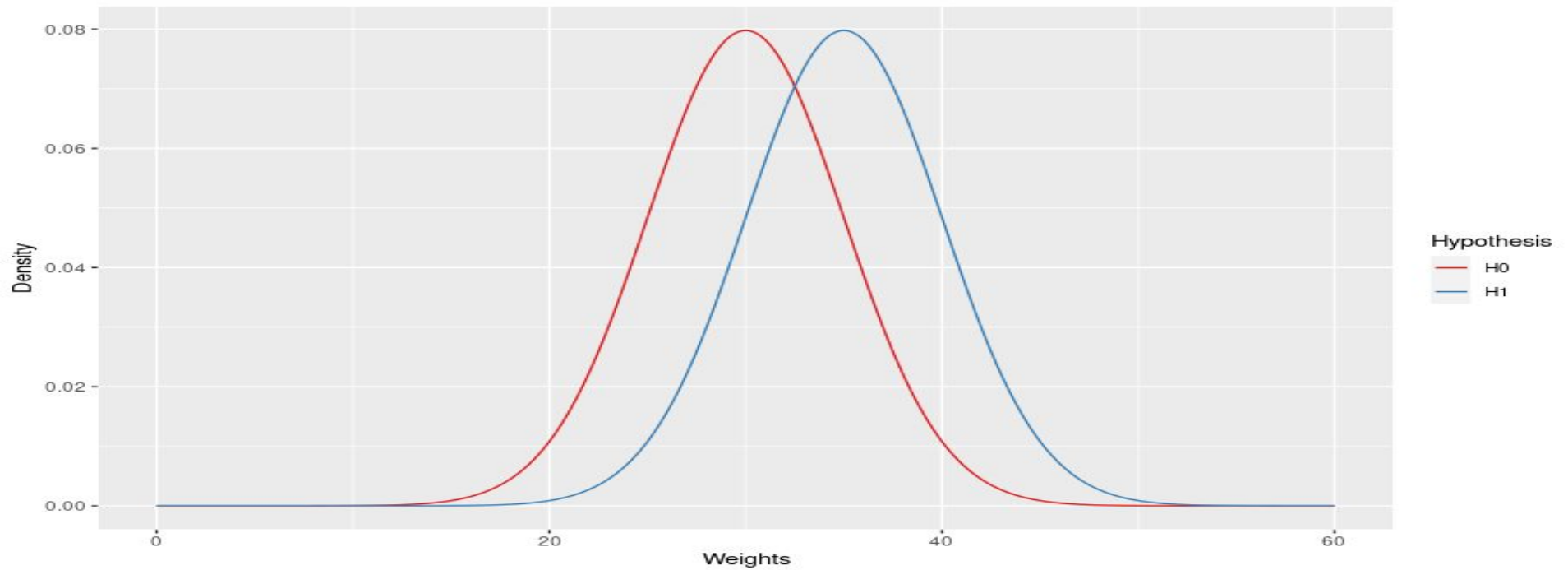
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma}$$

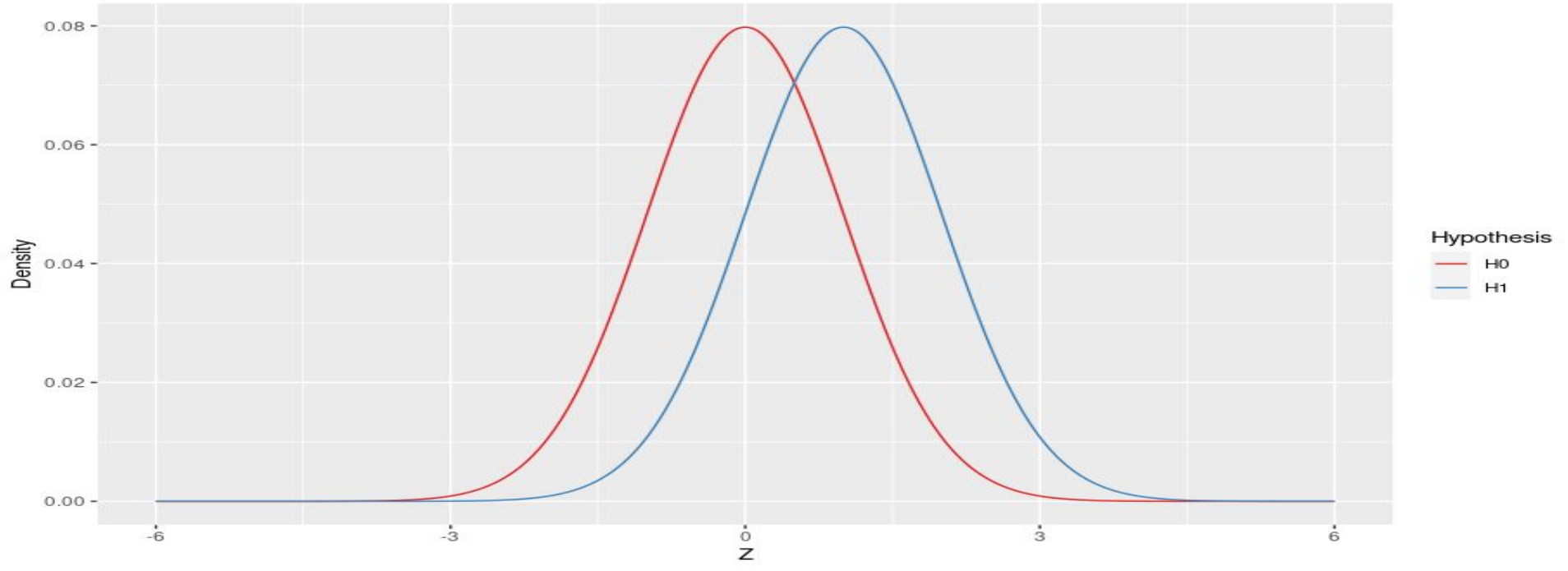
The ***z***-statistic is the **realisation** of that random variable

$$z = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} = \sqrt{5}\frac{34.772 - 30}{5} = 2.13$$

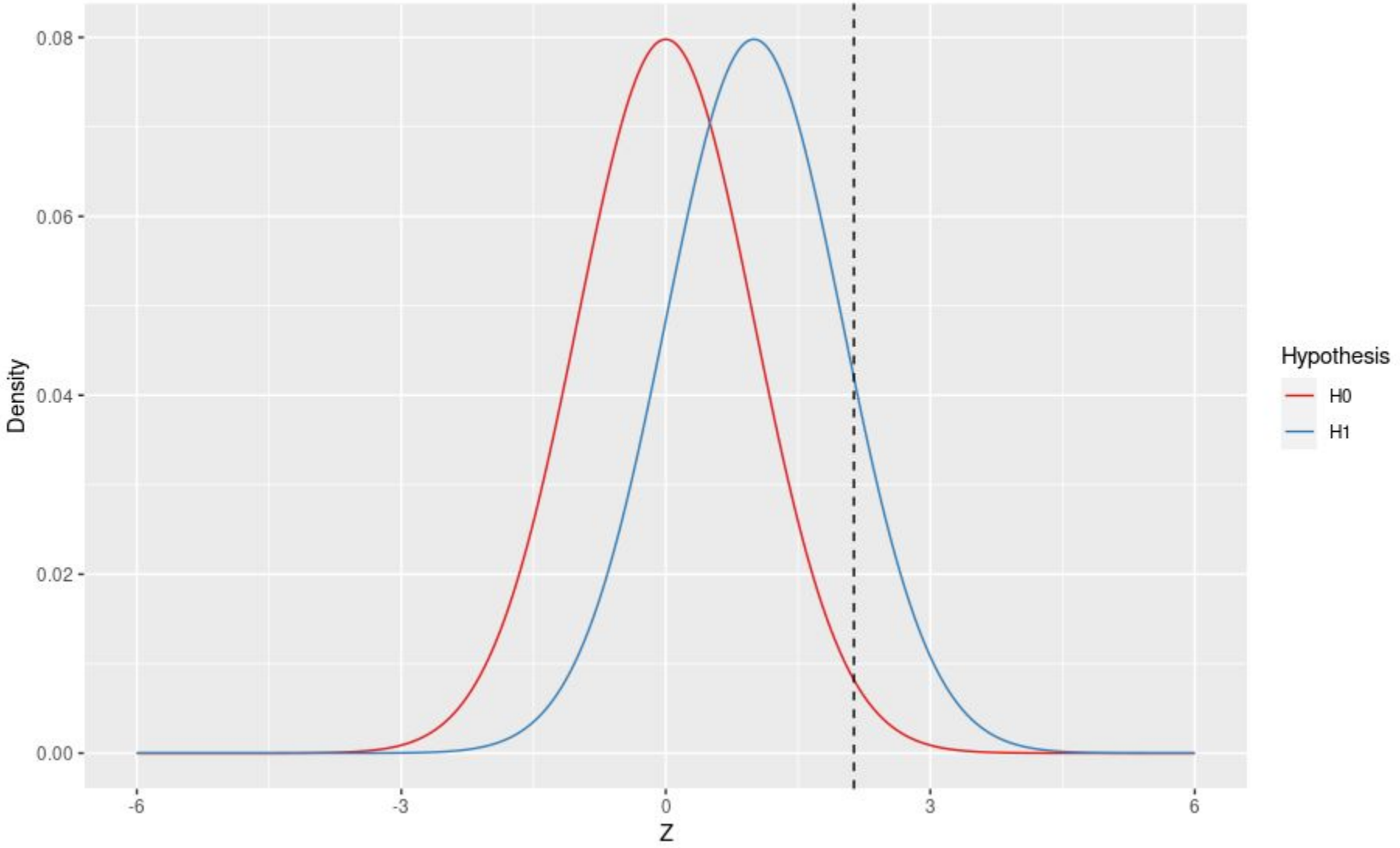
A test statistic is a random variable - with a certain probability density function




0.08
0.06
0.04
0.02
0.00
Density
0
20
40
60
Weights
Hypothesis
H0
H1


0.08
0.06
0.04
0.02
0.00
Density
-6
-3
0
3
6
Z
Hypothesis
H0
H1

In each experiment, we observe one realisation of the test statistic



The alpha level

Imagine you are in the universe in which the null hypothesis is true.

How often - on average - do you want to falsely reject the null hypothesis?

The value you choose is the alpha level

Calculating the critical threshold

The alpha level aka the Type-I-Error is defined as

$$\alpha = \int_{z_{1-\alpha}}^{\infty} f_{H_0}(z)dz = P(Z \geq z_{1-\alpha}) = 1 - F_{H_0}(z_{1-\alpha})$$

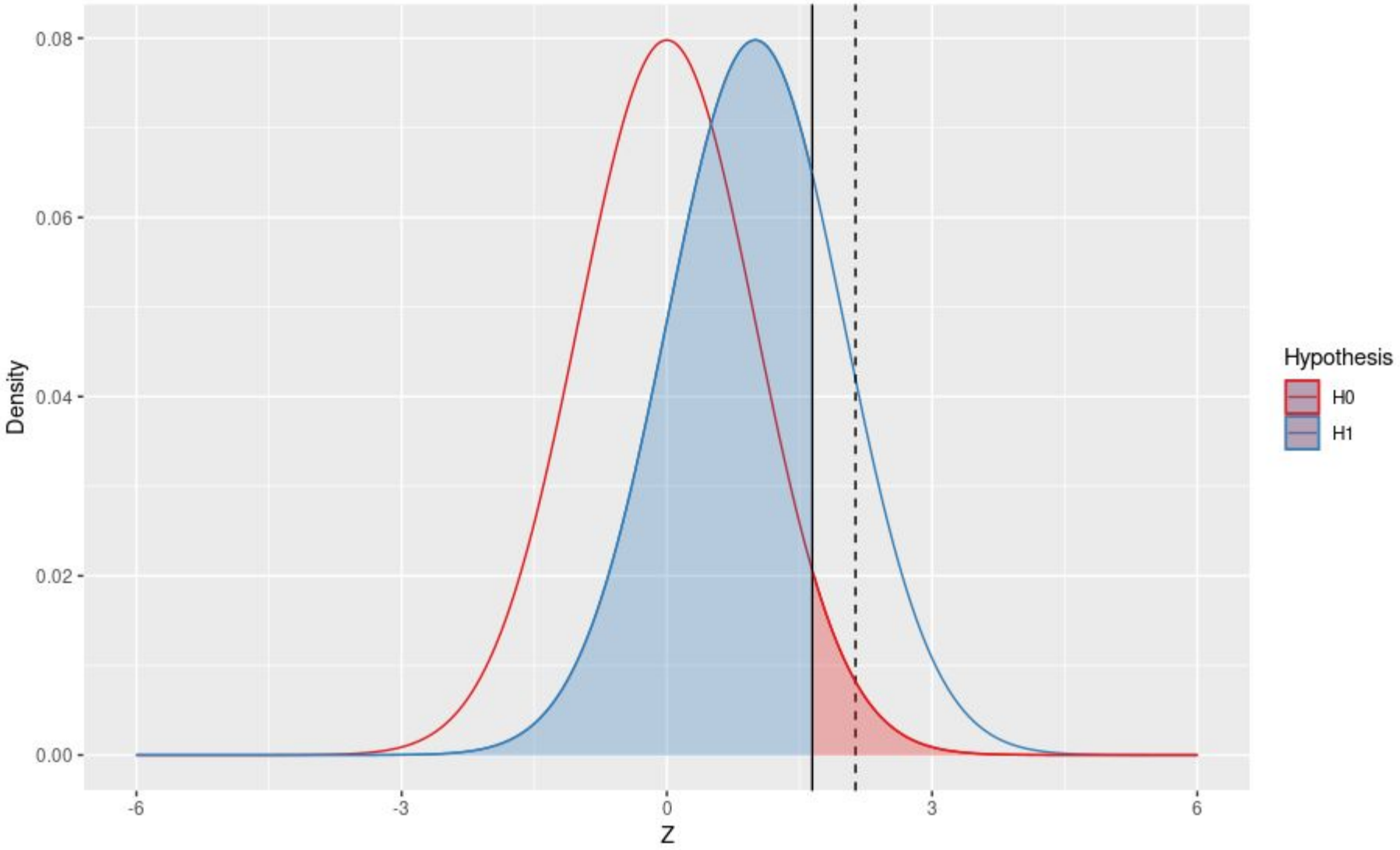
which can be rewritten as

$$F_{H_0}^{-1}(1-\alpha) = z_{1-\alpha}$$

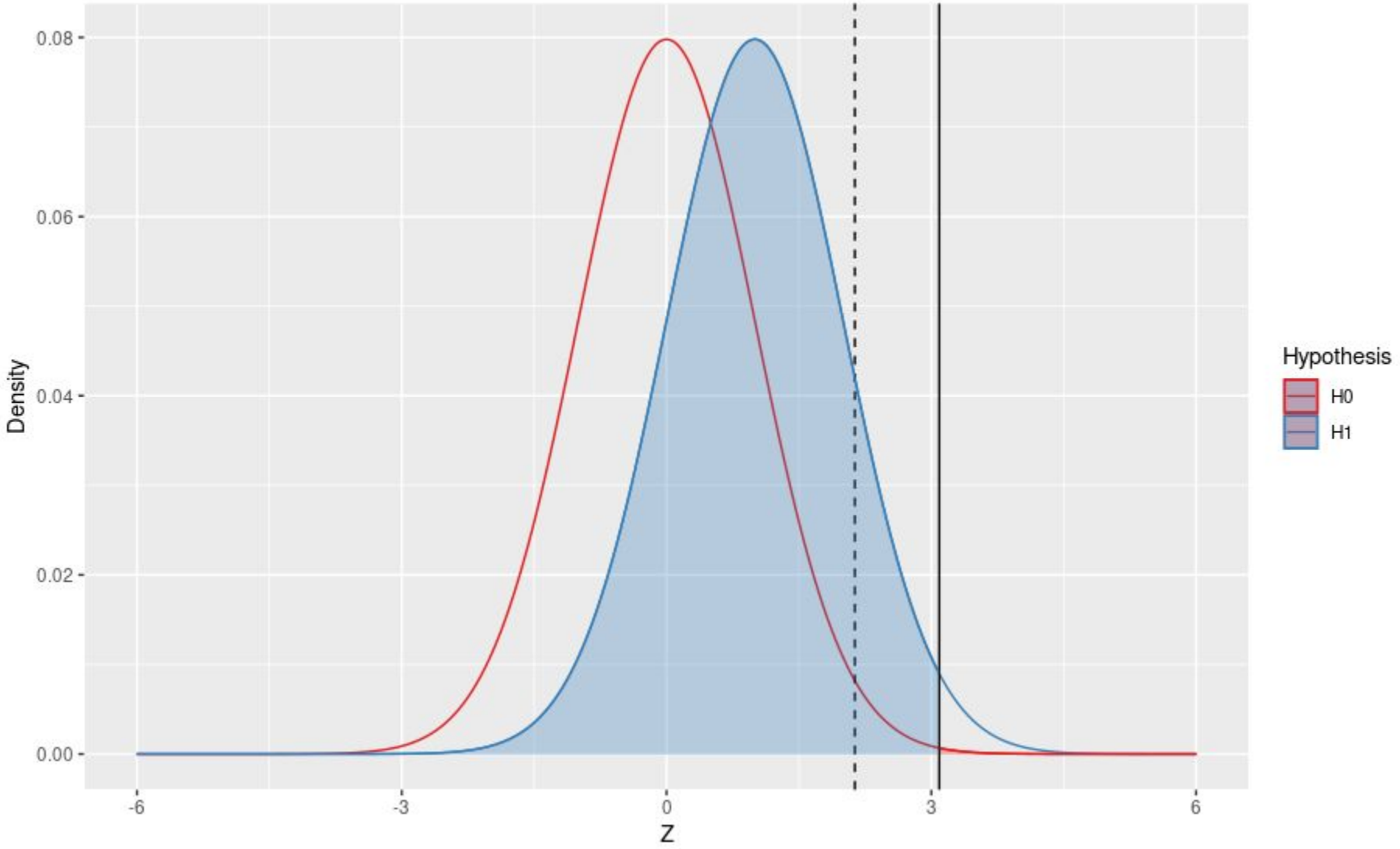
In other words: the critical threshold with which to compare the value of the test statistic directly related to the distribution function of the *Z*-statistics under the null hypothesis.

In yet other words: The critical threshold is a function of the alpha level which is representation of the Type-I-error you are willing to conduct.

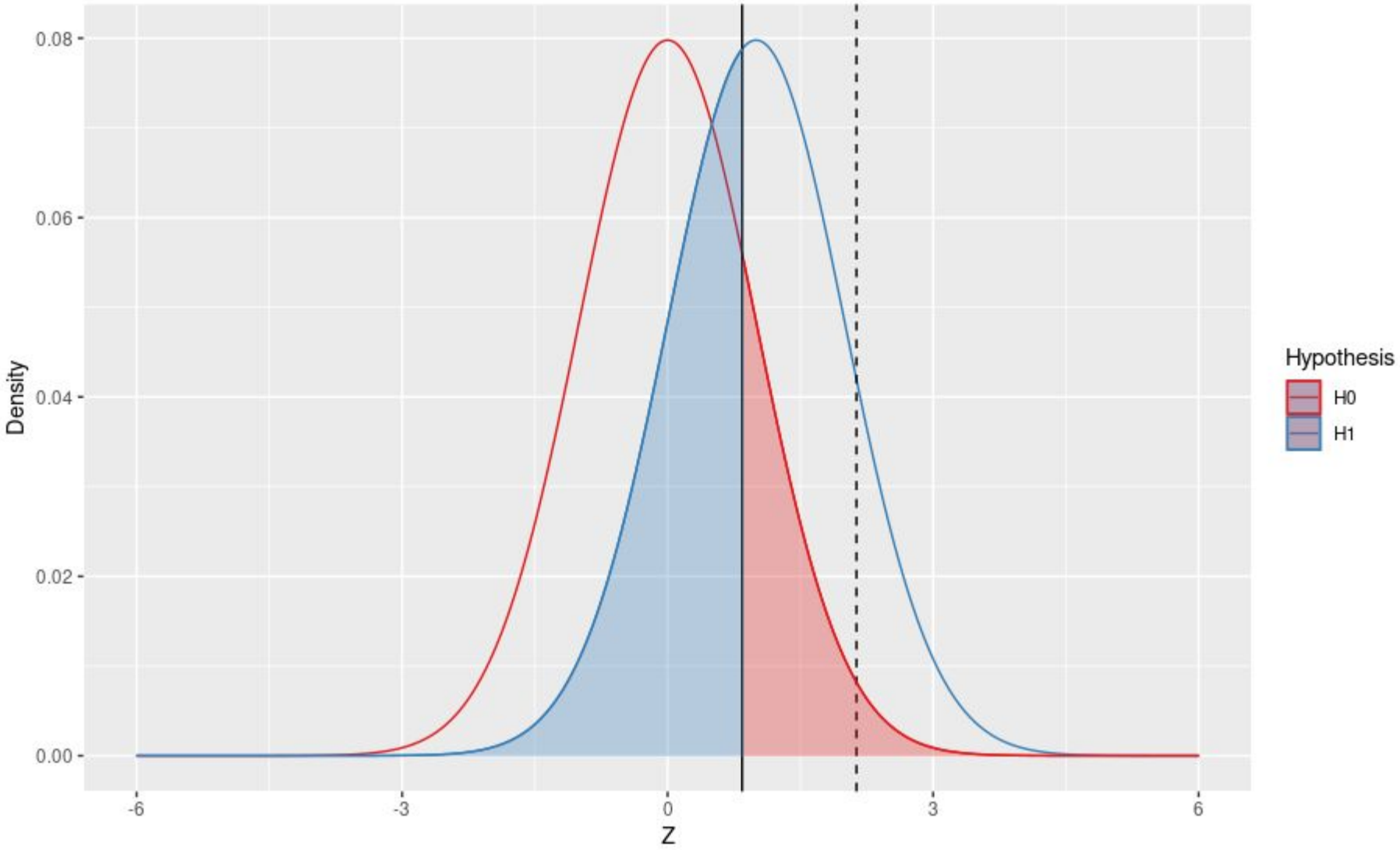
alpha = 0.05



alpha = 0.001



alpha = 0.2



Significance tests are based on the fact that test statistics are random variables

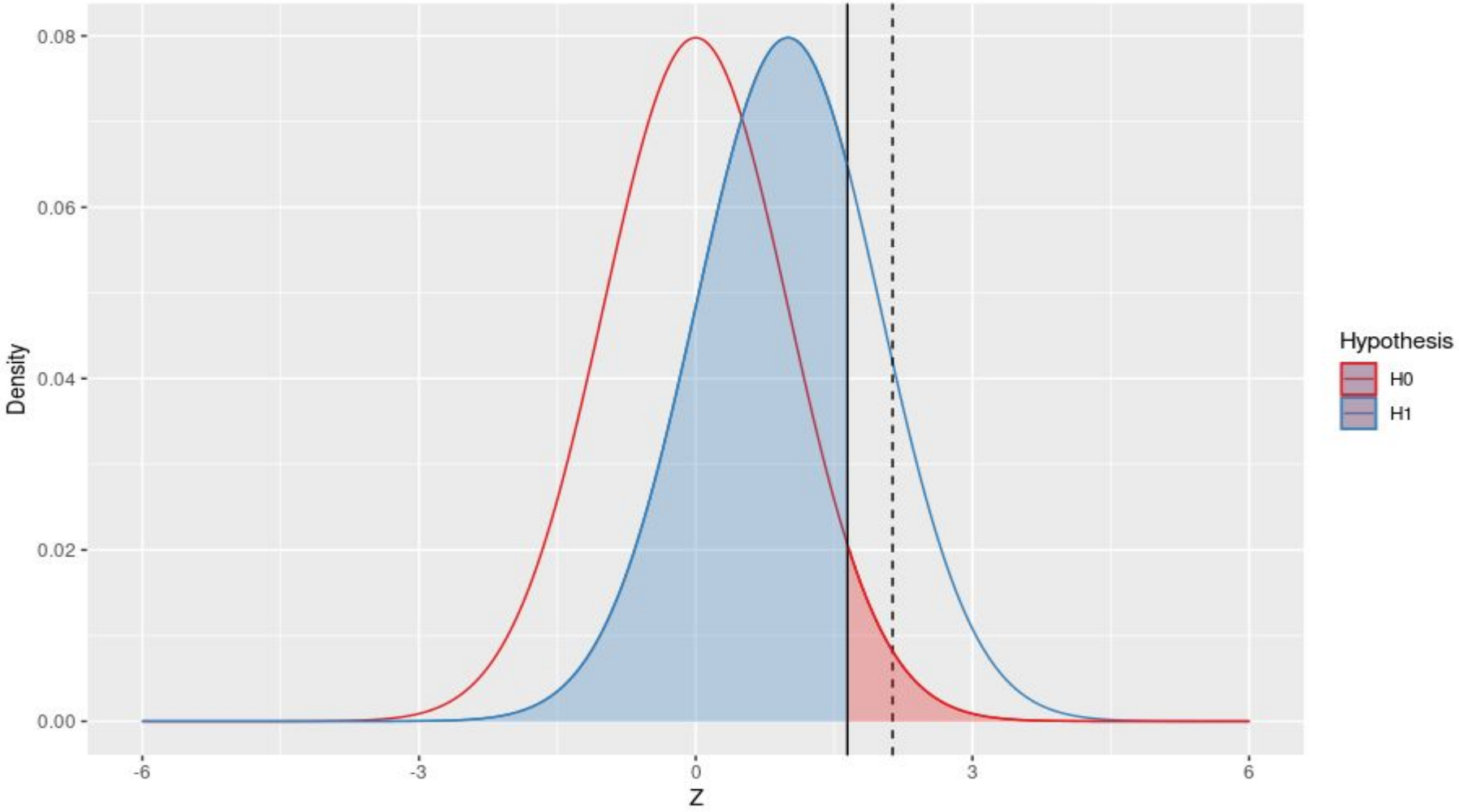
When performing a statistical test in a frequentist setting, you in principle conduct the following procedure:

- Choose a test statistic (that fits your experimental design AND the data AND your assumptions)
- Calculate the test statistics
- Compare it to your **pre-specified** threshold value based on your alpha-level aka Type-I-error rate.
    - If it is above the threshold value: reject the null hypothesis
    - If it is below the threshold value: do not reject the null hypothesis

How to find the threshold value?

- Decide how often - on average - you want to be wrongly rejecting the null hypothesis when performing a hypothesis test.
- Calculate the threshold value based on this percentage and the probability density function of the test statistics under the null.

Calculating the critical threshold - you can do the same for the alternative hypothesis!



Calculating the critical threshold - you can do the same for the alternative hypothesis!

The beta level aka the Type-II-Error is defined as

$$\beta = \int_{-\infty}^{-z_{1-\beta}} f_{H_1}(z)dz = P(Z \leq -z_{1-\beta}) = F_{H_1}(-z_{1-\beta})$$

which can be rewritten as

$$F_{H_1}^{-1}(\beta) = -z_{1-\beta}$$

Note: the beta-level is nothing else than 1 - Power. Put differently:

Power = 1 - Beta.

Putting everything together

If we put everything together (the exercise is left to the audience), we get

$$z_{1-\alpha} = \sqrt{n}\frac{\mu_1 - \mu_0}{\sigma} - z_{1-\beta}$$

which can be rewritten as

$$n = \sigma^2 \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\left(\mu_1 - \mu_0\right)^2}$$

Hence, we have an equation that links the 6 relevant parameters of our testing setting to each other.

Power Calculations: Cookbook recipes for sample sizes

In short, power calculation leverages the relationship of the following six elements that are part - in different forms - of every frequentist test:

- n — sample size: number of **true** replications in the experiment
- $z_{1-\alpha}$ — f(alpha level): rate at which one falsely rejects the null hypothesis **if it is true**
- $z_{1-\beta}$ — f(beta level): rate at which one falsely rejects the alternative hypothesis **if it is true**
- μ_0 — effect size under H_0: unstandardized size of the effect on the outcome of interest if the null hypothesis is true
- μ_1 — effect size under H_1: unstandardized size of the effect on the outcome of interest if the alternative hypothesis is true
- σ^2 — variability: variability of the outcome of interest

$$n = \sigma^2 \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\left(\mu_1 - \mu_0\right)^2}$$

What if you don't know effect sizes and variability?

Problem: In the exploratory setting, you do not know variability and effect sizes (by definition - why else do you need to explore?).

Solution: work with the standardized effect size.

In our previous example, this would be: $$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Hence this

$$n = \sigma^2 \frac{(z_{1-\alpha} + z_{1-\beta})^2}{(\mu_1 - \mu_0)^2}$$

can be simplified into this:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{d^2}$$

Power Calculations: Cookbook recipes for sample sizes

Pros:

- If all elements are in place, pretty straightforward (cookbook recipe)
- Some terrific software packages out there that can help.
- If done correctly, forces you to think clearly about your data and your assumptions.

Cons:

- Many terrible software packages out there (especially online) that do more harm than good.
- Great danger of “cargo cult statistics”: calculating sample sizes based on a test that does in no way fit the experimental design.
- Not all test statistics are equally easy to use as basis for power calculations because not all of them have such nice mathematical properties as the *z*-statistic
- (In the exploratory setting, focus should be less on testing and more on effect size estimation.)

The Resource Equation Method: Correct by chance

How to plan sample size without empirical information?

One suggestion: Resource Equation Method (see e.g. Mead et al. 2012, Statistical Principles for the Design of Experiments, Chapter 0.4)

Basic idea: Equate degrees of freedom available for treatment factors (T), blocks (B) and random error (E) to total sample size minus 1

$$\mathrm{df}_T + \mathrm{df}_B + \mathrm{df}_E = N - 1$$

Recommendation by Mead et al. (2012), p. 546:

- keep error degree of freedom between 12 and 15, but always above 10 and below 20.
- If error degree of freedom is above 20, ask more questions (i.e. increase number of treatments)
- If error degree of freedom is below 10, increase N

Remember the fundamental design equation:

Fundamental experimental design equation:

Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”)

What use are blocks?

Fundamental experimental design equation without blocking (but assuming proper randomisation):

Outcome	=	Treatment Effect + Large Noise (aka “Random Error”)

The “noise” (or “error”) of your model describes the amount of imprecision of of your outcome measure. Can be reduced by **prudently** including biological or technical factors as blocking factors in the design and the analysis.

Outcome	=	Treatment Effect + Blocking Factors (biological or technical) + Small Noise (aka “Random Error”)

An example for the resource equation method

$$\mathrm{df}_T + \mathrm{df}_B + \mathrm{df}_E = N - 1$$

If we have

- 1 comparison: 1 treatment diet vs. 1 control diet which you compare to each other
- 5 blocks (for example: 5 different times of the year)
- 30 mice

Then we have:

- $1 + (5-1) + \mathrm{df}_E = 30 - 1 \rightarrow \mathrm{df}_E = 24$

Good news! We can ask more questions aka test more diets. Let us increase the number of diets to 5 and compare each of them with the control (5 comparisons). Then we have:

- $5 + (5-1) + \mathrm{df}_E = 30 - 1 \rightarrow \mathrm{df}_E = 20$

IMPORTANT: Multiplying the number of comparisons by 5 does usually not mean multiplying the sample size by five.

Why between 10 and 20?

From Mead et al. (2012), p. 546

Now, to obtain a good estimate of error, it is necessary to have at least 10 df, and many statisticians would take 12 or 15 df as their preferred lower limit, although there are situations where statisticians have yielded to pressure from experimenters and agreed to a design with only 8 df. The basis for the decision is most simply seen by examining the 5% point of the *t*-distribution, and using sufficient df so that increasing the error df makes very little difference to the significance point, and hence to the interpretation.

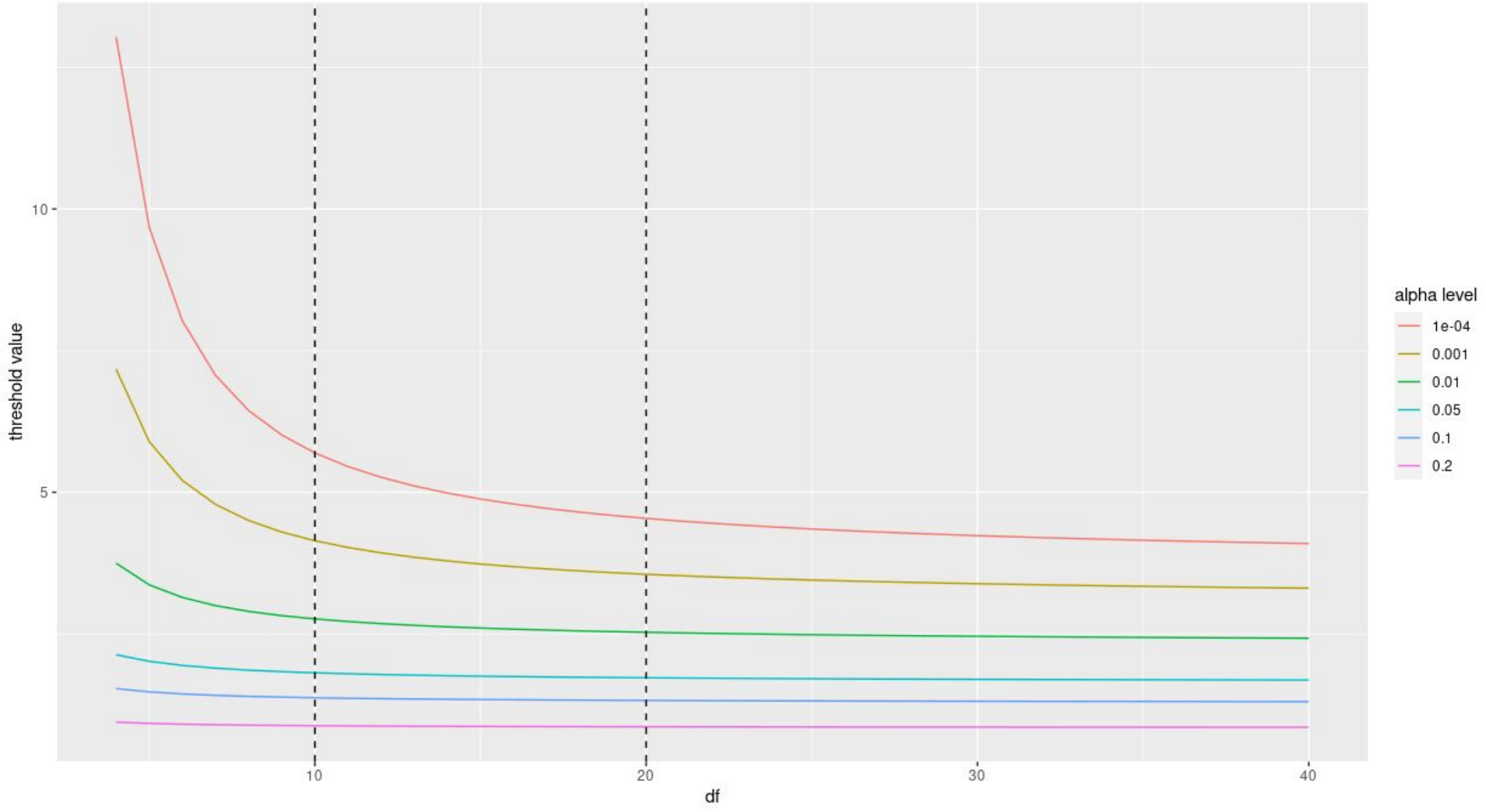
The implications for design are obvious. It is necessary to have at least 10 df for error, to estimate σ^2 and, in choosing the number of treatment question df, we should not normally allow E to fall below 10. But equally, if E is allowed to be large, say greater than 20, then the experimenter is wasting resources by not asking enough questions. If the experiment has too many df for error, then ways should be found of asking more questions about the treatments.

Let us have a look

For a one-sample t-test, the resource equation reduces to:

$$\mathrm{df}_E = N - 1$$

Critical values for one-sample t-test



But: That is only part of the story

Error degrees of freedom do not only affect threshold values but also the value of the test statistic.

Remember the *z*-statistic:

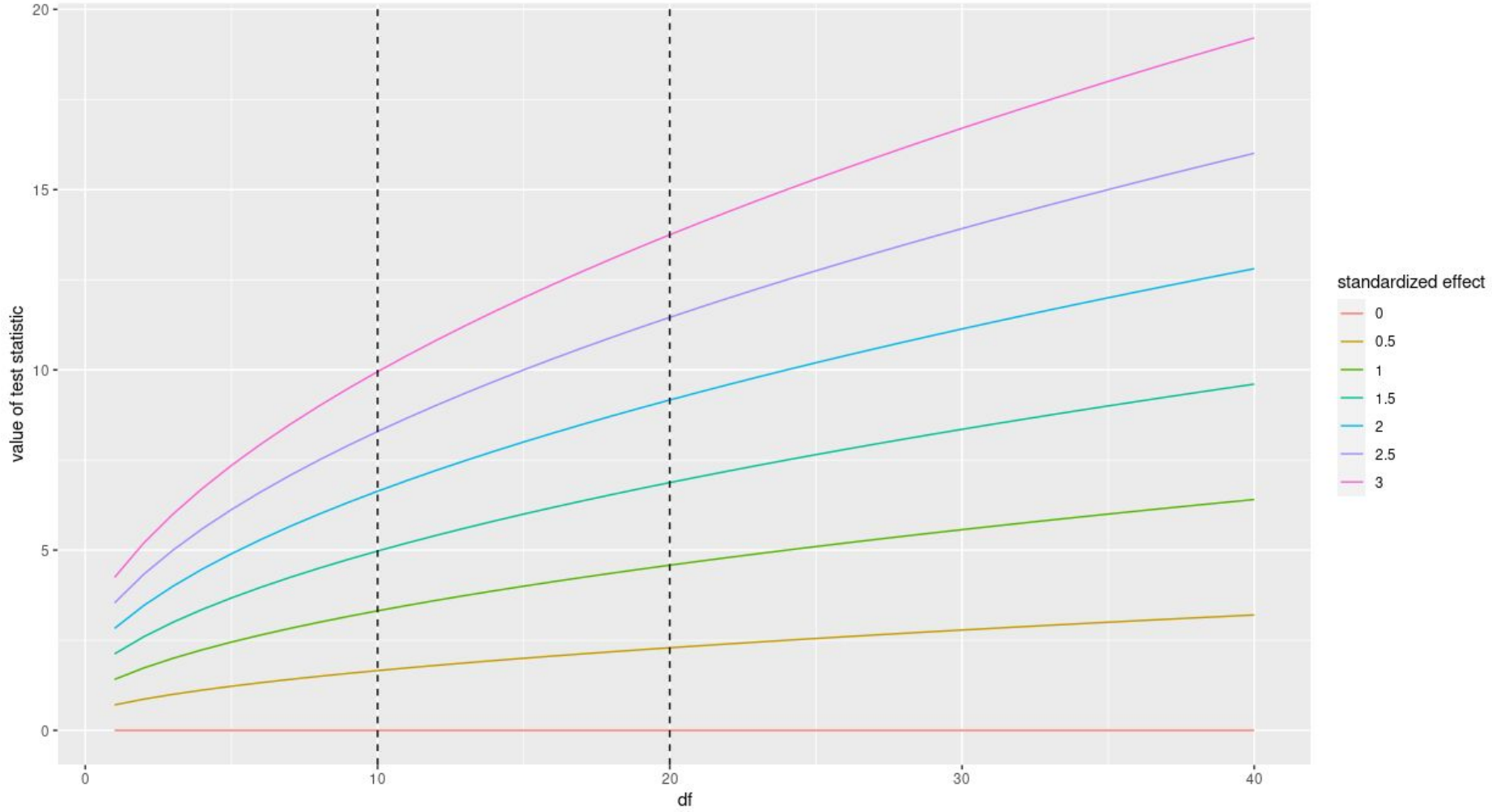
$$Z = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} \sim \mathcal{N}(0,1); \quad \text{if } H_0 \text{ is true}$$

The larger *n,* the larger any given value of the *z*-statistic and the easier it is to reach the pre-specified threshold value.

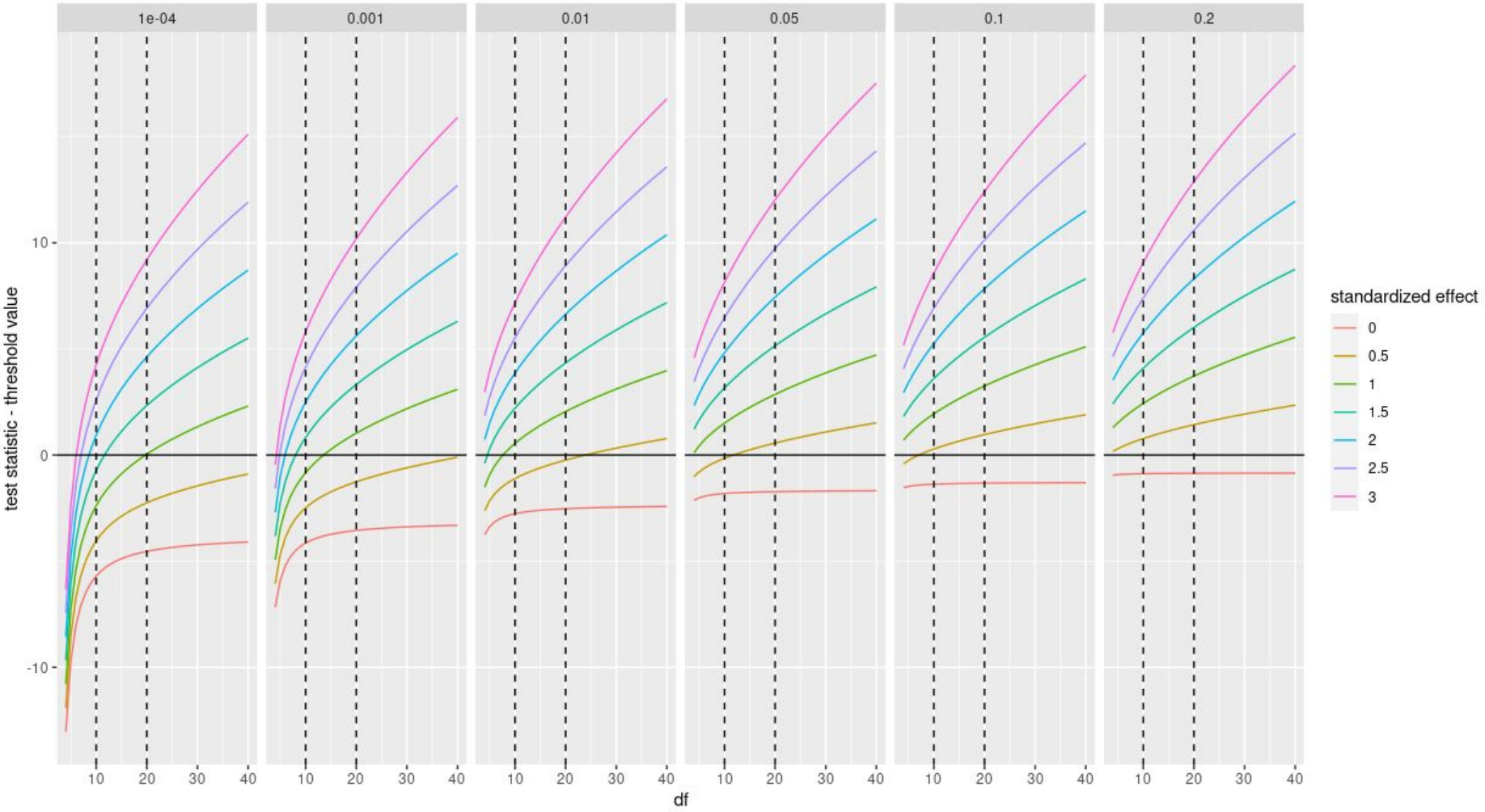
The same holds true for the *t*-statistics

$$t = \sqrt{n}\frac{\bar{X} - \mu_0}{s} \sim t_{n-1}; \quad \text{if } H_0 \text{ is true}$$

t-statistic increases with increasing error degrees of freedom (~sample size)



Does it matter? It depends



The Resource Equation Method

Pros:

- It forces you to think about your treatment and blocking factors early on in the experiment.
- Easy rule of thumb for a first guesstimate of the sample size you need if your design is set.
- Makes clear that doubling the number of treatments does not necessitate doubling the number of animals.

Cons:

- Great danger of “cargo cult statistics”: calculating sample sizes that does not fit well with your design
- If you are capable to set up a proper experimental design you can get a more precise estimate based on that. I you are not capable of doing so, you probably should not be using the resource equation method in the first place.

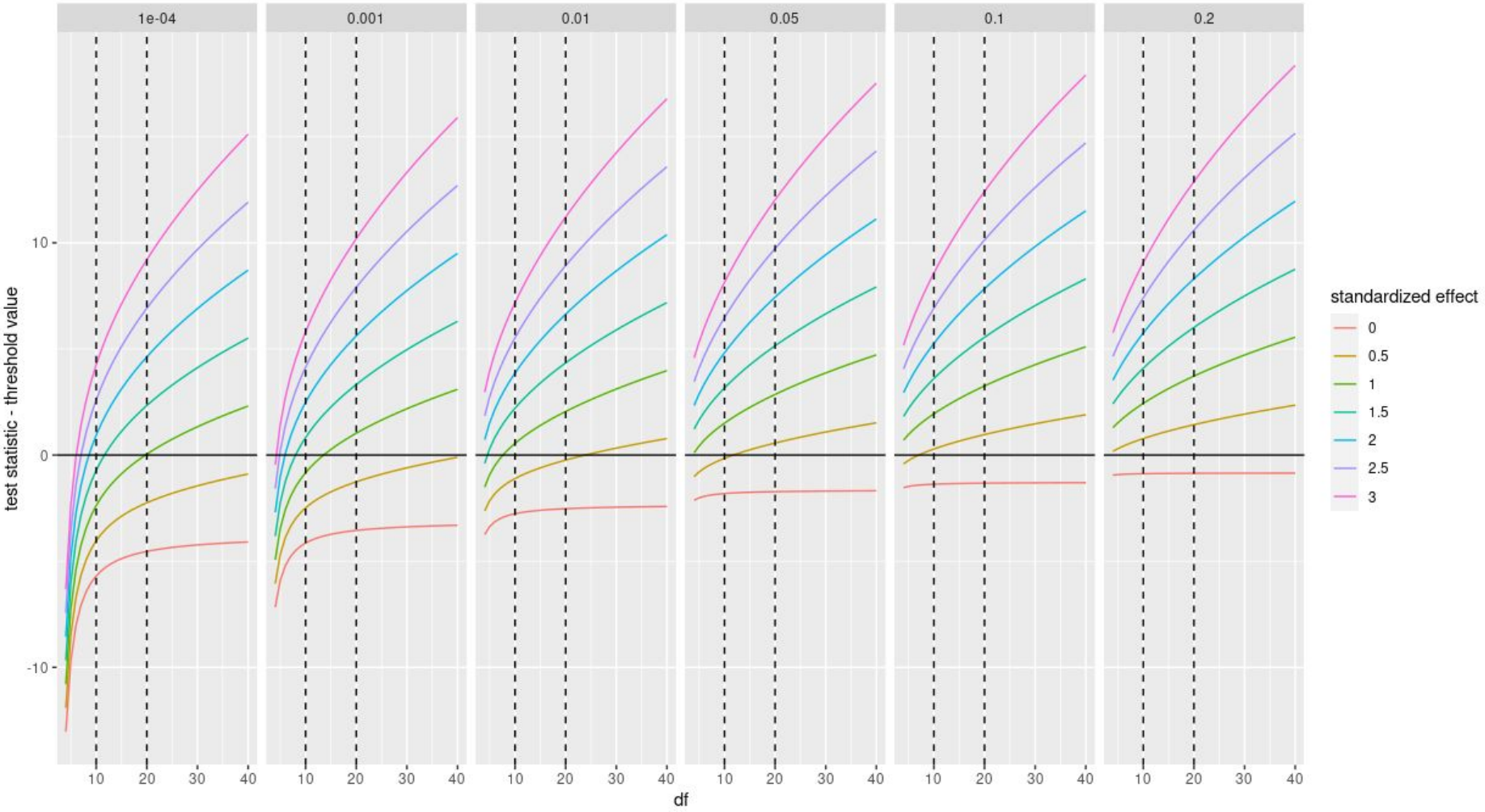
Simulation-based methods: If you don’t have a recipe, you need to improvise

Simulation-based methods: If you don’t have a recipe, you need to improvise

In short:

- Given a design, choose an appropriate test statistic (ideally: non-parametric and robust)
- Simulate your experiment based for a wide range of parameters (effect size, sample size, variability, Type-I and Type-II error) and conduct a hypothesis test after each run.
- Inspect your results visually to see how your experimental setup behaves in different scenarios
- Pick a sample size which suits:
  - your effect size of interest
  - your Type-I error
  - your Type-II error

Simulations - it is not rocket science



Simulation-based methods: If you don’t have a recipe, you need to improvise

Pros:

- If well done, gives a lot of insight into your experimental system.
- Allows you to fine tune and optimize your experimental setup to specific scenarios without needing to spend resources in the lab.
- Analysis is already fully prepared before the experiment has started.
- Can be a good check whether sample size estimations achieved by other means are robust under different scenarios.

Cons:

- Requires quite a lot of knowledge in programming and statistics.

Previous experience: A self-contradictory justification

Previous experience: A self-contradictory justification

Researchers sometimes refer to “previous experience” to justify samples sizes. The general reasoning goes as follows:

- We do not know enough information to perform a power calculation (or other formal sample size estimations).
- However, we successfully conducted experiments in the past with sample size n_{past}.
- Hence, sample size n_{past} is a good sample size for future experiments.

Often, “successful” is used interchangeably with “significant” or “publishable” which is in itself highly problematic due to publication bias, p-hacking, HARKing etc.

But using previous experience as basis for sample size calculations is problematic on a more fundamental level.

Previous experience: A self-contradictory justification

More fundamentally: Claiming to not have enough information to conduct a sample size calculation but using sample sizes from previous experiments as basis for future experiments leads to a contradiction

- **Either** past experiments are too different from future experiments to allow any form of extrapolation -> then this claim extends to the sample size.

- **Or** sample size are informative with regard to sample size -> but then you would need to be able to explain **why** this does not hold for any other property of the experiment (effect size, variability, co-variates etc) that would allow for more formal sample size estimation.

-> You cannot have your cake and eat it, too.

Previous experience: A self-contradictory justification

Pros:

- Easy.

Cons:

- Justification is self-contradictory.
- Can leave researchers over- or underpowered without having any indication in which scenario they are in.
- Does not take into consideration differences in experimental design between experiments.

Should never be accepted as justification in an exploratory setting.

KISS: Why “Keep it Simple, Stupid” is not so simple, but rather stupid

KISS: Why “Keep it Simple, Stupid” is not so simple, but rather stupid

Described by Festing (2017), On determining sample size in experiments involving laboratory animals

Approach, in a nutshell:

- Let the scientist make a provisional estimate of the sample size based on common sense / previous experience and/or the resource equation method
- Calculate the effect size you would be able to detect with this sample size given a certain power and a significance threshold using power analysis.
- If this effect size does not seem to be large enough, they can increase the sample size.
- Then, they can legitimately explain their choice in terms of power analysis.

KISS: Why “Keep it Simple, Stupid” is not so simple, but rather stupid

The problems:

- First, KISS assumes power calculation cannot be done because of lack of information about effect size and variability. However, KISS simultaneously assumes that the researchers *do* have an idea (at least “upon reflection") about what effect size they are interested in -> contradiction.

- Second, if the researchers have idea about effect size of interest - why do KISS? Why not directly perform power calculation?

KISS: Why “Keep it Simple, Stupid” is not so simple, but rather stupid

Pros:

- Emphasizes the exploration of different scenarios: Gateway to simulation-based approaches

Cons:

- Either unnecessarily complicated or useless: If researchers have a notion of the effect size of interest, they can go to power calculations straight away. If they do not have such a notion, then they cannot follow the KISS routine.

The Fermi approximation: A good starting point

The Fermi approximation: A good starting point

Described by Reynolds (2019). “When power calculations won’t do: Fermi approximation of animal numbers”

A set of heuristics - not based on formal statistical considerations but rooted in practical considerations.

Instead of relying on statistical test theory as a basis for sample size criteria, it bases the estimation on other information that are also relevant to plan and conduct an experiment involving animals.

This predominantly includes feasibility assessments such as checking how many animals can realistically be handled by the researchers in a project during a certain time frame or how many animals are needed to harvest enough tissue material for subsequent experiments in vitro.

An example of the Fermi approximation in action

Table 4 | Feasibility calculations (proposed animal numbers versus processing capacity)

Sub-problems	Items	Calculations
Model		
• Processing capacity • Costs	• Number of trained personnel • Total study duration • Number of working hours • Duration of procedure • Cost per procedure • Study budget	• Processing capacity N = (number of animals processed by one person in 1 h) x (number of working hours per work day) x (number of work days per year) x (number of trained personnel) • Processing rate (N per unit time) = (total number requested)/((number of trained persons) x (number of working hours per day) x (number of working days per year)) • Budget assessment = Total budget - ((cost per animal) x (number of animals requested))
Example 1: An investigator requested 500,000 mice for a 3-year project. Is this project feasible?		
• Processing capacity • Costs	• Two trained personnel were identified in the protocol • Assumed working days and hours: 350 days per year; 8–10 hours per day • Procedure time for each animal: 30–60 min	• Number of animals processed per person (lower bound): = 350 working days per year x 3 years x 8 h x 1 processed animal per hour (procedure time 60 min) = 8,400 • Number of animals processed per person (upper bound) = 350 days per year x 3 years x 10 h x 2 processed animal per hour (procedure time 30 min) = 21,000 • Processing capacity (two people) = 16,800–42,000 animals • Processing rate (N/min) = 500,000 animals/ (2 persons x 10 h x 60 min/h x 350 days per year x 3 years) = 0.39 = 1 animal every 3 min per person. • Conclusions: projected numbers were unrealistic, and the protocol could not be approved without modification.

The Fermi approximation: A good starting point

Pros:

- Makes explicit the lower and upper limits of the sample size given **practical** limitations.
- Can be a good check whether sample size estimations achieved by other means are feasible

Cons:

- If used as a basis, **you cannot perform null hypothesis significance testing for inferential purposes** (again: you cannot have your cake and eat it, too)

3. Take Home Messages

The test statistic always has to fit the statistical design!

- Sample size calculation is - usually - depending on the test statistic.
- The experimental design, the treatment scheme you adopt and the data of your outcomes decide which test is suitable.
- NEVER start your sample size calculation with the test statistic. Always do a proper design first and then base the test statistics on that.
- ALWAYS approach a statistician during the design stage of your experiment.

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."
R.A. Fisher

Planning: PREPARE before you ARRIVE

- Follow PREPARE guidelines when planning your experiments:
  - Smith et al. 2017, PREPARE: guidelines for planning animal research and testing
- Follow some basic rules for effective statistical practice when preparing data collection and analysing it:
  - Kass et al. 2016, Ten Simple Rules for Effective Statistical Practice
- Follow ARRIVE guidelines when reporting your experiments (updated in 2020)
  - Percie du Sert et al. 2020, Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0

Some text books I can recommend

- For those who have (almost) no statistical knowledge and want to improve their understanding of fundamental principles of statistical design:
  - Stanley Lazic, Experimental Design for Laboratory Biologists
  - Bate and Clark, The Design and Statistical Analysis of Animal Experiments

- For those who have a basic statistical understanding and want to learn more about efficient experimental design:
  - George Casella, Statistical Design
  - George Box, Improving almost anything

Servan Grüninger, MSc UZH & MSc EPFL
Applied Statistics Group, Dept. of Mathematics, UZH
servan.grueninger@math.uzh.ch
@SGruninger, www.servangrueninger.ch

Appendix slides (NOT presented)

4. How to plan, execute and analyse a study

4. How to plan, execute and analyse an experiment

A. What does it mean to design an experiment?

B. Do you even know what your question is? How to transform scientific hunches into solid study designs.

C. What could possibly go wrong? Biases and how to prevent them.

-> Slides are adapted from joint slides from LTK 13 (created together with Prof. Dr. Reinhard Furrer & Dr. Bernadetta Tarigan) and from personal slides from LTK 22 (held together with Dr. Maike Heimann and Dr. Phil Bugnon).

4.A. What does it mean to design an experiment?

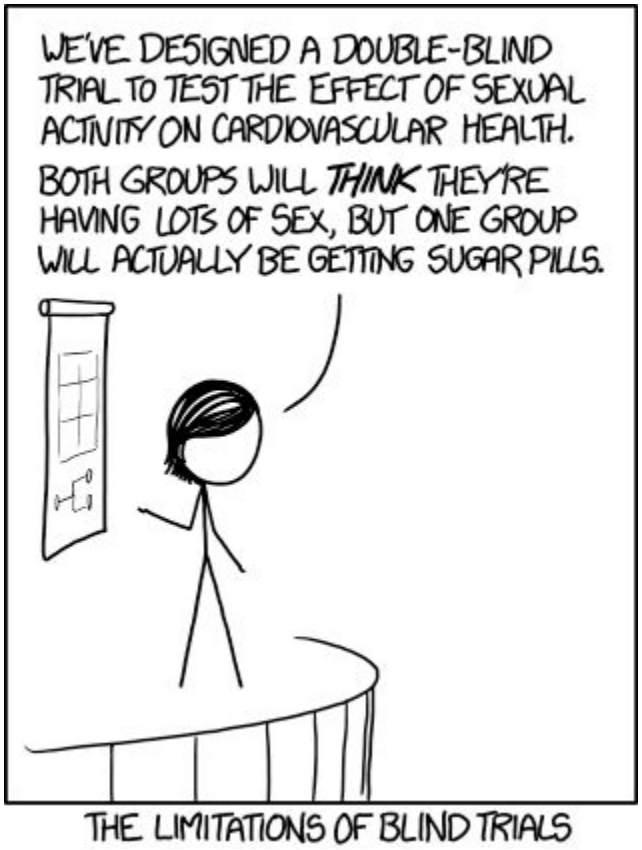
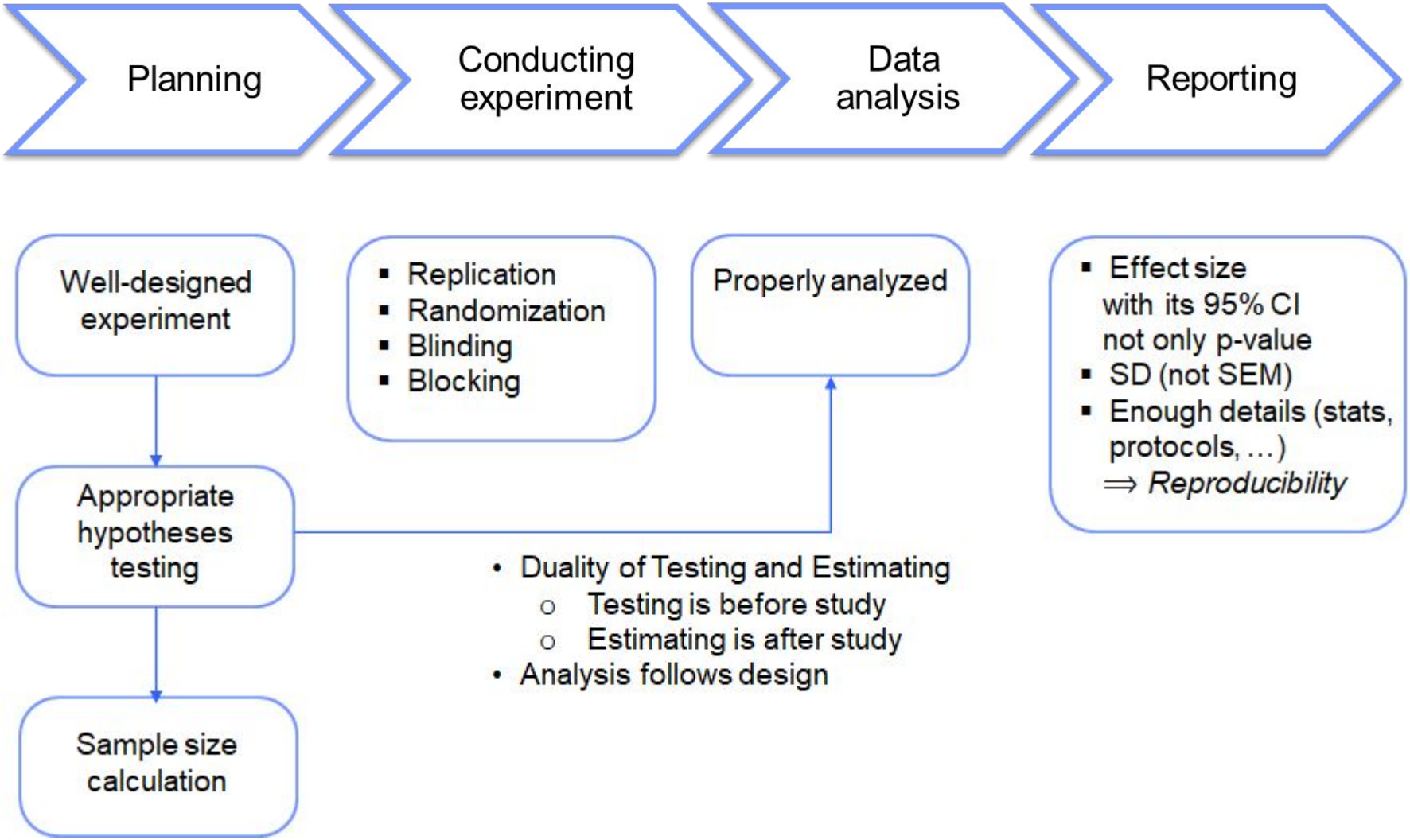


Image: xkcd comics (https://xkcd.com/1462/)

Details of the four stages of experiment



What does it mean to design an experiment?

Fundamental experimental design equation:

Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”)

Conceptually easy, but details depend on many factors:

- Exploratory or confirmatory experiment
- Knowledge about factors and effect sizes
- Available resources

Characteristics of well-defined experiment

Characteristics	How to do it
Clear objective	PICO-B method
Clear definition of EUs	Think about the smallest unit to which you can apply a different treatment
Unbiased	Randomized, Blinding
High precision (low variability)	Replication, Blocking
Able to estimate uncertainty	Replication, Randomized
Wide range of applicability	Blocking (deliberate variation)
Simple	Protect against mistakes

Characteristics of well-defined experiment

Characteristics	How to do it
Clear objective	PICO-B method
Clear definition of EUs	Think about the smallest unit to which you can apply a different treatment
Unbiased	Randomized, Blinding
High precision (low variability)	Replication, Blocking
Able to estimate uncertainty	Replication, Randomized
Wide range of applicability	Blocking (deliberate variation)
Simple	Protect against mistakes

4.B. Do you even know what you're asking?

How to transform scientific hunches into solid study designs?

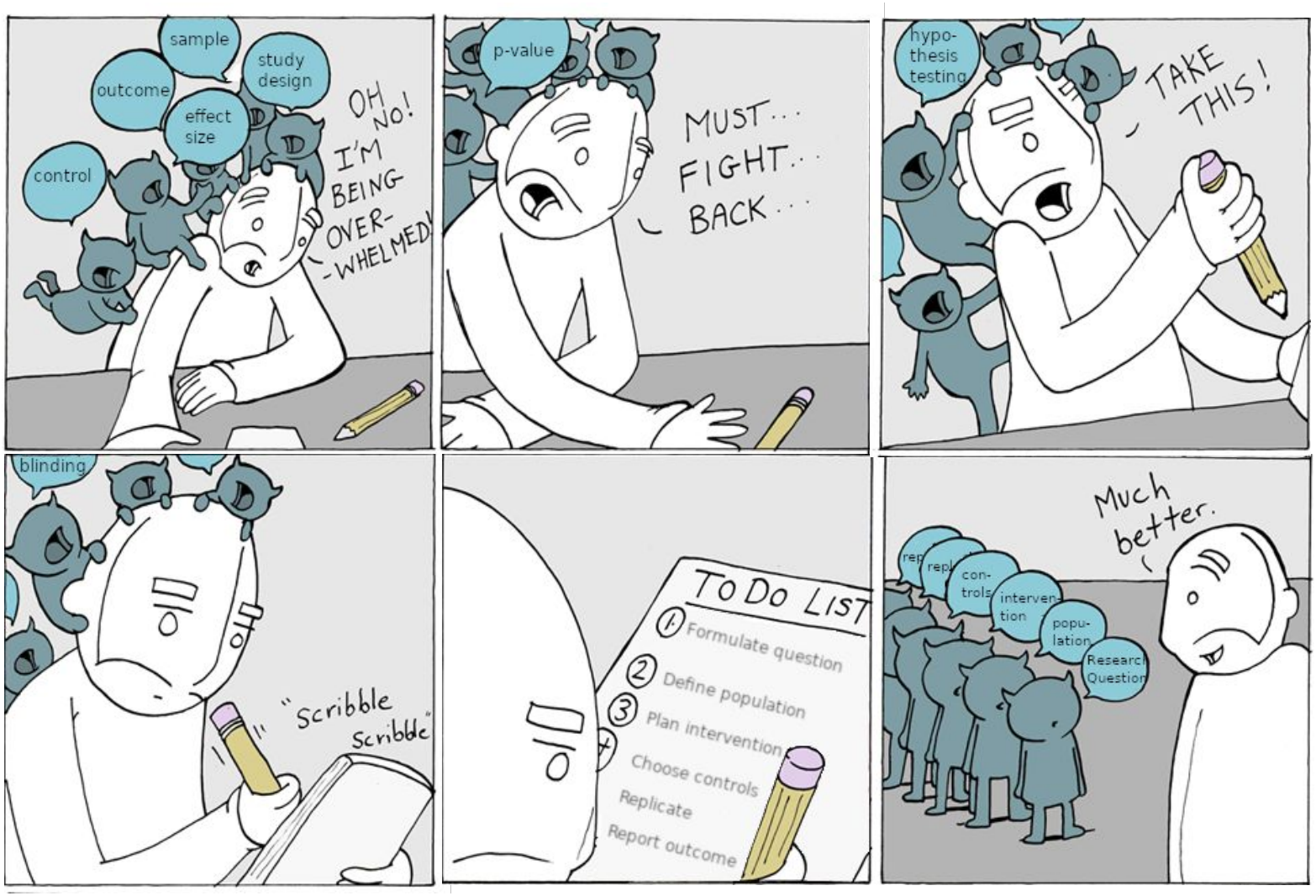


Image: adapted from Lunarbaboon (http://www.lunarbaboon.com/comics/overwhelmed.html)

Good planning is key

- clear objective
- well-defined experim. unit
- simple design
- methods correspond to design
- pre-registration

How to formulate experimental objectives?

***PICO-B* method**	
Population	*Experimental Units (EUs)*
Intervention	*Predictor*
Control	*Predictor*
Outcome	*Response*
Blocking	*Confounders*

How to formulate experimental objectives?

***PICO-B* method**	
Population	*Experimental Units (EUs)*
Intervention	*Predictor*
Control	*Predictor*
Outcome	*Response*
Blocking	*Confounders*

What is “N”?

Types of units to consider

- **Biological unit** (BU) of interest: the entity about which you make inferences, e.g.
  - litter of mouse
  - individual mouse
  - organs
  - parts of organs (such as brain areas)

- **Experimental unit** (EU): the entity that is randomly and independently assigned to one of the treatment levels., e.g.
  - a BU of interest
  - groups of BU
  - parts of BU
  - a sequence of observation of a BU

- **Observational unit** (OU): the entity on which measurements are taken.

adapted from: Lazic (2016). Experimental Design for Laboratory Biologists. Chapter 3, p. 96

What is “N”? -> The experimental unit (EU)

The EU is the unit which has **to be replicated** in an experiment!

Sample size “N” is the **number of proper replications** aka the number of EU.

Increasing the number of OU does **not** increase the sample size unless EU = OU (so-called “pseudo-replication”).

The experimental unit (EU)

How to check whether you identified the correct experimental unit?

1. The EU were independently randomised -- following the chosen randomization scheme such as CRD, RCBD, Split-Plot, etc. -- to the treatment conditions.

2. The treatment must be independently applied to each EU without spill-over to other EUs.

3. The EUs must not **systematically** influence each other **with regard to the experimental outcome.**

For more details, see, Lazic et al. (2018), What exactly is “N” in cell culture and animal experiments? and the NC3R’s entry on experimental units: https://eda.nc3rs.org.uk/experimental-design-unit

How to formulate experimental objectives?

***PICO-B* method**	
Population	*Experimental Units (EUs)*
Intervention	*Predictor*
Control	*Predictor*
Outcome	*Response*
Blocking	*Confounders*

Blocking

Blocking involves dividing the experiment into **a series of (roughly) homogeneous mini-experiments** according to predefined and experimentally relevant criteria.

Reasons to block:

- practical: easier to manage because of technical constraints such as lab equipments, operators, etc.
- statistical: to reduce variability

When you **know** about sources of variability, you should make sure that your intervention and control groups are homogeneous with regard to these sources **and include them into your analysis** in order improve your statistical efficiency.

What use are blocks?

Fundamental experimental design equation without blocking (but assuming proper randomisation):

Outcome	=	Treatment Effect + Large Noise (aka “Random Error”)

The “noise” (or “error”) of your model describes the amount of imprecision of of your outcome measure. Can be reduced by **prudently** including biological or technical factors as blocking factors in the design and the analysis.

Outcome	=	Treatment Effect + Blocking Factors (biological or technical) + Small Noise (aka “Random Error”)

Don’t overdo it with blocks!

The following will often lead to disaster:

Outcome	=	Treatment Effect + Biological Factor 1 + Biological Factor 2 + Biological Factor 3 + Biological Factor 4 + Biological Factor 5 + Biological Factor 6 + Technical Factor 1 + Technical Factor 2 + Technical Factor 3 + Technical Factor 4 + Technical Factor 5 + Technical Factor 6 + Small Noise (aka “Random Error”)

Reasons:

- Risk to introduce dependencies
- Risk to introduce colliders
- Risk of overfitting
- Instead of gaining power, you start losing power (i.e. need more experimental units to answer your question).

Rule of thumb: Add blocking factors that you know have an effect on your outcome variable, randomise the rest.

PICO-B for concrete examples

PICO-B		Example 1	Example 2 Metabolic study
Population	*Experimental Units (EUs)*		
Intervention	*Predictors*		
Control	*Predictors*		
Outcome	*Response*		
Blocking	*Confounders*		

Example 1: A fictional but realistic example

You work with a specific mouse model for a severe bowel disease.

The animal research committee demands that you administer pain killers to your animals.

However, those pain killers of which you know the effect on your mice are out of the question because they would interfere with your research question.

You are left with one drug that has proven to be effective in other bowel disease models but which has never been tested in your specific model.

You neither know whether it is effectively reducing pain nor whether its effect could interfere with your scientific results.

Hence, you are tasked to conduct a pilot study to assess the effects of this drug on a range of different biological parameters.

How do you proceed?

PICO-B for concrete examples

PICO-B		Example 1	Example 2 Metabolic study
Population	*Experimental Units (EUs)*	Mice (specific bowel disease model) of both sexes	
Intervention	*Predictors*	Pain killer	
Control	*Predictors*	No pain killer (negative control) Standard pain killer (positive control)	
Outcome	*Response*	- Severity of bowel disease (e.g. MEIC) - Pain level - Weight loss	
Blocking	*Confounders*	- (Sex) - Weight	

Example 2: Metabolic study

Exploratory questions	Confirmatory questions
Does the types of carbohydrate intake affect rodent metabolism (for how long, with regard to which metabolism measures)?	Does slowly absorbed carbohydrate diet (SAC diet) increase the body fat percentage by at least 15% compared to rapidly absorbed carbohydrate diet (RAC diet) on male mice consume in duration of 38 weeks?

PICO-B for concrete examples

PICO-B		Example 1	Example 2 Metabolic study
Population	*Experimental Units (EUs)*	Mice (specific bowel disease model) of both sexes	Mice or rats: -of different strains -both sexes
Intervention	*Predictors*	Pain killer	Diet 1 / diet 2
Control	*Predictors*	No pain killer (negative control) Standard pain killer (positive control)	Normal-diet
Outcome	*Response*	- Severity of bowel disease (e.g. MEIC) - Pain level - Weight loss	- Weight gain - Body fat (%) - Energy intake - Insulin AUC - Glucose AUC
Blocking	*Confounders*	- (Sex) - Weight	- (Sex) - Weight

4.C. What could possibly go wrong?

Biases and how to prevent them

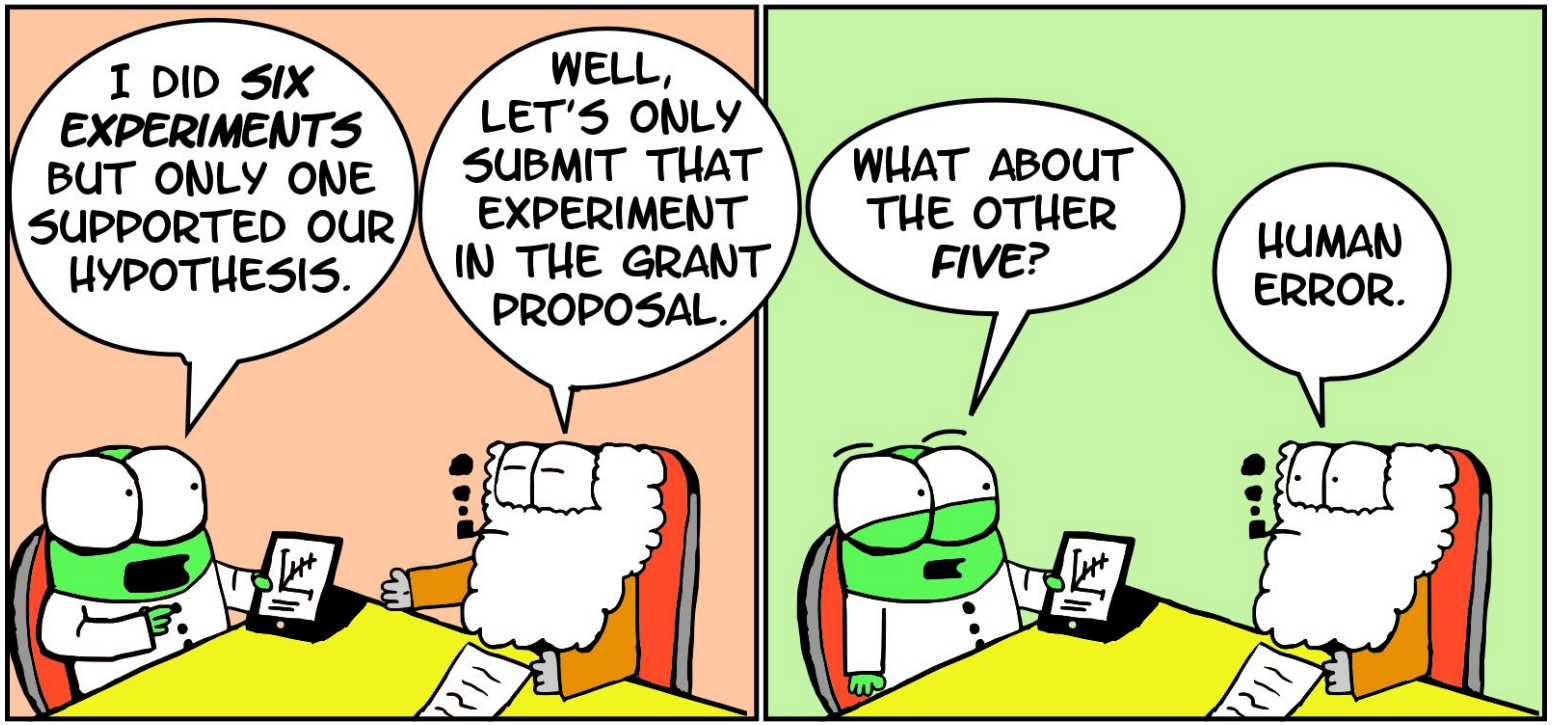


Image: The Upturned Microscope
(https://theupturnedmicroscope.com/comic/logical-fallacies-confirmation-bias)

How to prevent biases during the experiment?

- clear objective
- well-defined experim. unit
- simple design
- methods correspond to design
- pre-registration

- Replication
- Prevent biases
- Correct randomization
- Appropriate blinding

Recall: characteristics of well-defined experiment

Characteristics	How to do it
Clear objective	PICO-B method
Clear definition of EUs	Think about the smallest unit to which you can apply a different treatment
Unbiased	Randomized, Blinding
High precision (low variability)	Replication, Blocking
Able to estimate uncertainty	Replication, Randomized
Wide range of applicability	Blocking (deliberate variation)
Simple	Protect against mistakes

Why these characteristics?

The fundamental equation

Fundamental experimental design equation:

Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”)

- **Treatment effect(s)** are the effect(s) caused by the manipulations or interventions of the experimenter.
- **Random Error** is not a mistake but the **variation** in the outcome that cannot be explained or attributed to the treatment effect or other effect(s), also called the experimental error.
- **Biological effects** are differences that arise from intrinsic properties of the samples and are not actively manipulated by the experimenter
- **Technical effects** are the properties of the experimental system that can influence the outcome: of little interest but may affect the outcome

Any study is subject to uncertainty/error

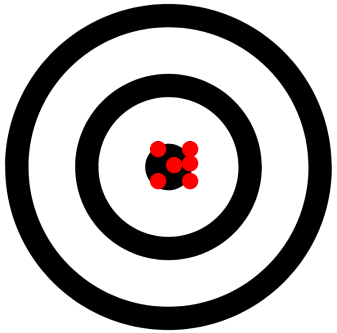
Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”) + Bias (aka “Systematic Error”)

	Source of uncertainty/error	
	Random error almost unavoidable	**Systematic error / bias** can be eliminated
Implication	*Imprecision*	*Low trueness*

Source: ISO 5725-1:1994(en) Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions (https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:en)

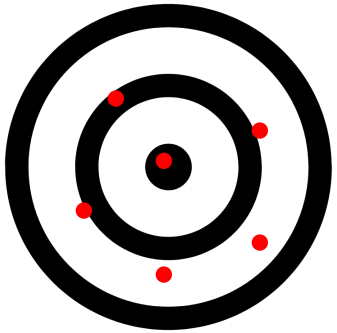
Imprecision and low trueness

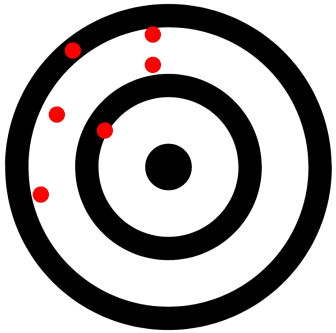
high trueness
high precision



low trueness
high precision

high trueness
low precision





low trueness
low precision

large variability = large random error → low precision
large bias = large systematic error → low trueness

Any study is subject to uncertainty/error

Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”) + Bias (aka “Systematic Error”)

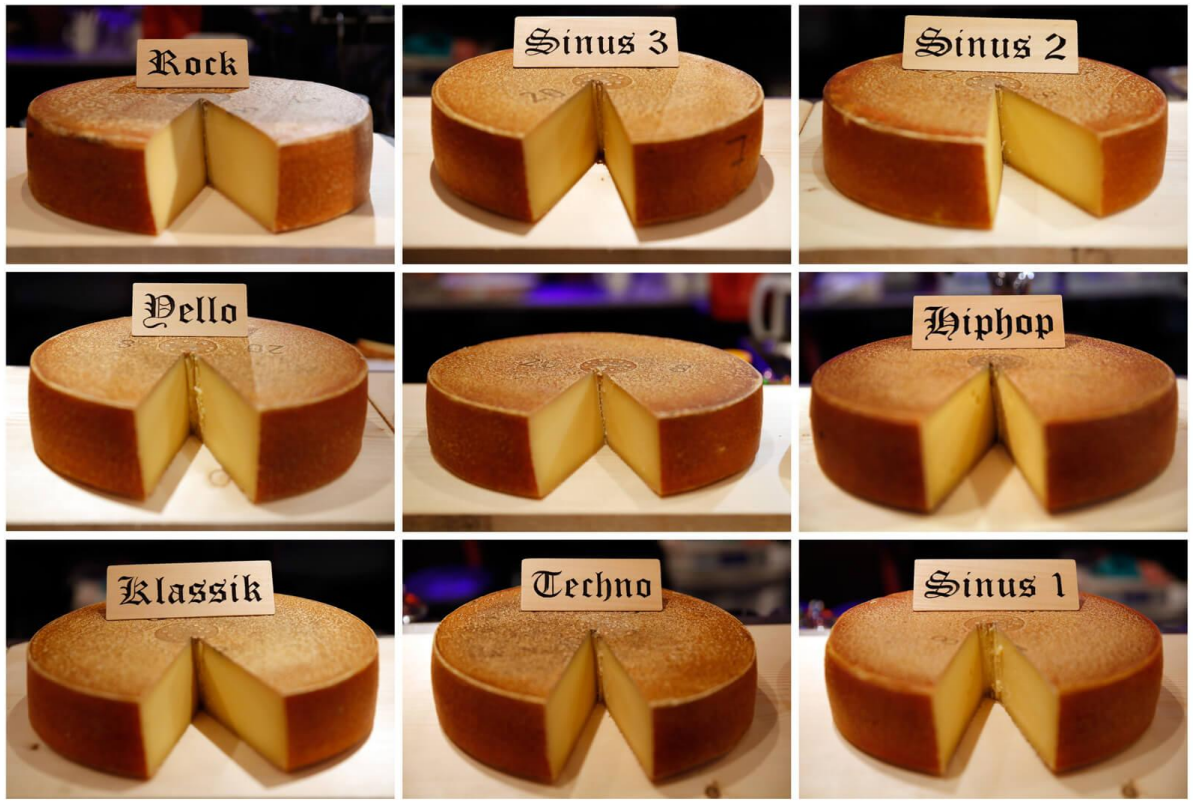
	Source of uncertainty/error	
	Random error almost unavoidable	**Systematic error / bias** can be eliminated
Implication	*Imprecision*	*Low trueness*

-> to measure imprecision, we need replicates!

Source: ISO 5725-1:1994(en) Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions (https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:en)

Without replicates, the results stink

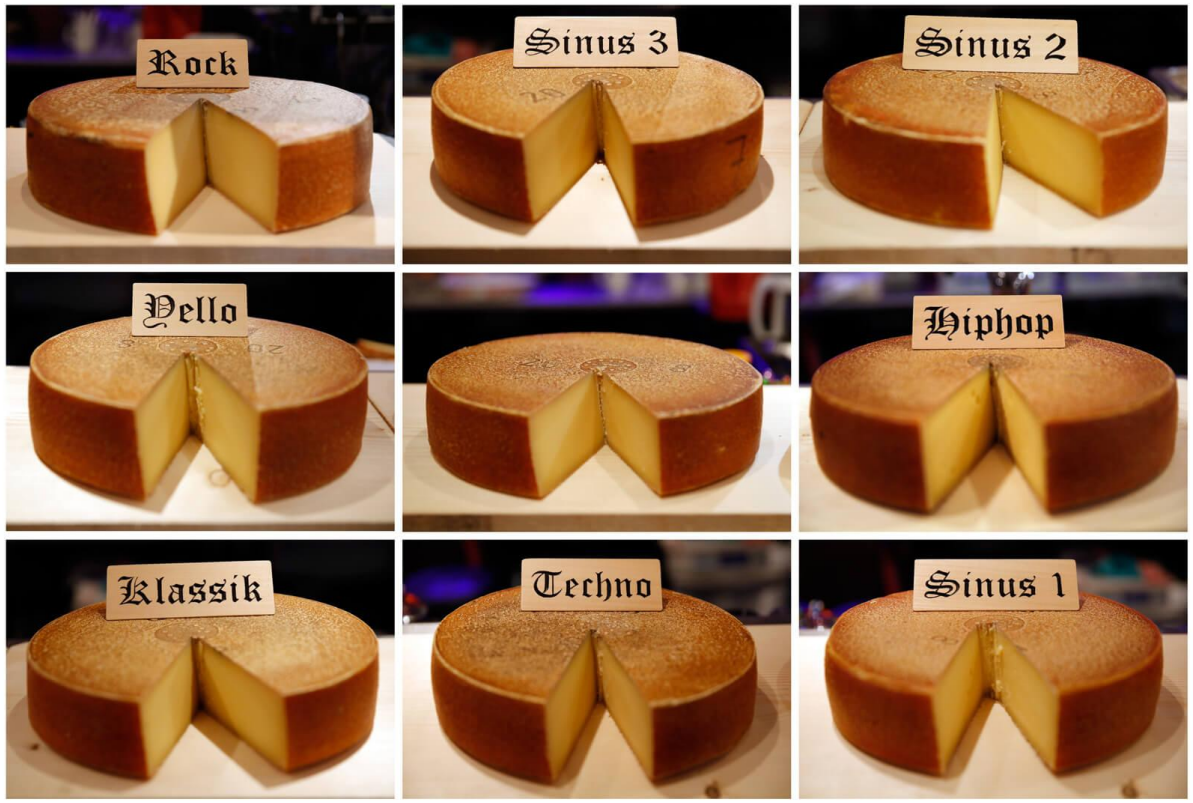
Can hip-hop music make cheese tastier? Yes, says science!



Newly Swissed (https://www.newlyswissed.com/cheese-in-sound/)

Without replicates, the results stink

And to push the idea even further, the project applied eight different types of sounds, one each per cheese wheel:



Newly Swissed (https://www.newlyswissed.com/cheese-in-sound/)

Any study is subject to uncertainty/error

Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”) + Bias (aka “Systematic Error”)

	Source of uncertainty/error	
	Random error almost unavoidable	**Systematic error / bias** can be eliminated
Implication	*Imprecision*	*Low trueness*

-> to remove bias we must properly plan and conduct study!

Source: ISO 5725-1:1994(en) Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions (https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:en)

Page 121

Bias (aka systematic error)

- Bias is any **variation that systematically occurs** in the result
- In other word, a deviation from the truth in the result
- Bias leads to low trueness estimates of the treatment effects to the true effects
- Bias may be introduced at the design of analysis phase of a study
- We consider a couple of biases that can happen at the design phase

see also: van der Worp et al. 2010, Can Animal Models of Disease Reliably Inform Humans?
Catalogue of Bias (https://catalogofbias.org/biases/confounding/)

Selection bias

It is a bias caused by non-random allocation of EUs to treatment groups

Examples:

- Allocating the less healthy animals to the high-dose group
- Allocating the more healthy subjects to intervention group
- Allocating the males to the intervention, the females to the control group

Solution: randomization

Performance bias

It is a bias caused by differences in care given to the subjects across treatment groups.

It can happen in two scenarios:

1. If researchers provide -- intentionally or unintentionally -- unequal care to subjects in different groups
2. If subjects in different groups behaved differently

Examples:

- The technician in laboratory animal gives *unequal husbandry care* for the mice in the intervention group than in the control group
- The animals were *not randomly housed* during the experiment
- A study in a weight-loss trial of a special counselling programme compared to a usual care in general practice. As participants in the control group were *disappointed* for being offered usual care instead of the new helpful program, the study concluded that their reaction to disappointment may introduce performance bias.

Solution: blinding, randomization

Observer/detection bias

It is a bias caused when the person assessing the outcome knows which treatment group the subject is assigned

Examples:

- Was the outcome assessor blinded?

  When assessing animal behavior, it is human nature to want to see a positive effect in your experiment

- Were animals selected at random for outcome assessment?

Solution: blinding, randomization

Attrition bias

Attrition means a reduction or decrease in numbers

Attrition occurs when participants leave during a study

Systematic differences between people who leave the study and those who continue can introduce bias into a study's results – this is attrition bias

Examples:

- Were incomplete outcome data adequately addressed?
- Study of psychosocial factors among patients with cardiac conditions showed that those who fully completed the study differed in clinical and psychosocial features from those who dropped out before the study ended. Such differential attrition could have biased the study's results.

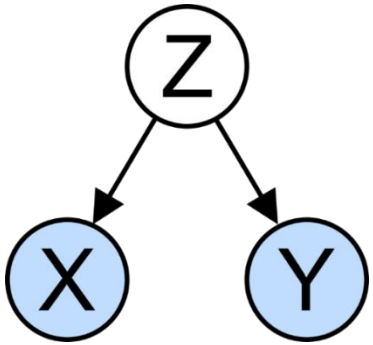
Solution:

- use the so-called "intention-to-treat" analysis (as opposed to "as-treated" analysis)
- take into account dropout rate in sample size calculation

Confounding bias

It is a bias due to another factor that distorts the relationship between treatment and outcome

Confounder is a variable that influences both the dependent variable and independent variable, causing a spurious association



Examples:

- Recall the metabolism study where the treatment levels are two types of certain diets and the outcome is the body weight gain. We can think that the treatment effect could be confounded by the initial weight.
- Can you think of an example?

Solution: blocking, randomization

Summary of biases

Type of Bias	Description	Solution
Selection	Bias caused by non-random allocation of EUs to treatment groups	Randomization
Performance	Bias caused by differences in care given to subject across treatment groups	Blinding, randomization
Observer/ Detection	Bias caused when the person assessing the outcome has knowledge of treatment assignment	Blinding
Attrition	Bias caused when participants drop out from a study and loss to follow-up across treatment groups	Intention-to-treat analysis, add dropout rate in sample size calculation
Confounding	Bias due to another factor that distorts the relationship between treatment and outcome	Blocking (reducing variability), randomization

What about your unknown knowns and unknown unknowns?

You will most likely never know all factors that influence your outcome variable.

Page 129

What about your unknown knowns and unknown unknowns?

Known Knowns Factors of which you are aware and of which you know the effect	**Unknown Knowns** Factors of which you are not aware but of which somebody else knows the effect
Known Unknowns Factors of which you are aware but of which you don't know the effect	**Unknown Unknowns** Factors of which you are not aware and of which nobody knows the effect
-> Handle with blocking & including factors in design and analysis	**-> Handle with randomisation and blinding**

Any study is subject to uncertainty/error

Outcome	=	Treatment Effect + Biological factors + Technical factors + Noise (aka “Random Error”) + Bias (aka “Systematic Error”)

	Source of uncertainty/error	
	Random error almost unavoidable	**Systematic error / bias** can be eliminated
Implication	*Imprecision*	*Low trueness*
Improved by	**Replications** **(least needed)** **Blocking**	**Randomization** of assignment **Blinding** **Blocking**

Source: ISO 5725-1:1994(en) Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions (https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:en)

Assessment of bias (risk of bias)

There are multiple helpful tools to help you assessing possible risks of bias, for example:

- Cochrane RoB page
- RoB tools
- Risk of Bias (RoB) tool in animal study: SYRCLE

A systematic assessment of different risk of bias tools can be found in Page et.al (2017)

Have we removed all the biases?

It’s impossible to control for all potential sources of bias on your outcome - simply because you might not even know what is influencing your outcome (“unknown knowns” and “unknown unknowns”).

Solution: randomize

Randomization

Assign experimental units to treatment groups by chance alone
→ independently from each other
→ following a pre-defined probabilistic assignment rule (often: identically with the same probability/chance)

Randomization ensures that – on average – the only systematic difference between the groups is the treatment.

How to randomize?

- Never do it by “hand”
- Use computer to randomize your experimental units

Type of randomization:

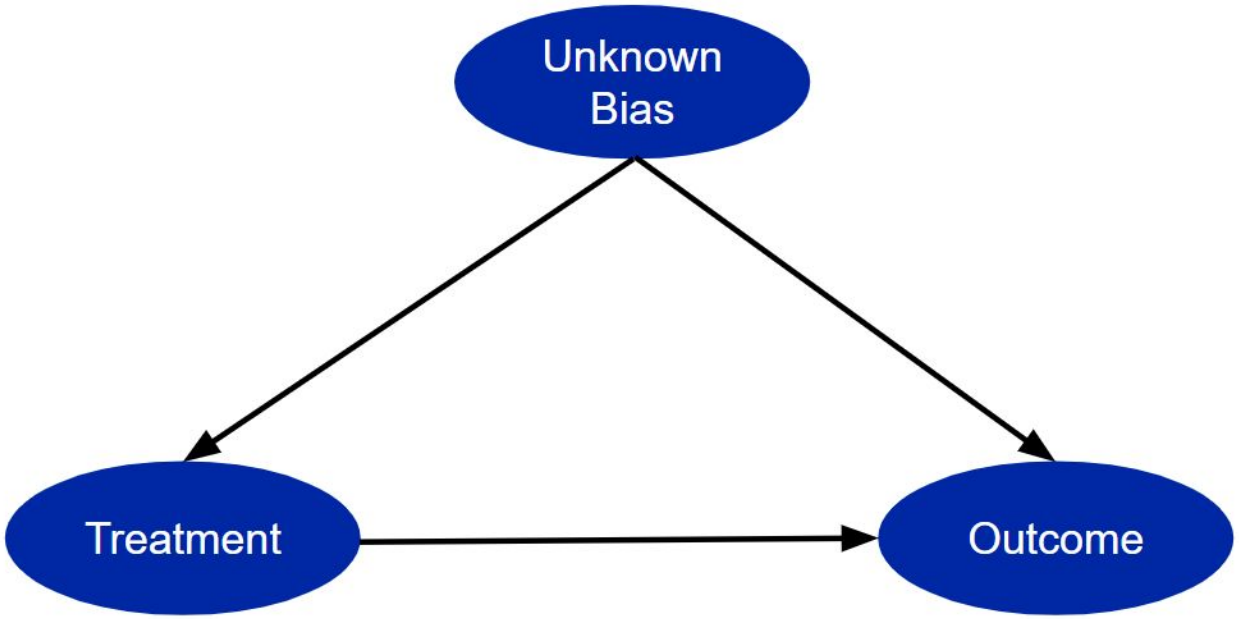
- Simple randomization:
  - strictly simple (might be unbalanced)
  - balanced
- Stratified/block randomization

Randomization - some common misunderstanding

- Randomization is **not** about “balancing” co-variables or factors.
- You **can** include co-variables and factors into your analysis even if you haven’t randomized.
- It is important **what** you randomize: You should randomize the treatment allocation to your experimental unit.
- Randomization “breaks” the link between a co-variable or a factor and your treatment. It does **not** break the link between a co-variable or a factor and your outcome.

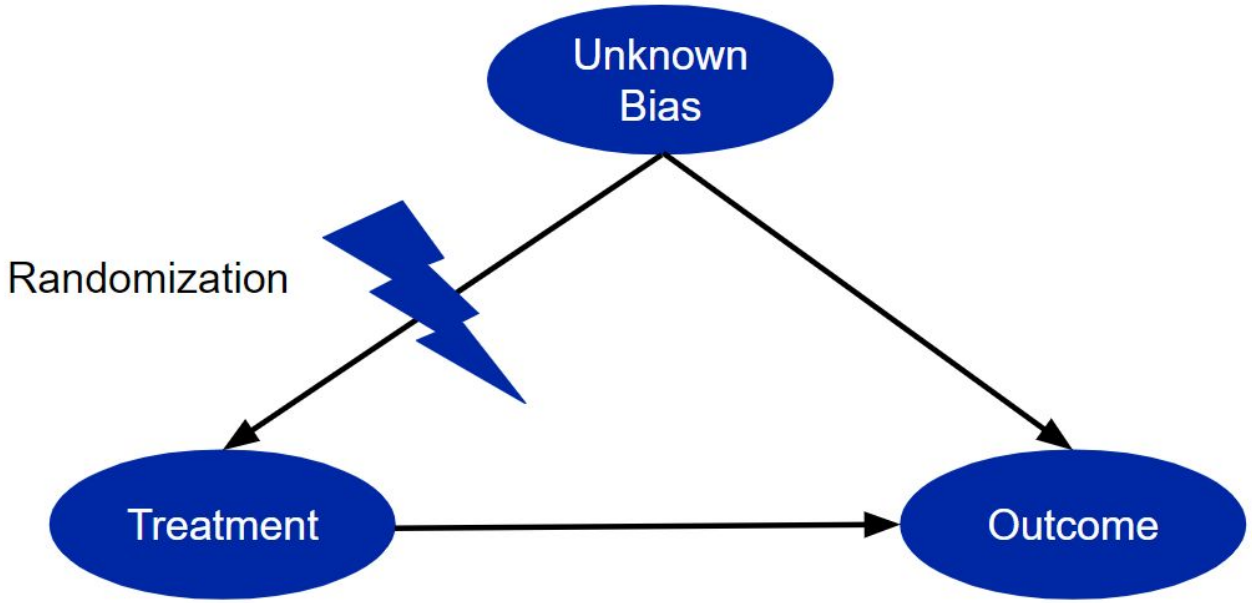
The inferential power of randomization

In this setup, you can't distinguish between the effect of the unknown bias and the treatment on the outcome. In this setup, you can't distinguish between the effect of the unknown bias and the treatment on the outcome.



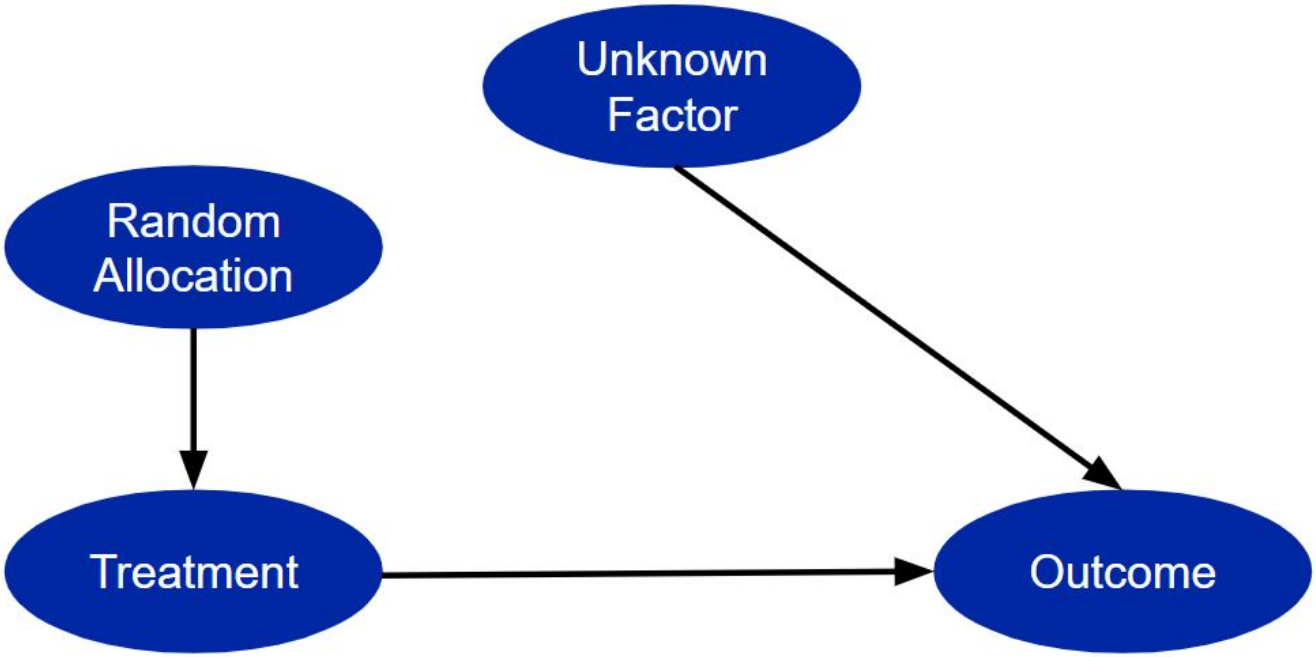
The inferential power of randomization

Randomization **of the treatment allocation to the experimental unit** breaks the link between the unknown bias and the treatment.



The inferential power of randomization

In this setup, you the effect of the unknown factor onto your outcome is still present, but you can distinguish between its effect on the outcome and the effect of the treatment on the outcome. Hence, the **bias** turned into a factor / co-variable that is **independent** of your treatment allocation.



Randomization

Assign experimental units to treatment groups by chance alone
→ independently from each other
→ identically with the same probability/chance

Randomization ensures that – on average – the only systematic difference between the groups is the treatment.

How to randomize?

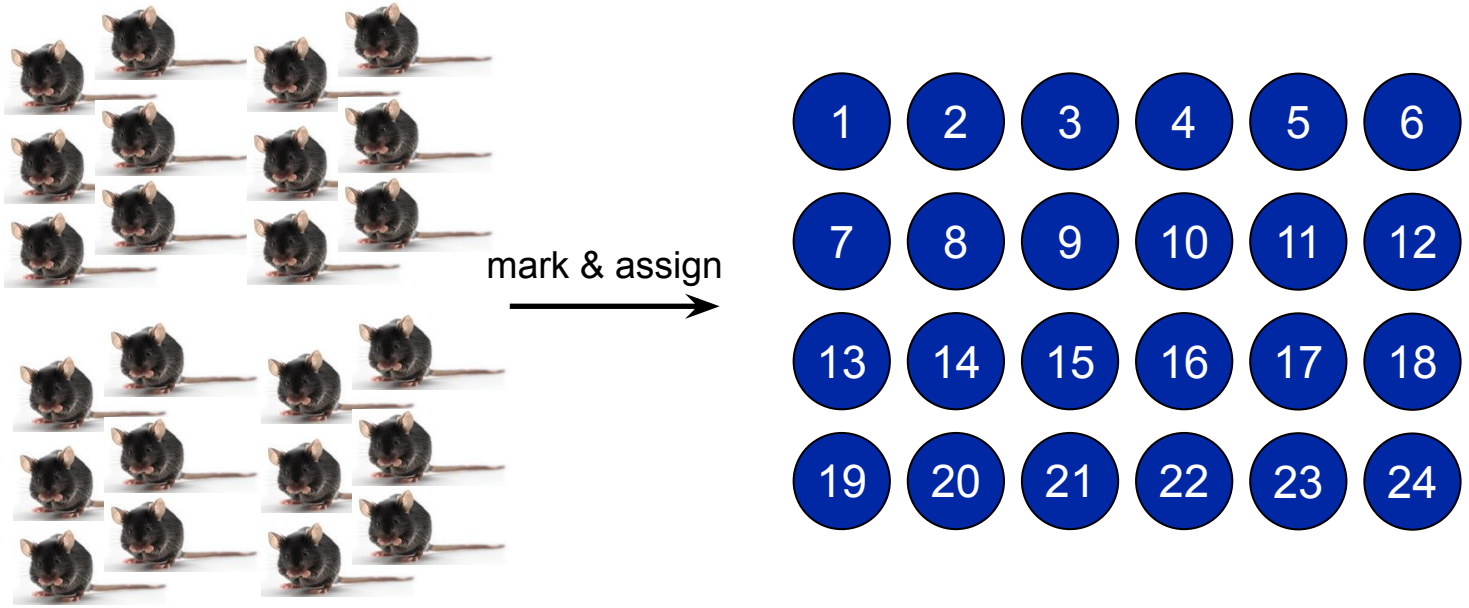
- Never do it by “hand”
- Use computer to randomize your experimental units

Type of randomization:

- Simple randomization:
  - strictly simple (might be unbalanced)
  - balanced
- Stratified/block randomization

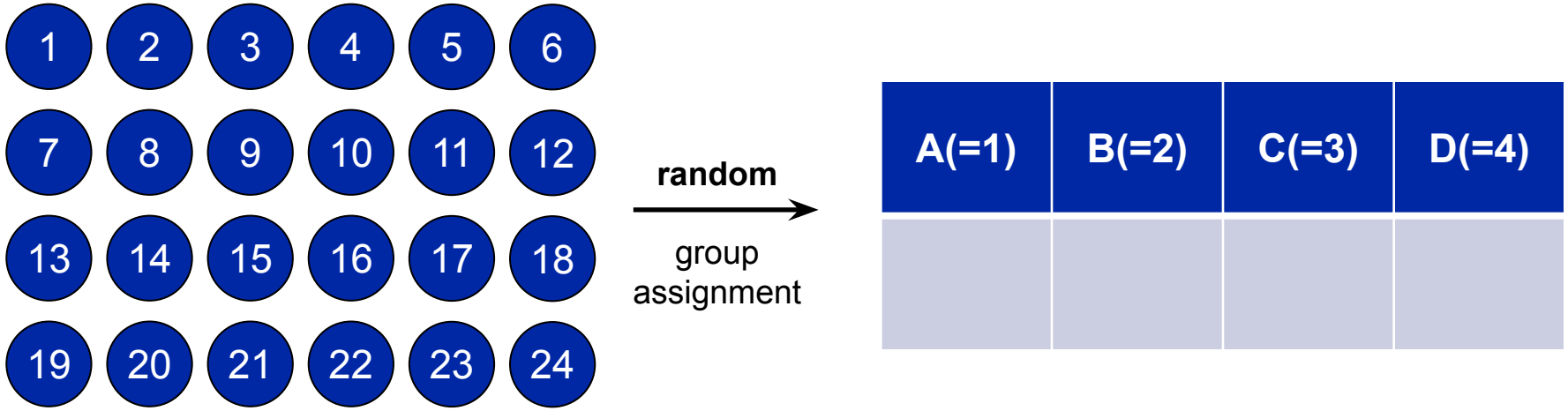
Simple randomization

- Remember example 2 (metabolic study) with 3 treatment groups and 1 control group; assume you have 24 mice.
- Goal: Assign each mouse randomly to one of the four groups



Simple randomization

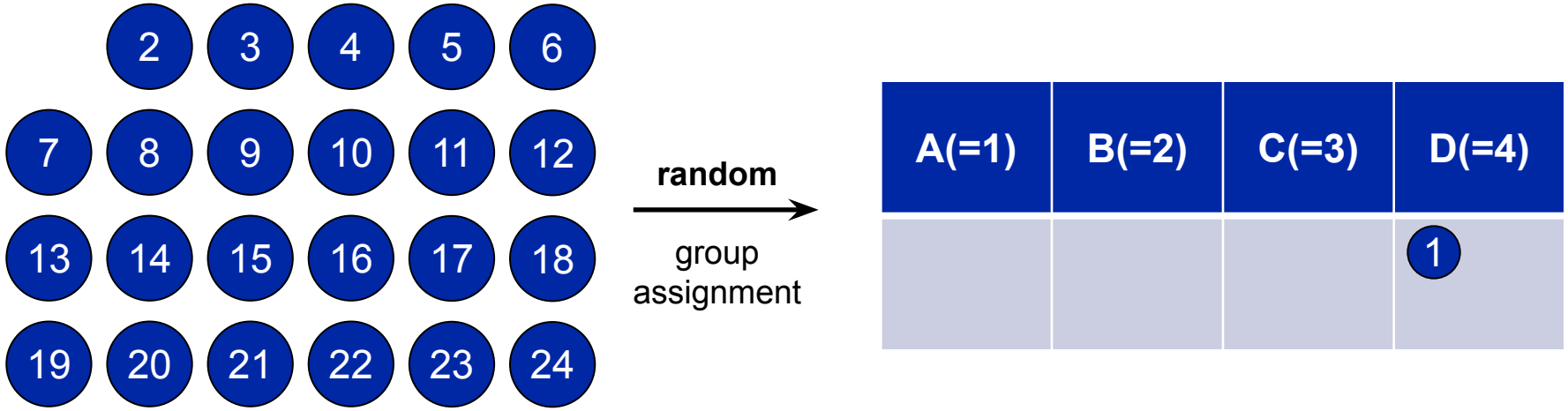
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

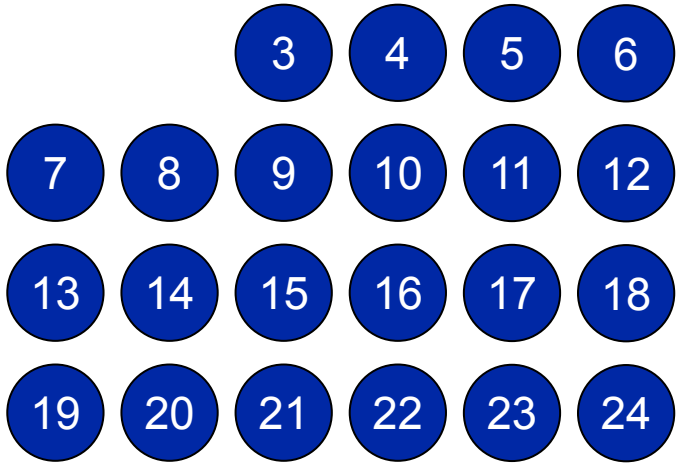
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

Take each mouse and randomly assign it to one of the four groups



random

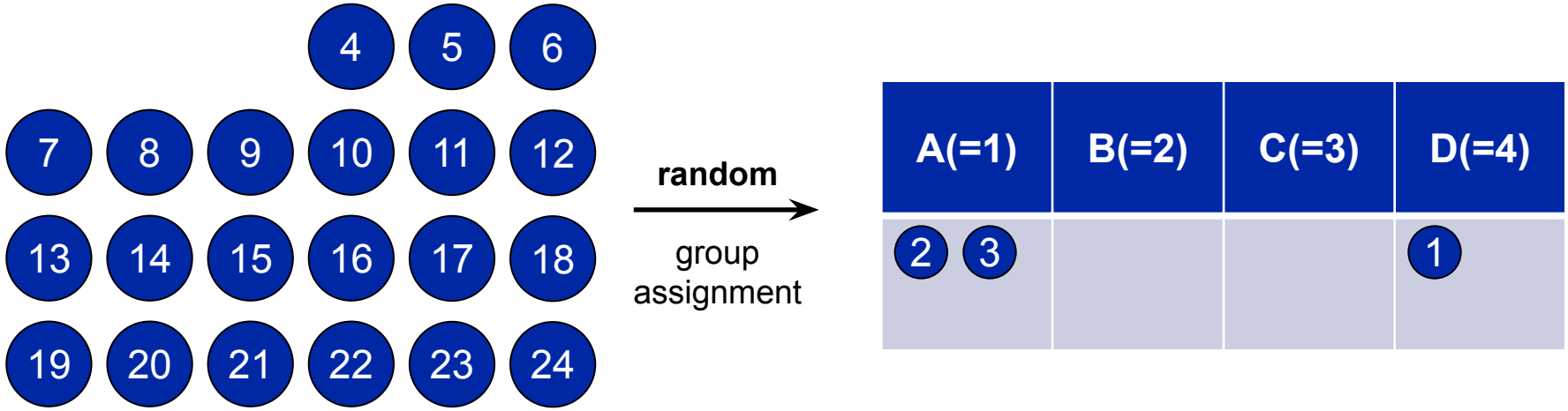
group assignment

A(=1)	B(=2)	C(=3)	D(=4)
2			1

In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

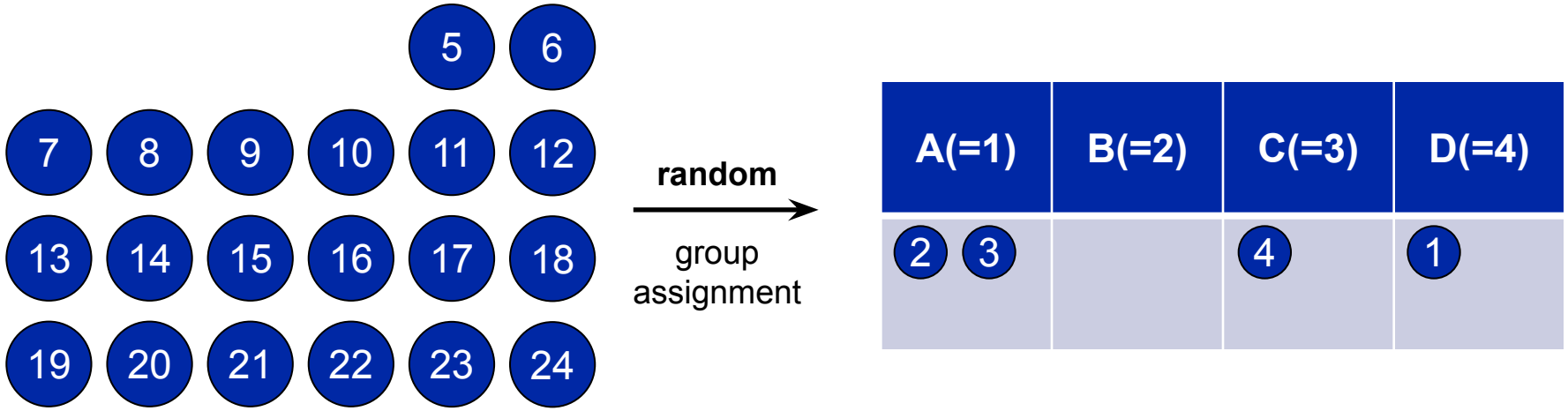
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

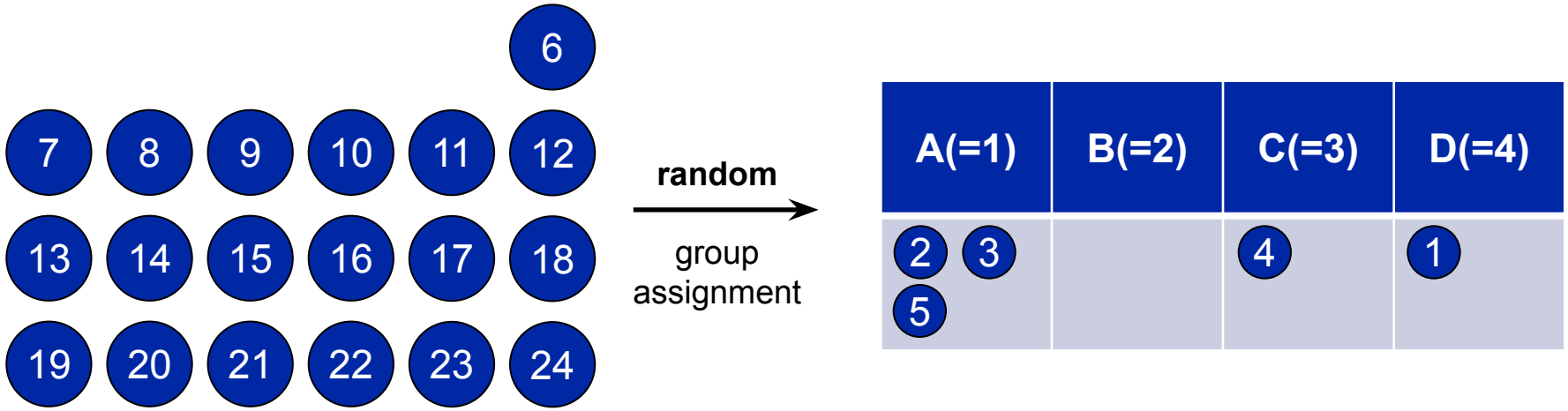
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

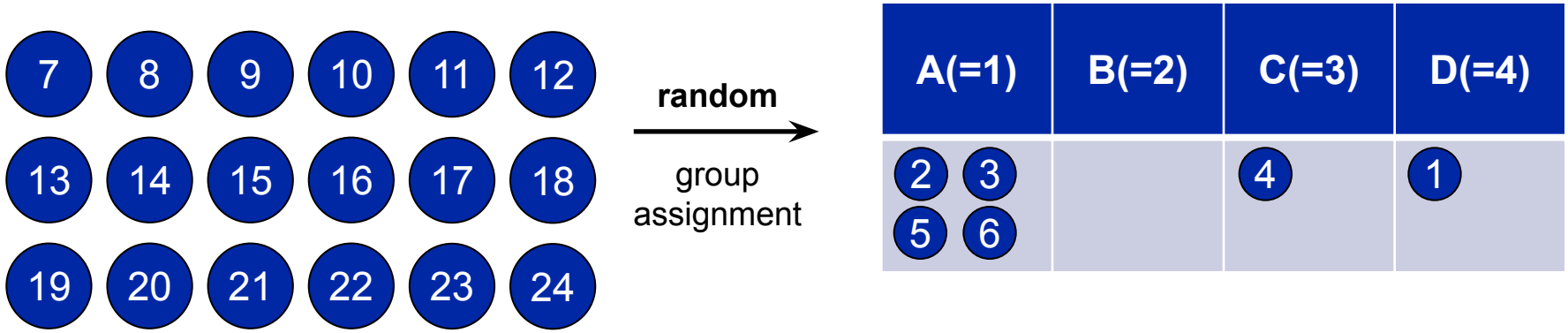
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

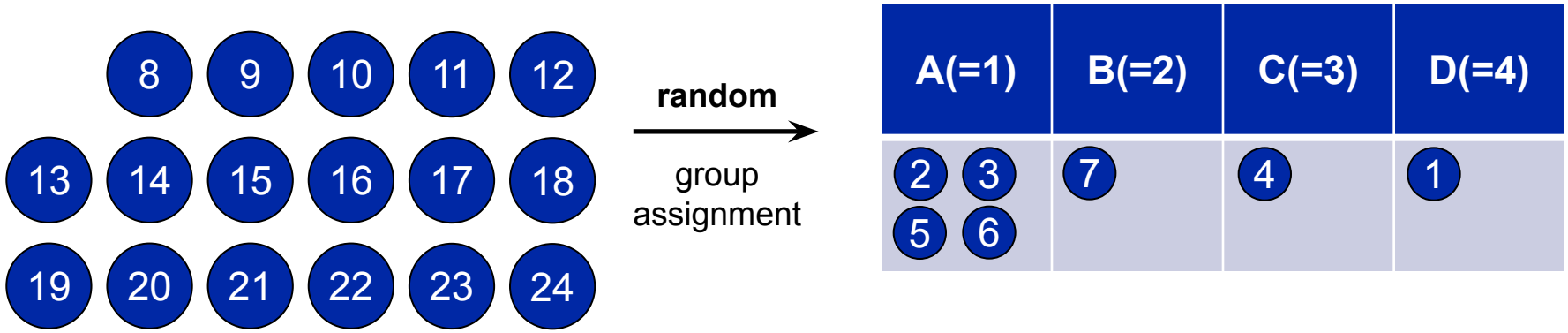
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

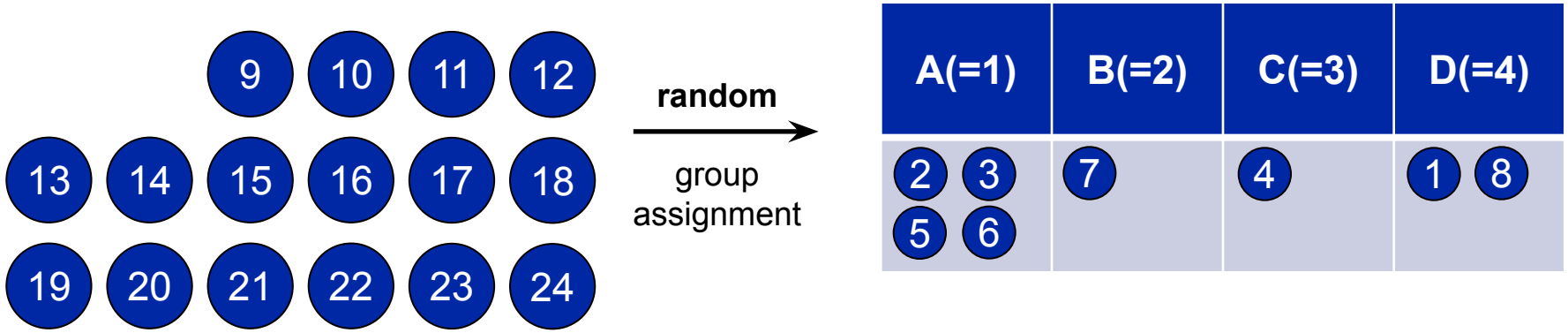
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

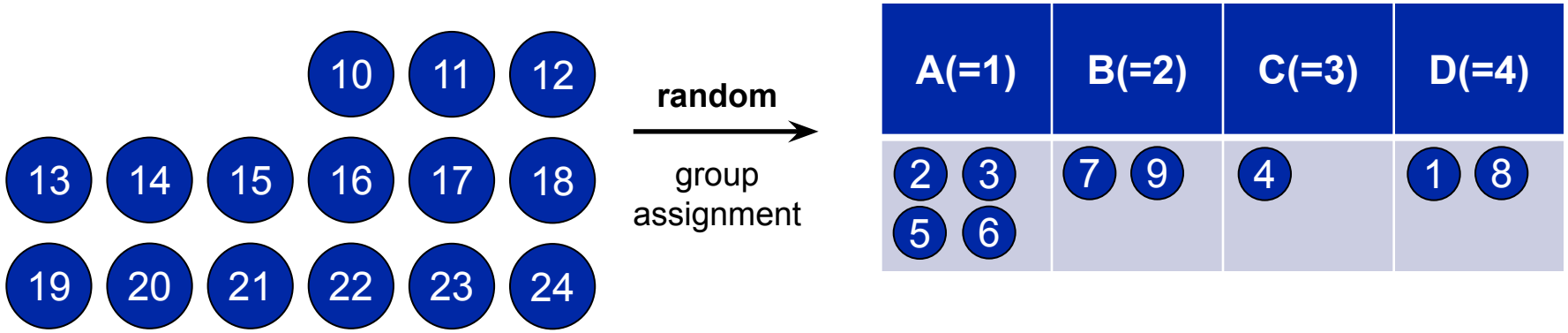
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

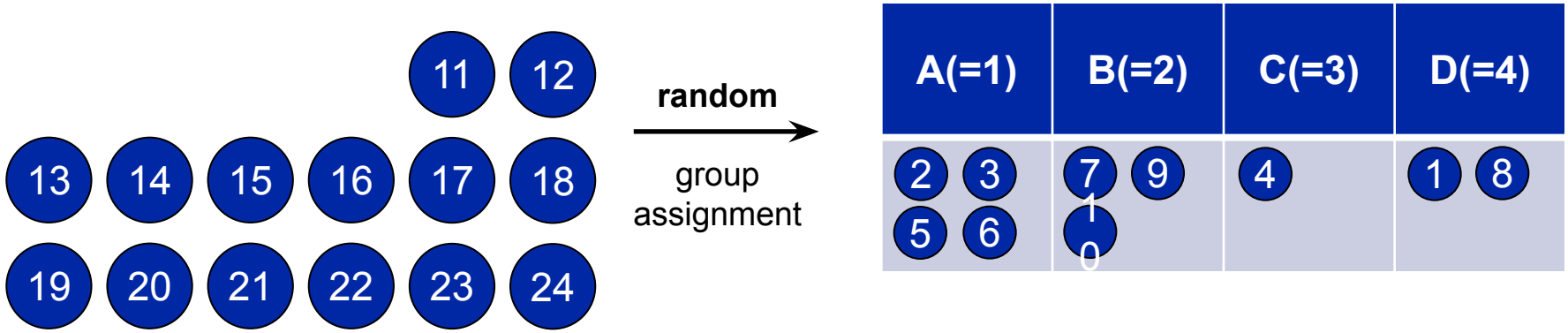
Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

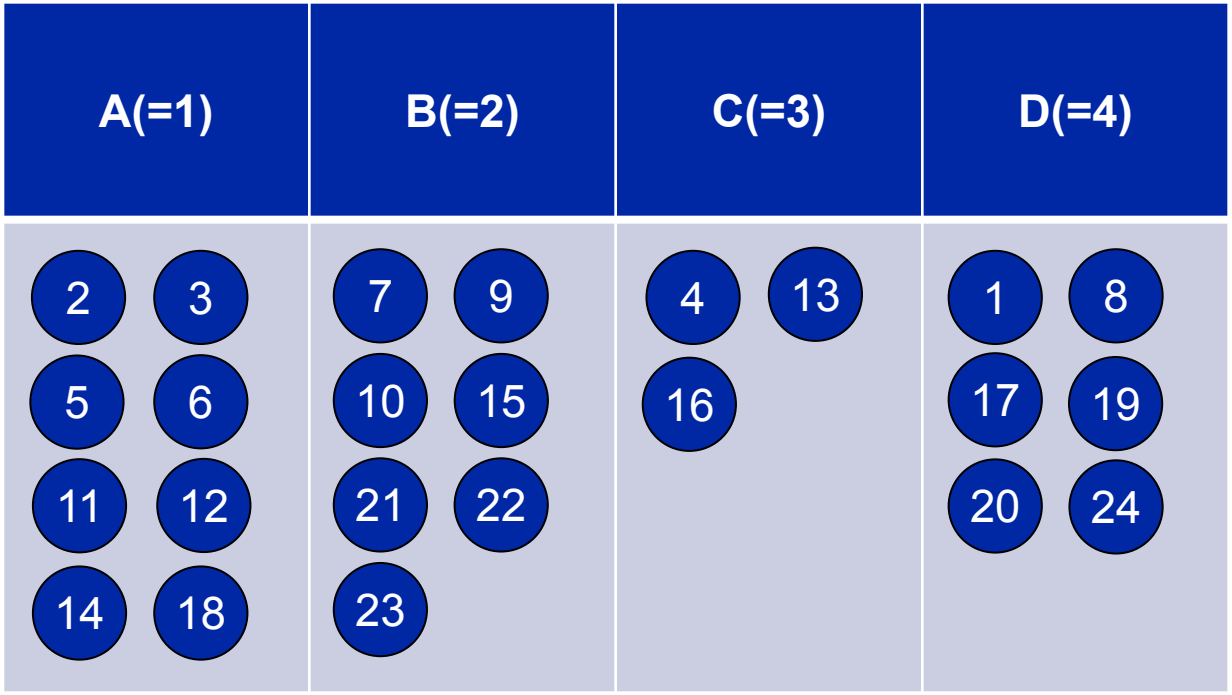
Simple randomization

Take each mouse and randomly assign it to one of the four groups



In R:
For each mouse: `sample.int(4,size=1,replace=F)`

Simple randomization

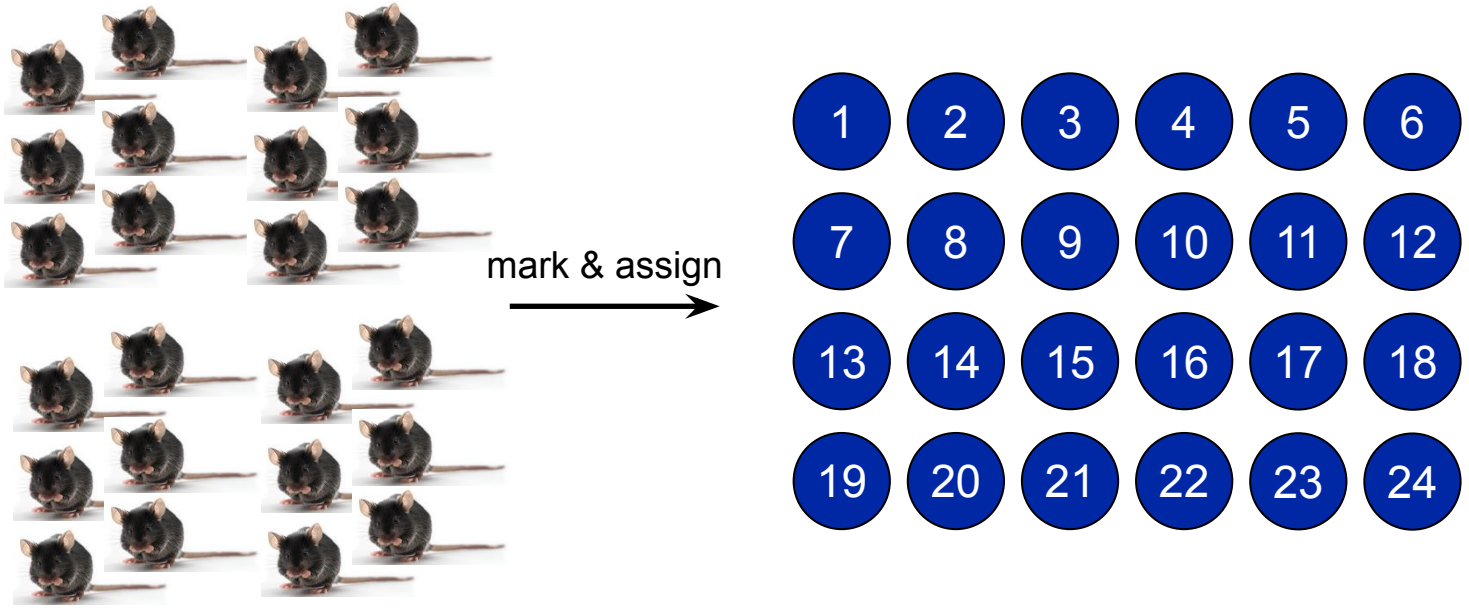


In R:

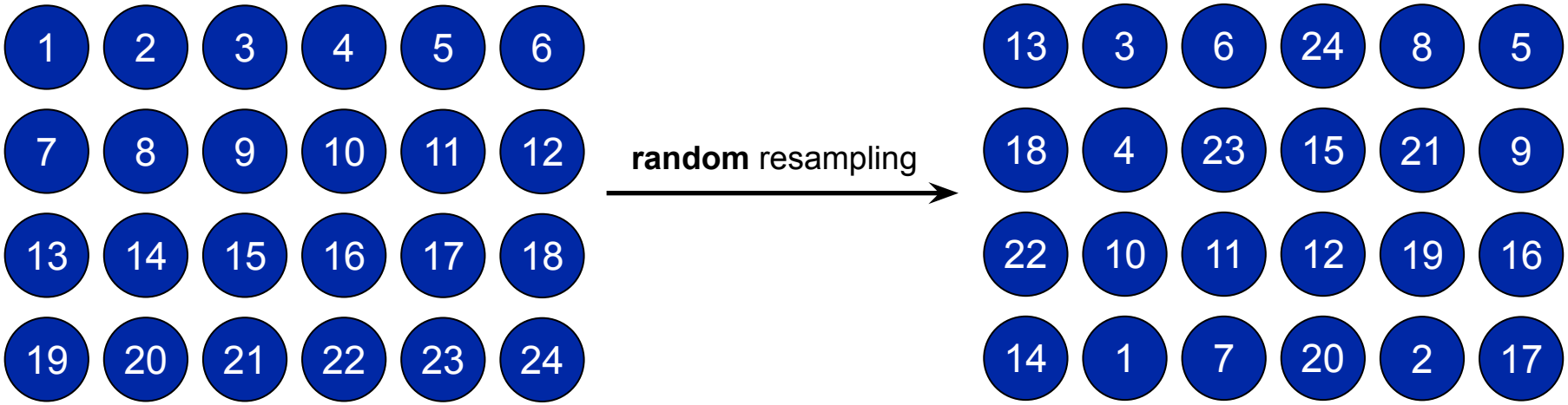
For all mice together: `sapply(1:24,function(x) sample.int(4,size=1,replace=F))`

Balanced randomization

- Remember example 2 (metabolic study) with 3 treatment groups and 1 control group; assume you have 24 mice.
- Goal: Assign each mouse randomly to one of the four groups, but ensure that group sizes are equal



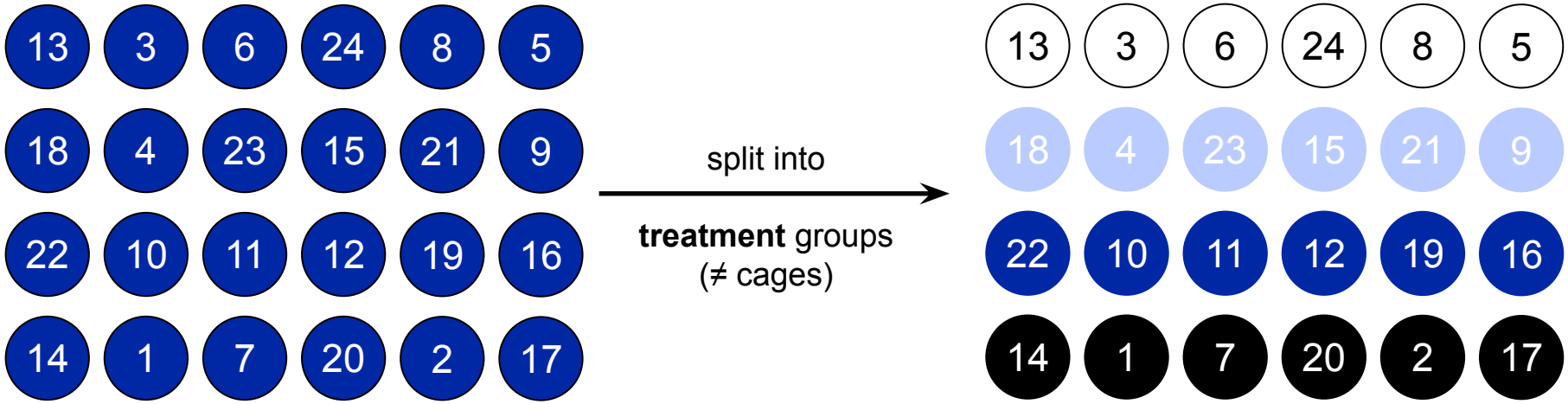
Balanced randomization



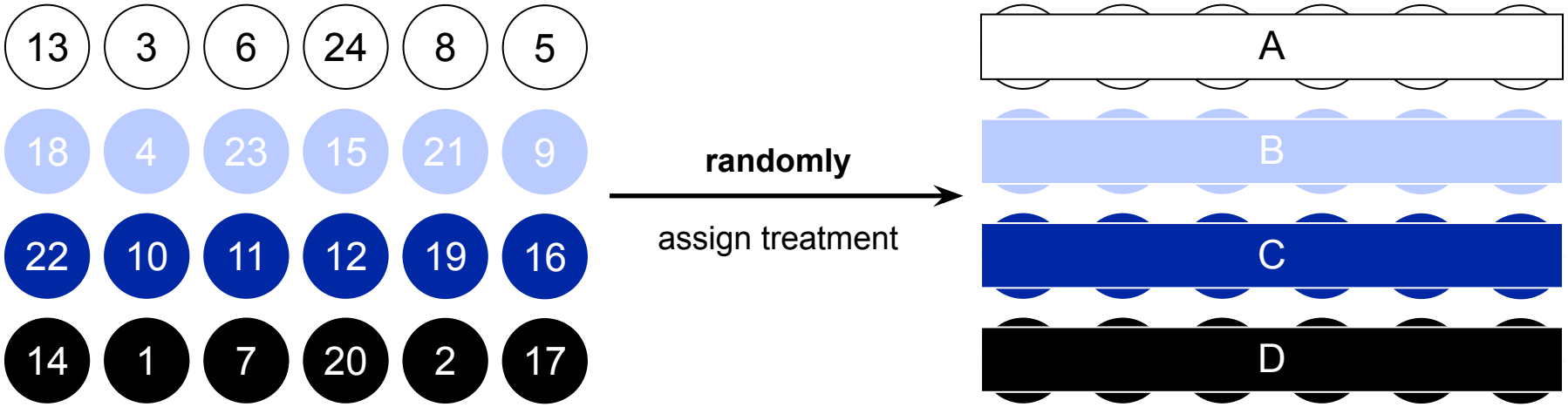
In R:

```
sample.int(n=24,size=24,replace=F)
```

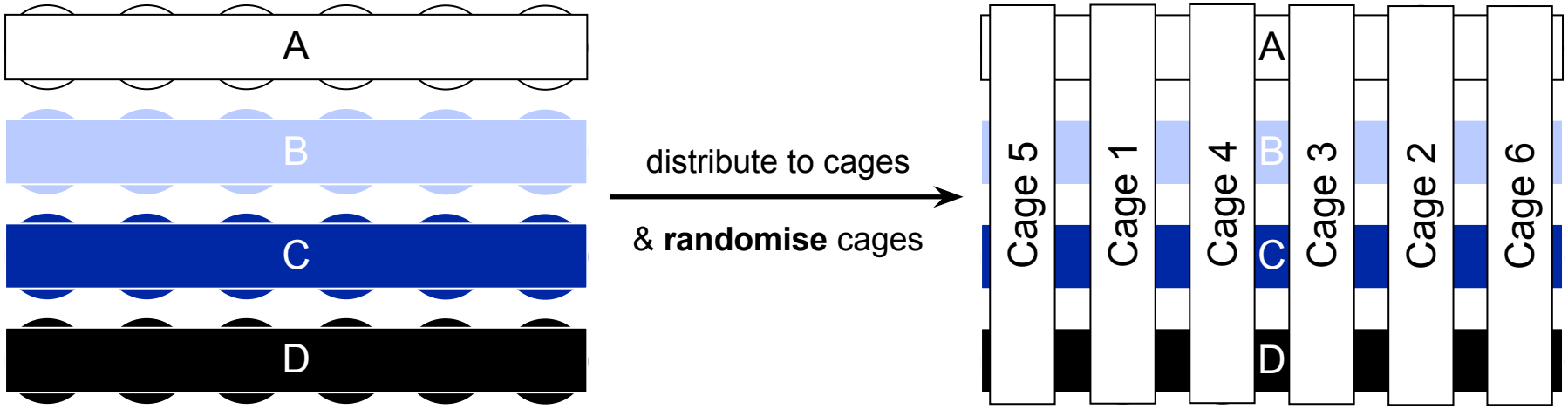
Balanced randomization



Balanced randomization

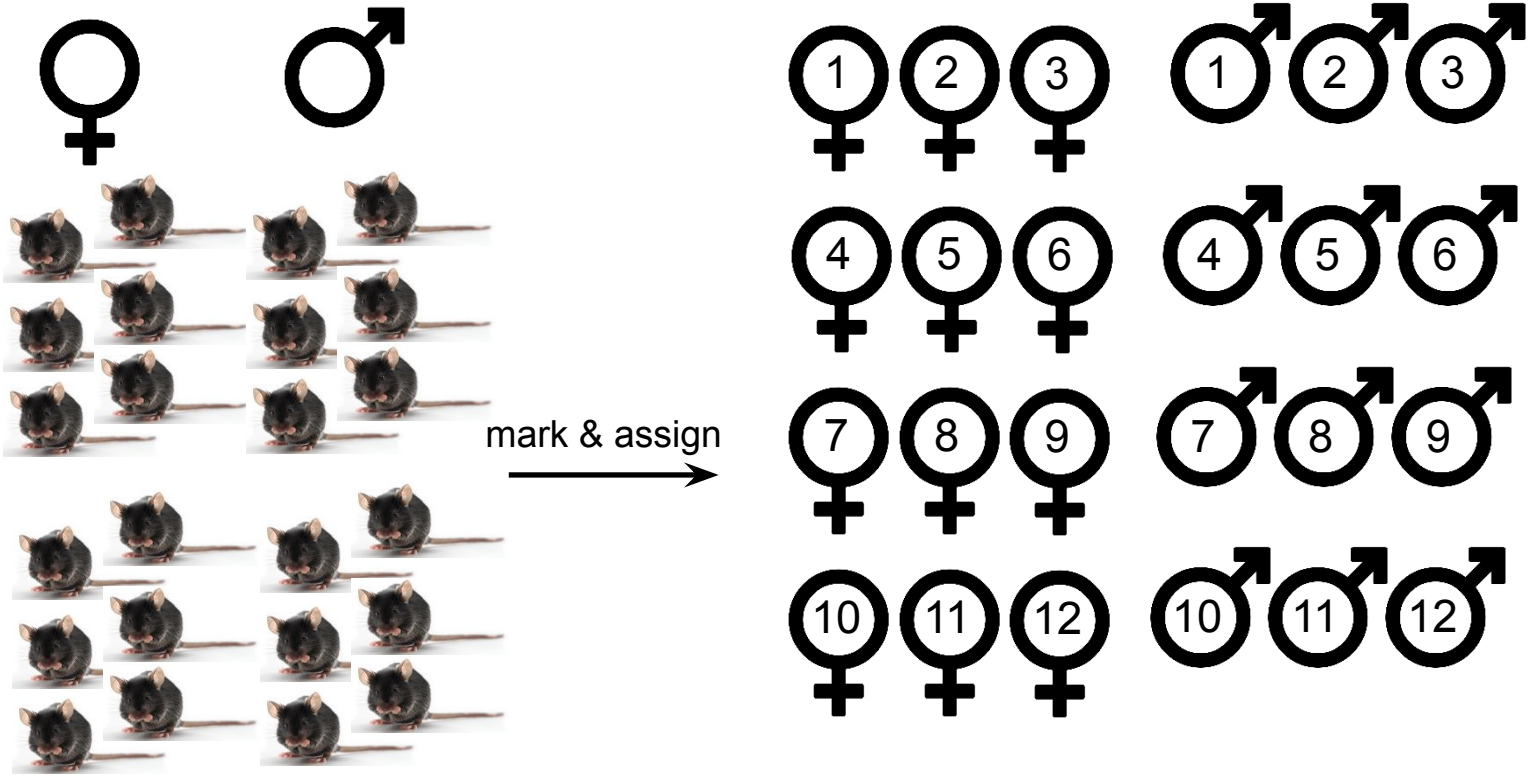


Balanced randomization

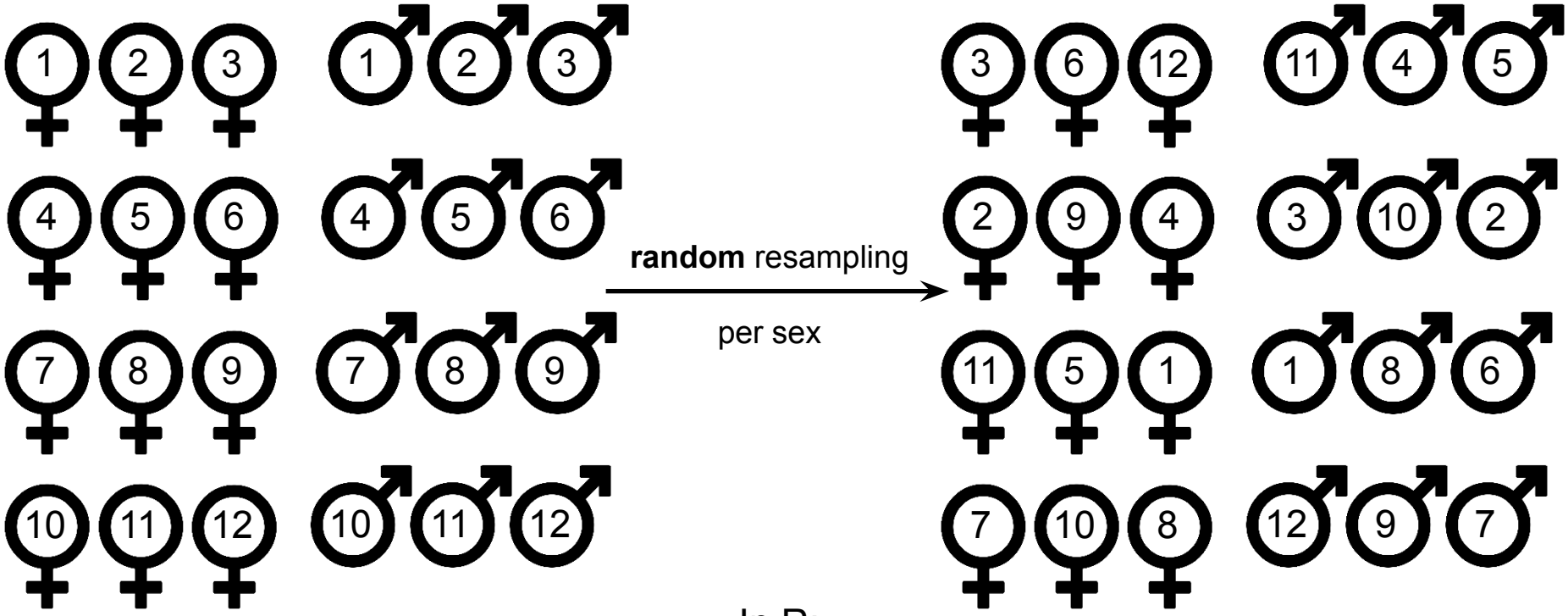


Stratified randomization

- Remember example 2 (metabolic study) with 3 treatment groups and 1 control group; assume you have 24 mice, 12 female and 12 male.
- Goal: Assign each mouse randomly to one of the four groups, but ensure that each group contains equal amount of both sexes (and all groups are equally sized)



Stratified randomization

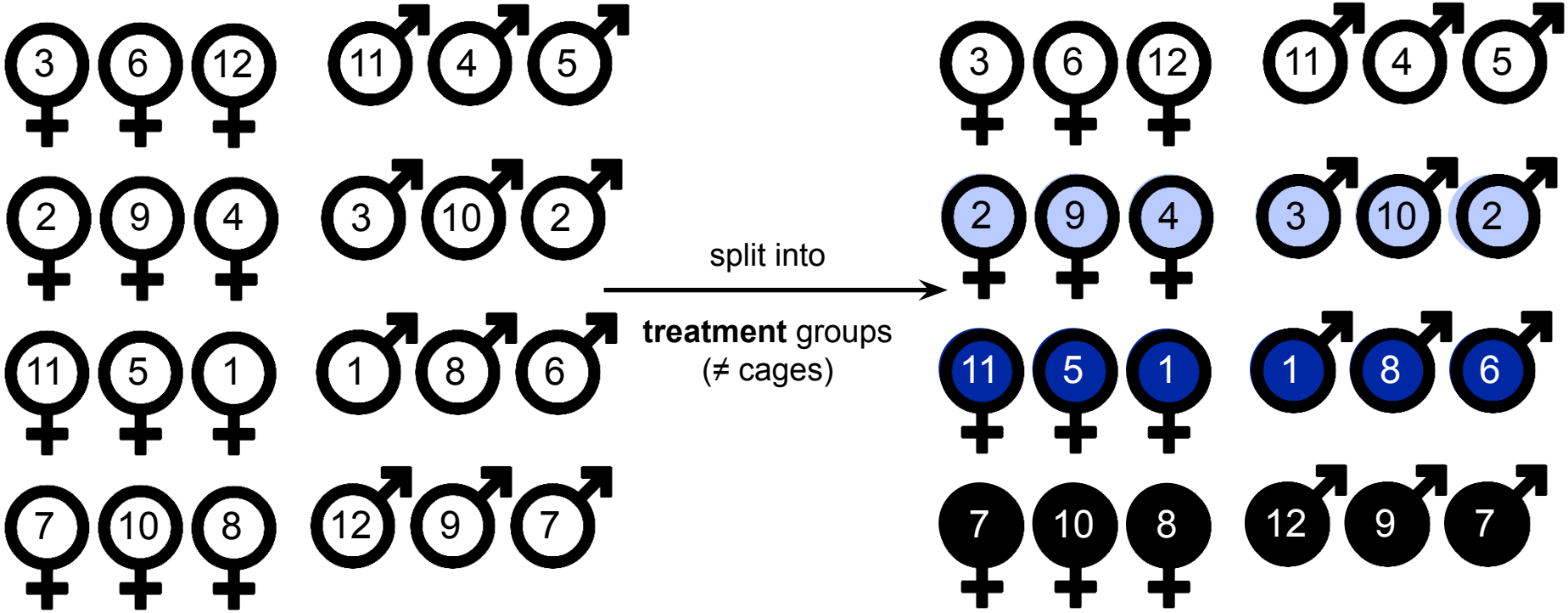


In R:

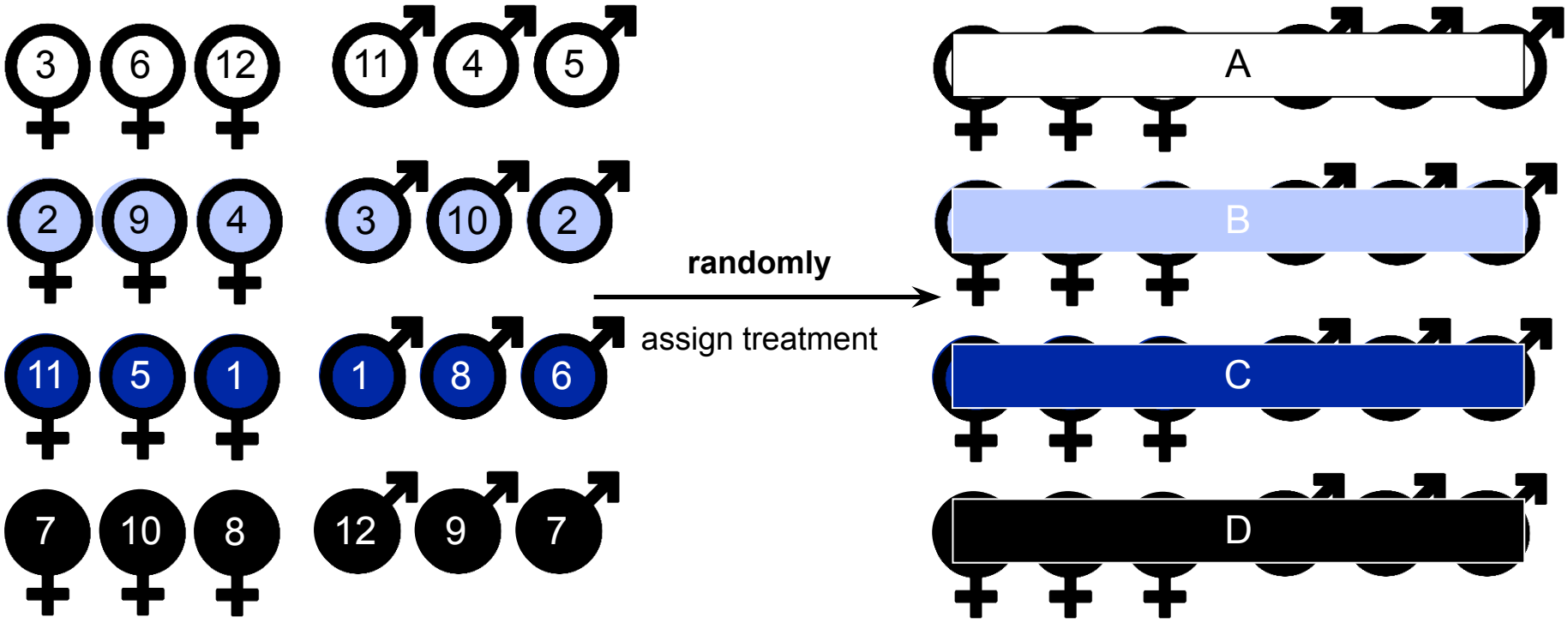
For each sex: `sample.int(n=12,size=12,replace=F)`

Stratified randomization

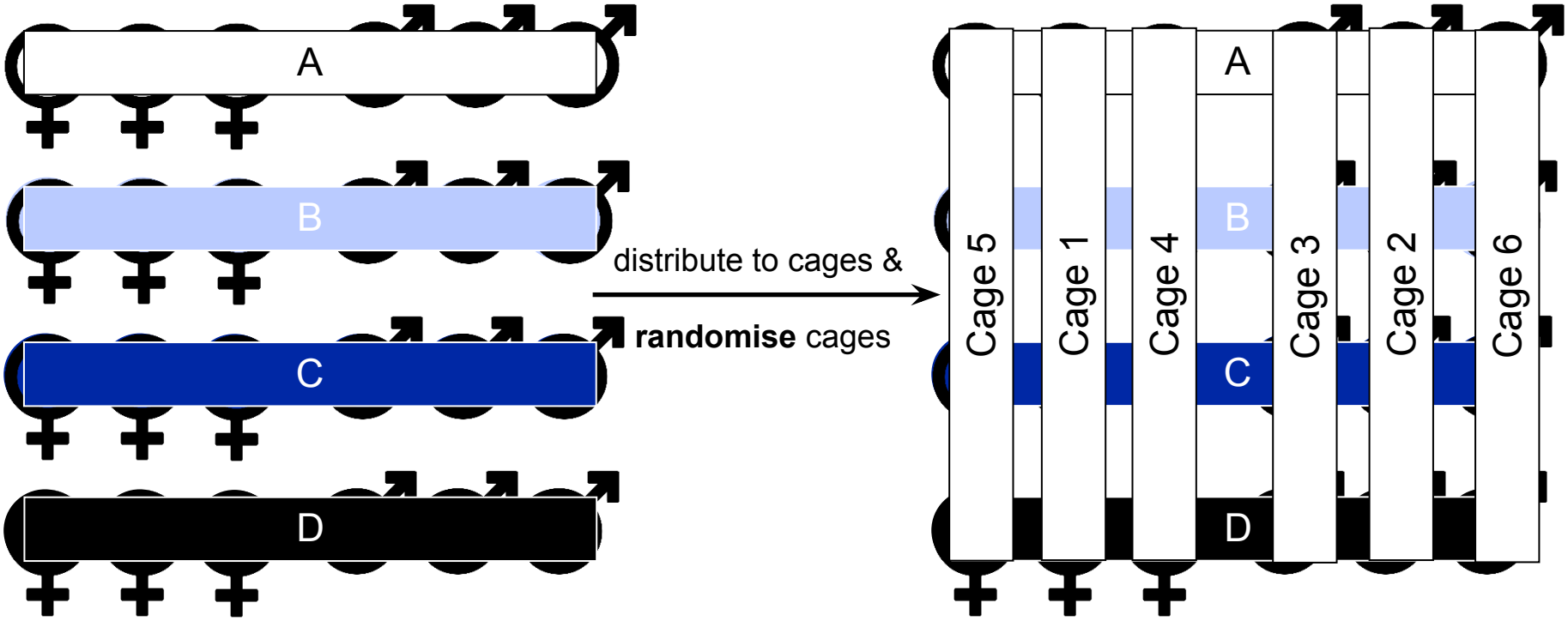
Stratified randomization (aka block randomization) is basically simple balanced randomization applied on the blocks/strata



Stratified randomization



Stratified randomization



Randomization – What if it's not so simple?

- In some cases, your treatment has to be allocated to the all animals of the same cage because co-housing of animals of different groups would lead to bias

- In these cases, you need to randomly distribute animals to cages and then randomly allocate one of the interventions to the cages.

- Be aware that in this case, the experimental unit is the cage, not the individual animal!

- This needs to be taken into consideration when calculating sample size and analysing the experiments. Two examples of how to do this:
  - Calculate sample size & analyse results on per-cage levels
  - Use hierarchical/nested/clustered designs

Design and analyse on a per-cage basis

Calculate sample size and analyse results on per-cage basis:

1. Estimate all relevant statistics (e.g. empirical mean, variance, biologically relevant effect size etc.) on the cage level with a given number of animals per cage.
2. Calculate number **of cages** that are needed to achieve the desired power.
3. When conducting a statistical test, use the mean measurement of each cage as the basis of your test.

- Advantage: Easy and straightforward, standard methods can be used
- Disadvantage: Interpretation of effect size is not straightforward; you lose a lot of information; in most cases, you do not address the biologically relevant question

Use hierarchical/nested/clustered designs

Calculate sample size and analyse results on a per-animal basis:

1. Estimate all relevant statistics (e.g. empirical mean, variance, biologically relevant effect size etc.) on the level of the individual animal.
2. Calculate number of animals that are needed to achieve the desired power.
3. Correct sample size for correlation of animals that are held within the same cage.
4. Analyse the data **including the cluster effect**!

- Advantage: Usually more powerful and insightful.
- Disadvantage: More complex -> Approach statistician for support!

Blinding

Conceal information about treatment groups from people involved in an experiment.

Blinding prevents a lot of systematic errors from biasing your results.

Who should be blinded?

- Experimenters administering treatments
- Experimenters assessing outcome
- Caregiver, animal facilities staff
- Others who interact with the animals

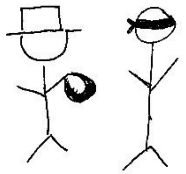
What information should be concealed?

- The experimental group when assessing the outcome
- The future treatment group of subjects during pre-treatment interventions such as inducing a lesion
- Previous values when multiple observations are made on the same experimental units or observational units
- The values of other outcomes from the same experimental units or observational units
- Information that provides clues about treatment groups

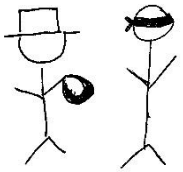
Blinding – It's easy!

1. Plan your experiment in an unblinded manner.

2. Ask a colleague to randomly assign your mice into different treatment groups and to give you a **coded** group allocation so that you do not kno which mice belong to the control group and which belong to the treatment group.

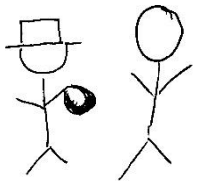


3. Ask a colleague to code all substances used for interventions etc. so that you do not know whether you are applying a treatment or a control substance.

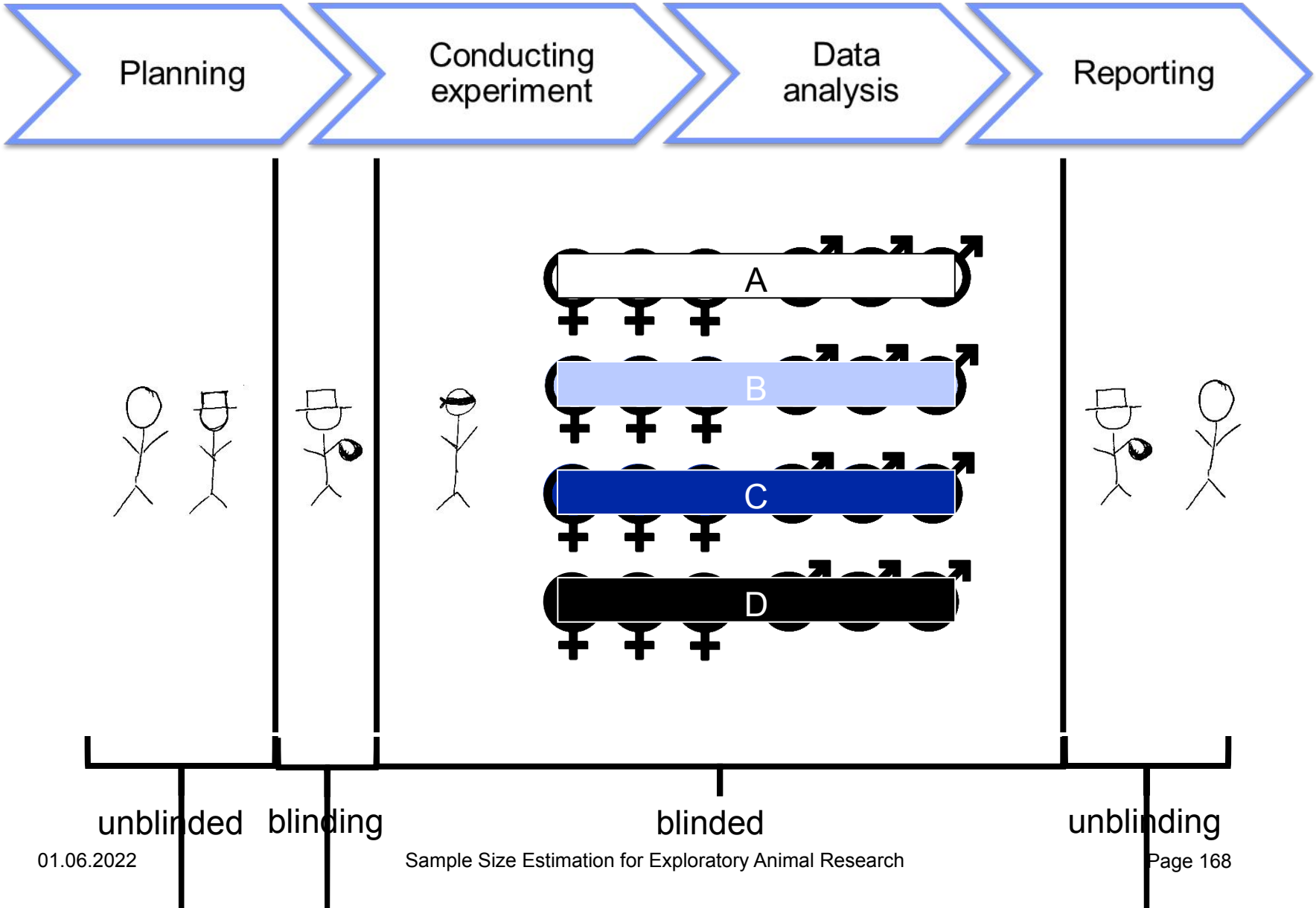


4. Keep blinding until the end of the experiment (including data analysis)!

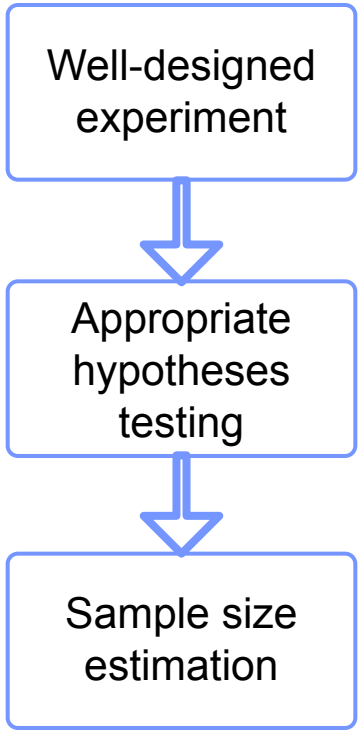
5. Unblind for reporting



Blinding – It's easy! (cont.)



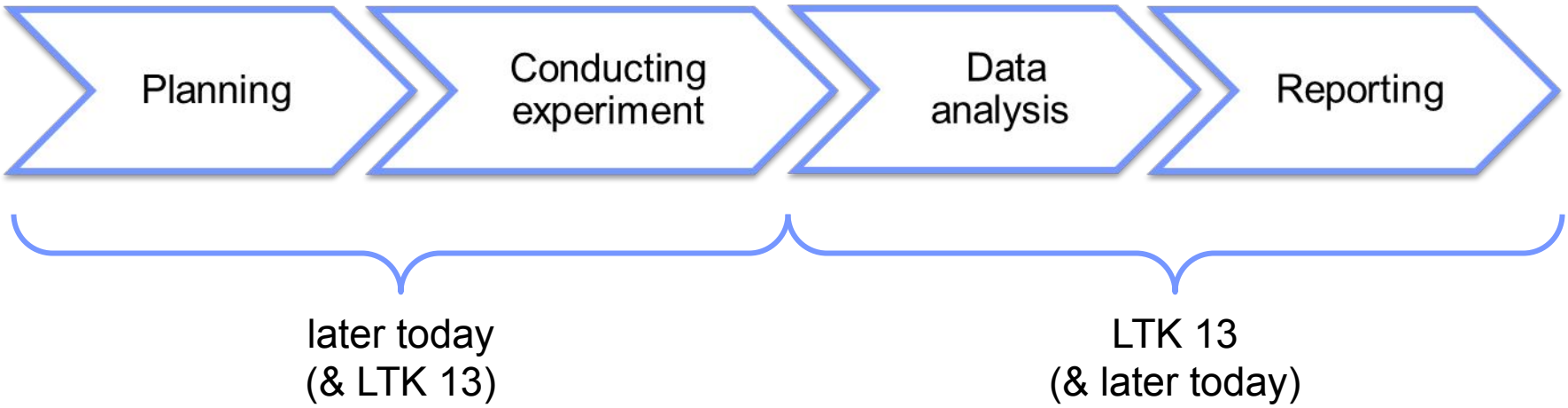
Recap: Planning for success



It provides the <u>strongest evidence</u> in support of <u>causal inference</u>

- Clear and focused hypotheses (PICO-B)
- Identification of source of variability/uncertainty
- Ensure unbiasedness (randomization, blinding)
- Choosing a good primary outcome
  - Objectively measured (biomarker vs subjective evaluation)
  - Measure close to where the action is
  - Unlikely to be missing or censored
  - High reliability
  - Normal distributed with constant variance
- Equal size of each group (balanced-design)
- Correct identification of experimental unit
- Wide range of applicability → blocking: deliberate variation
- Keep it simple but not simpler

Details of the four stages of experiment



Guidance document for animal research applications on
www.servangrueniger.ch/studydesign

Literature - How to plan, execute and analyse an experiment

Lazic, S.E. (2016). Experimental Design for Laboratory Biologists. Cambridge University Press. Chapter 3.

Lazic SE, Clarke-Williams CJ, Munafò MR (2018) What exactly is 'N' in cell culture and animal experiments?. PLOS Biology 16(4): e2005282. https://doi.org/10.1371/journal.pbio.2005282

Bate & Clark (2014), The Design and Statistical Analysis of Animal Experiments, 2014 https://www.cambridge.org/core/books/design-and-statistical-analysis-of-animal-experiments/BDD758F3C49CF5BEB160A9C54ED48706

Van der Worp, H. B., Howells, D. W., Sena, E. S., Porritt, M. J., Rewell, S., O'Collins, V., & Macleod, M. R. (2010). Can animal models of disease reliably inform human studies?. PLoS med, 7(3), e1000245.