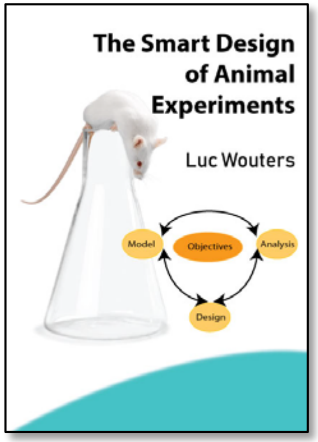
The Smart Design of Animal Experiments

Luc Wouters

September 2018

The Smart Design
of
Animal Experiments


The Smart Design of Animal Experiments

Luc Wouters

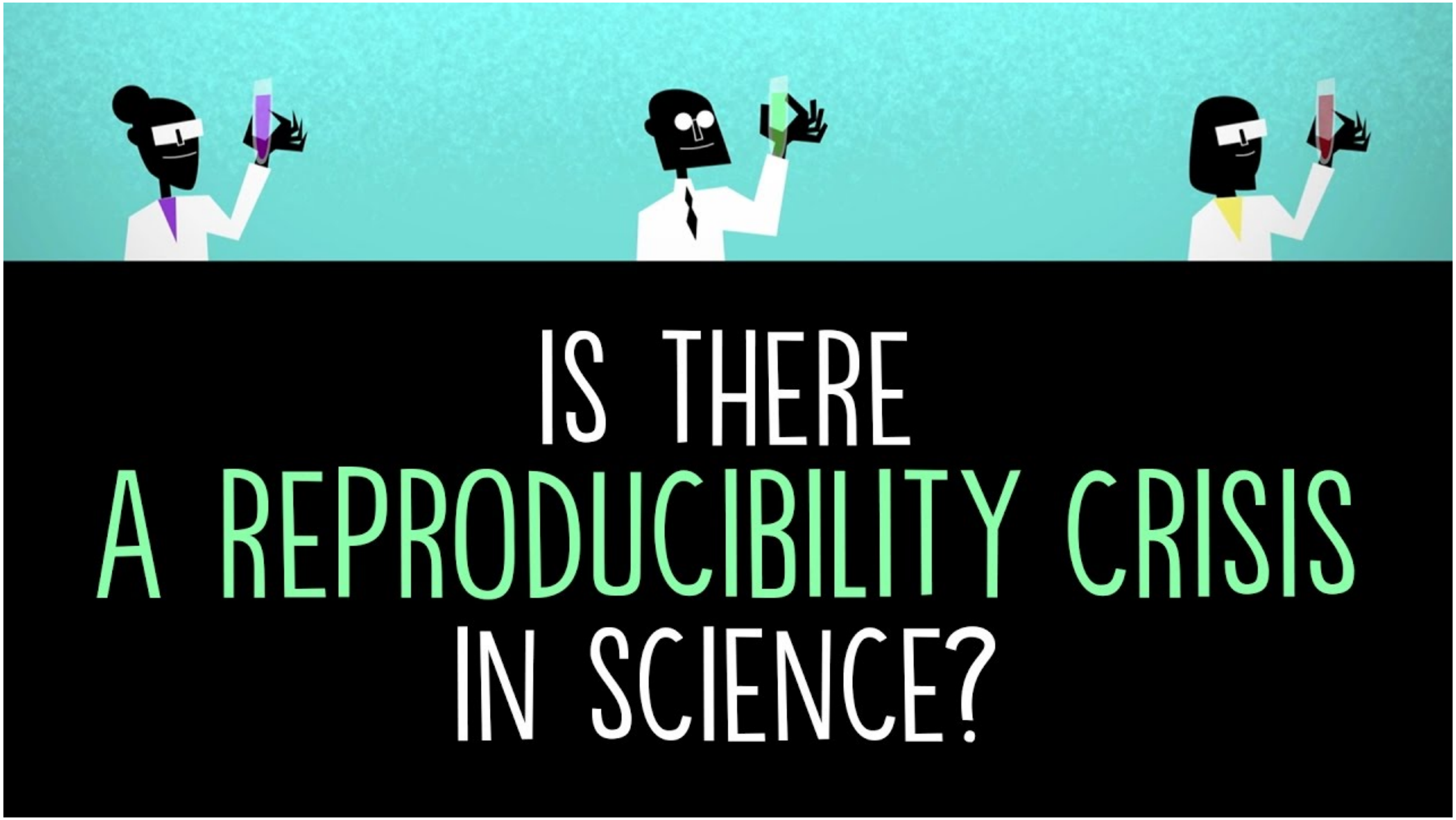
The Smart Design of Animal Experiments

- Day 1:
  - Introduction
  - Smart Research Design by Statistical Thinking
  - Planning the Experiment
  - Principles of Statistical Design
  - Common Designs in Biological Experimentation
- Day 2:
  - Common Designs in Biological Experimentation (cont.)
  - Sample Size and Power
  - The Statistical Analysis
  - The Study Protocol
  - The Research Report
  - Case Studies
  - Concluding Remarks

The Smart Design of Animal Experiments

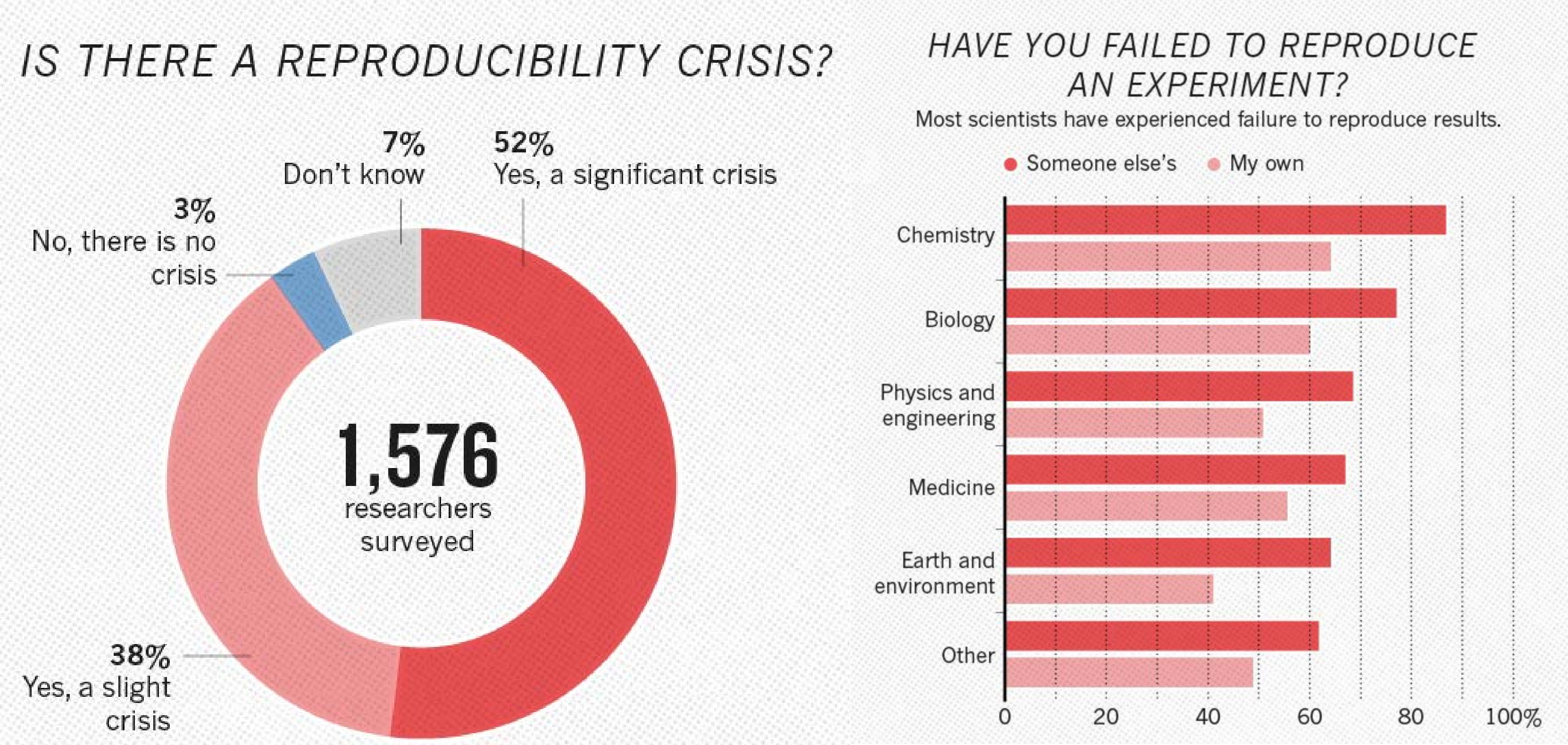
I. Introduction

Reliability of biomedical research
An issue?

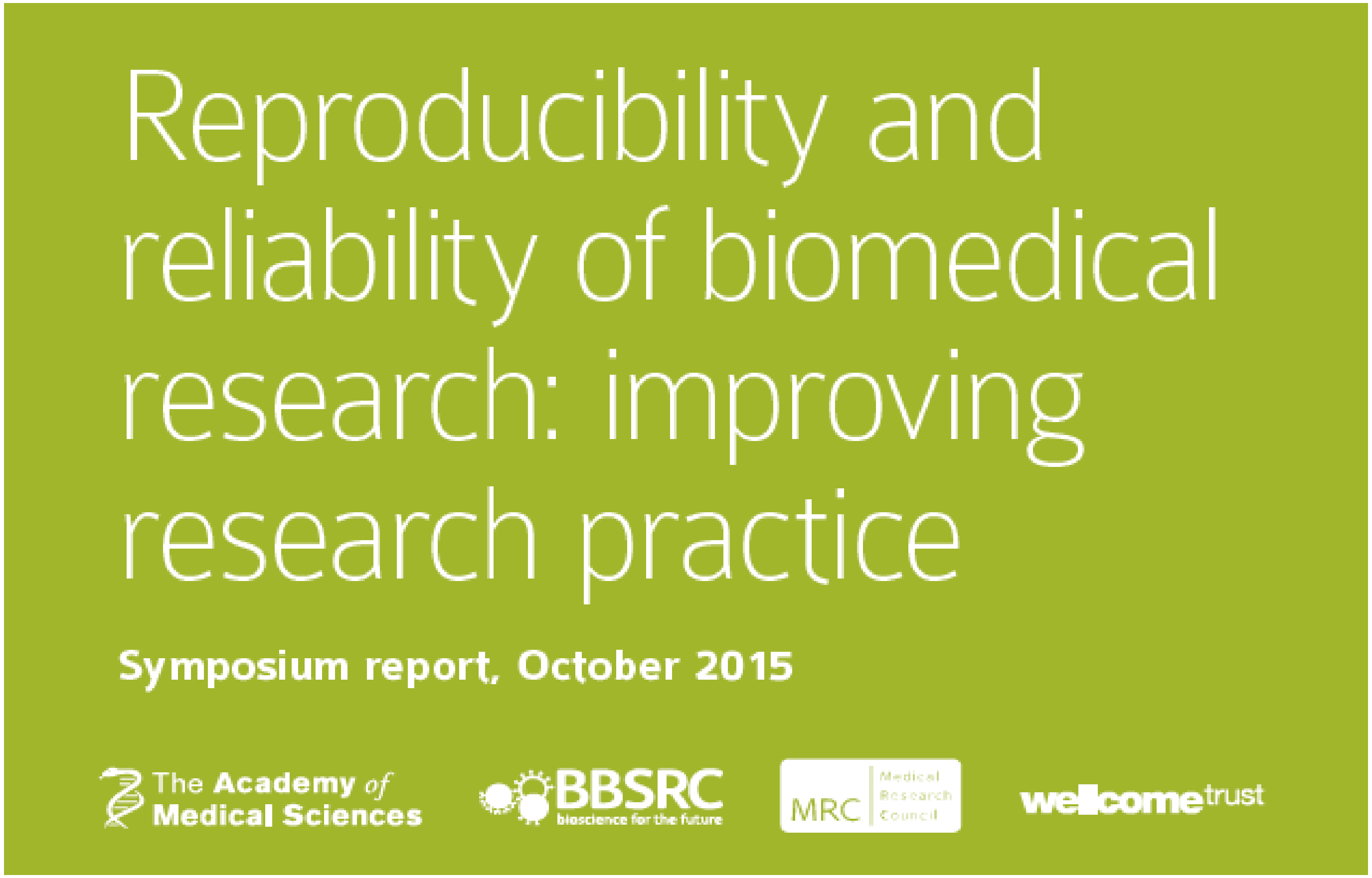


Reliability of biomedical research
A worrisome problem

Baker, M. (2016). Is there a reproducibility crisis? Nature 533, 452-456



Reliability of biomedical research
A worrisome problem



Reproducibility and reliability of biomedical research: improving research practice

Symposium report, October 2015

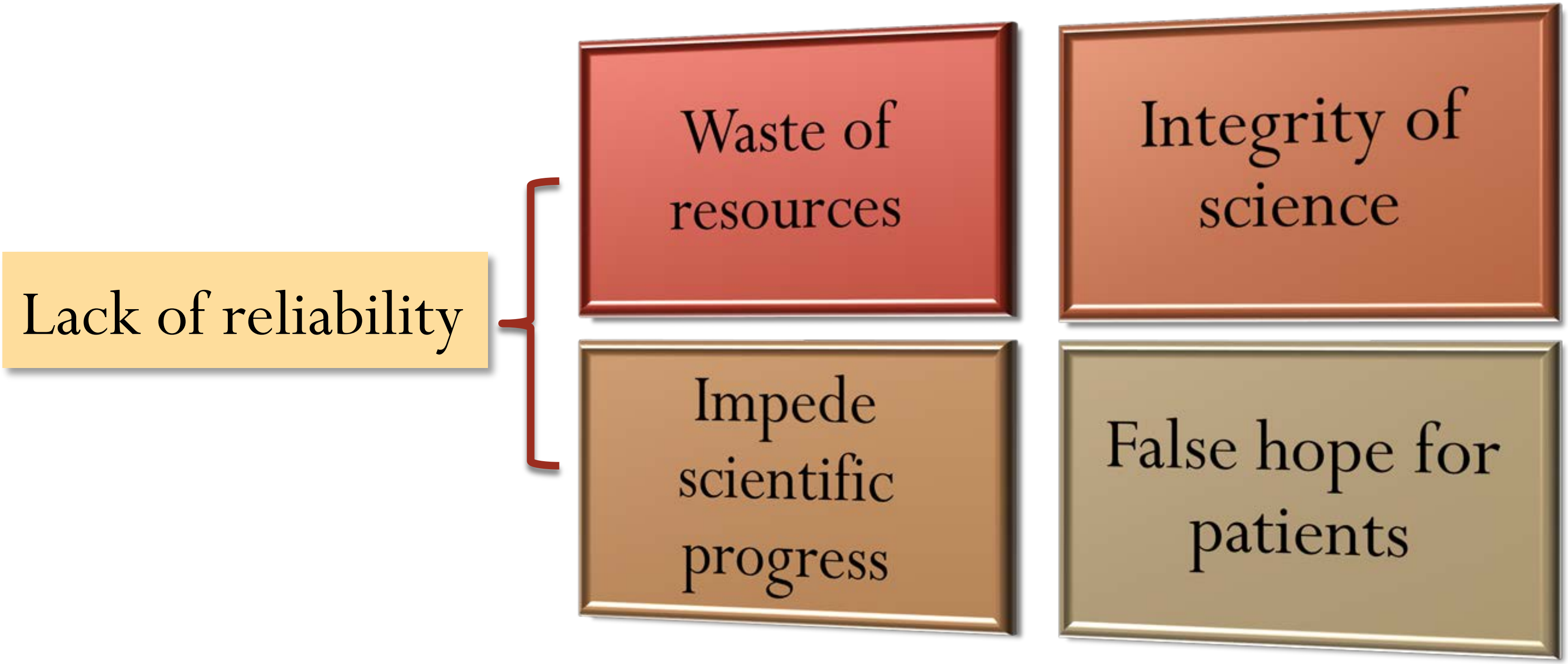
The Academy *of* Medical Sciences

BBSRC
bioscience for the future

MRC Medical Research Council

wellcometrust

Reliability of biomedical research
A worrisome problem



Lack of reliability

Waste of resources

Integrity of science

Impede scientific progress

False hope for patients

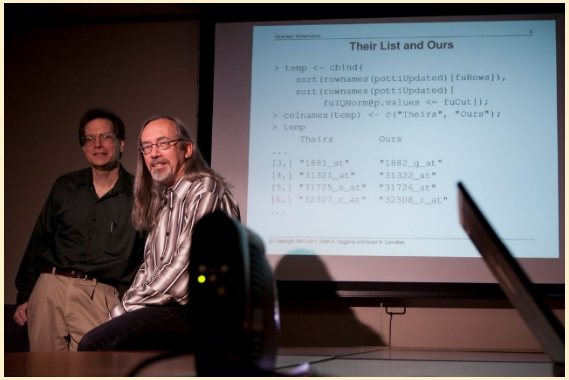
Issues in biomedical research
Reproducibility & Replicability

- **Reproducibility**
  Starting from the existing original data can we reproduce the same results, p-values, confidence intervals, tables and figures as reported.

- **Replicability**
  The replication of scientific findings using independent investigators, methods, data, equipment and protocol, i.e. can we repeat the entire experiment and obtain the same conclusions as reported

Issues in biomedical research
Reproducibility

- **Potti et al. (Nature Medicine 2006): Genomic signatures to guide the use of chemotherapeutics**
- Algorithm to predict which cancer patients will respond to chemotherapy
- The Economist:
  Choose your poison – *A new test picks the chemotherapy most suited to the tumour*
- Baggerly & Coombes unable to reproduce results:
  - ✓ poorly conducted data analysis
  - ✓ data mislabelled
  - ✓ samples duplicated
  - ✓ ambiguous coding



Their List and Ours

```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
         fuTQNorm@p.values <= fuCut]);
> colnames(temp) <- c("Theirs", "Ours");
> temp
      Theirs          Ours
...
[3,] "1881_at"      "1882_g_at"
[4,] "31321_at"     "31322_at"
[5,] "31725_s_at"   "31726_at"
[6,] "32307_r_at"   "32308_r_at"
...
```

© Copyright 2007-2011, Keith A. Baggerly and Kevin R. Coombes

- Potti et al. (Nature Medicine 2011): Retraction
  “*because we have been unable to reproduce certain experiments*”

Issues in biomedical research
Replicability

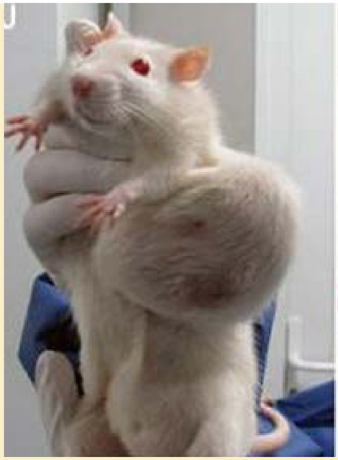
- **Scholl, et al. (*Cell*, 2009):**
  Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells.
- **Conclusion:** cancer tumours could be destroyed by targeting STK33 protein
- Amgen Inc.:
  24 researchers, 6 months labwork
- **Unable to replicate results** of Scholl, et al.

Issues in biomedical research
Replicability (cont.)

- Begley & Ellis (Amgen) identified 53 “landmark” studies in preclinical cancer research
- Replicate results in close collaboration with authors
- 47/53 studies findings could not be replicated
- Consistent with results of Prinz et al. only 25% of findings in target discovery could be validated

Issues in biomedical research
Acceptance

Issues in biomedical research Acceptance



- Séralini, et al. (*Food Chem Toxicol*, 2012):
  "Long term toxicity of a roundup herbicide and a roundup-tolerant genetically modified maize"
- **Conclusion:**
  GM maize and low concentrations of Roundup herbicide causes toxic effects (tumors) in rats
- Press conference
  e.g. Natural News: *Shock findings in new GMO study: Rats fed lifetime of GM corn grow horrifying tumors, 70% of females die early*
- Severe impact on general public and on interest of industry
  - ✓ Referendum labeling of GM food in California,
  - ✓ Bans on importation of GMOs in Russia and Kenya

Issues in biomedical research
Acceptance

The critiques:

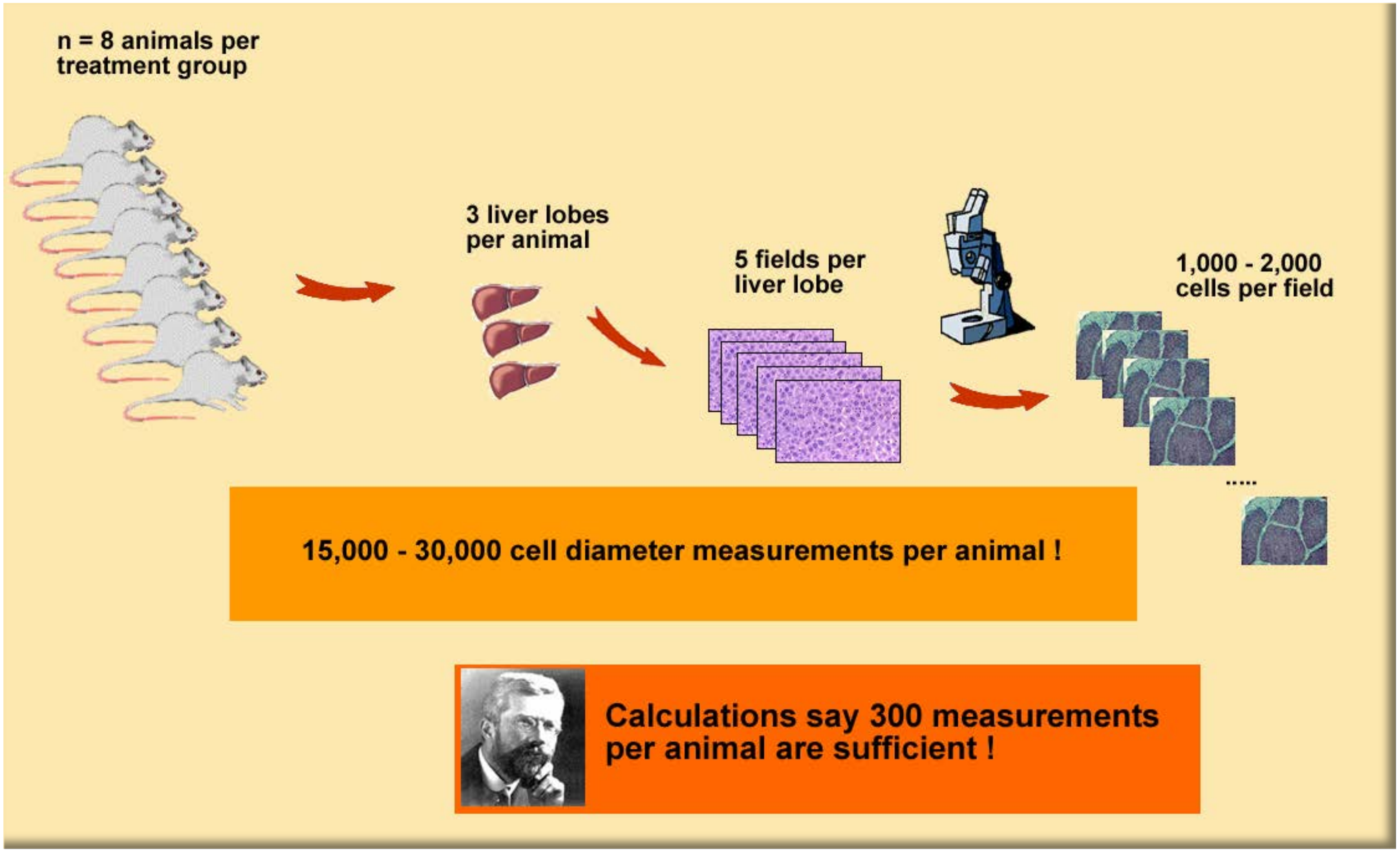
- Heavily critized

- European Food Safety Authority (2012) review:
  - ✓ inadequate design, analysis and reporting
  - ✓ wrong strain of rats
  - ✓ number of animals too small

- Paper withdrawn late 2013

- Republished 2014 in journal with less impact (Environmental Sciences Europe)

Issues in biomedical research
Efficiency



n = 8 animals per treatment group

3 liver lobes per animal

5 fields per liver lobe

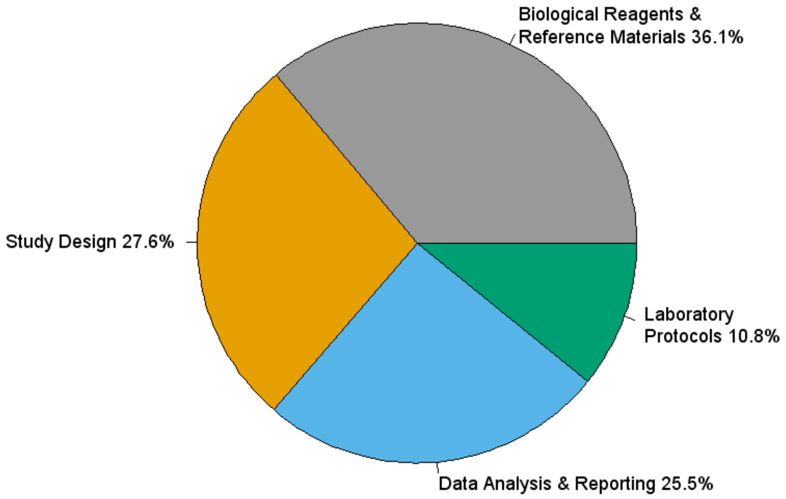
1,000 - 2,000 cells per field

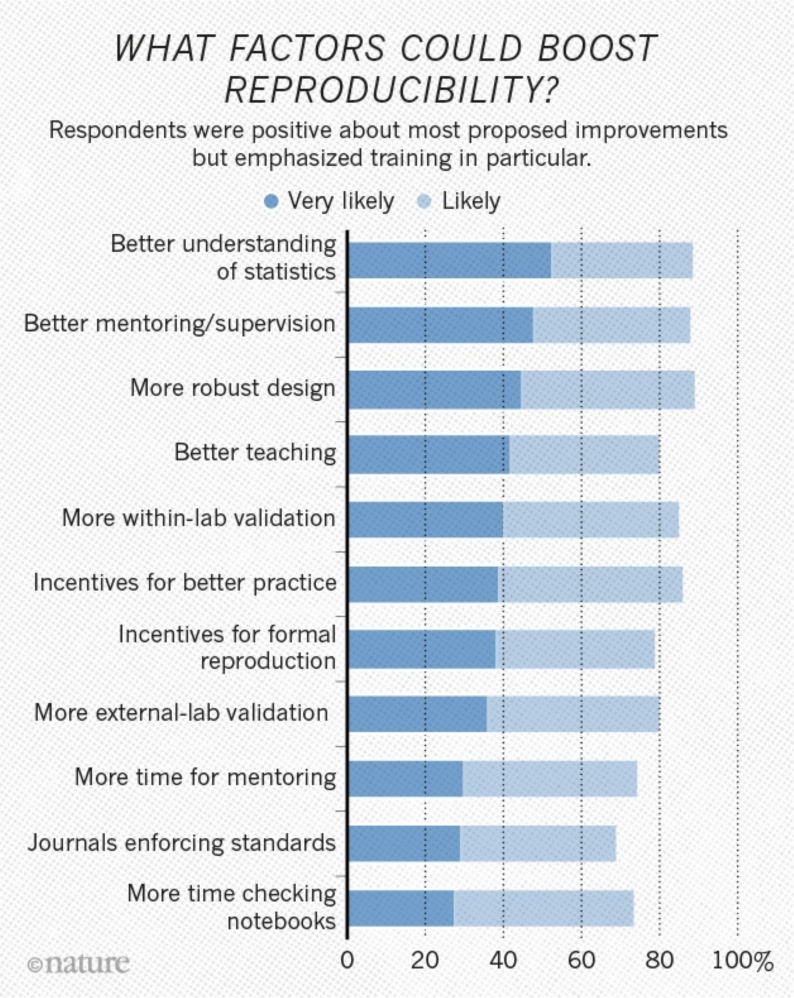
15,000 - 30,000 cell diameter measurements per animal !

Calculations say 300 measurements per animal are sufficient !

Issues in biomedical research
Some figures

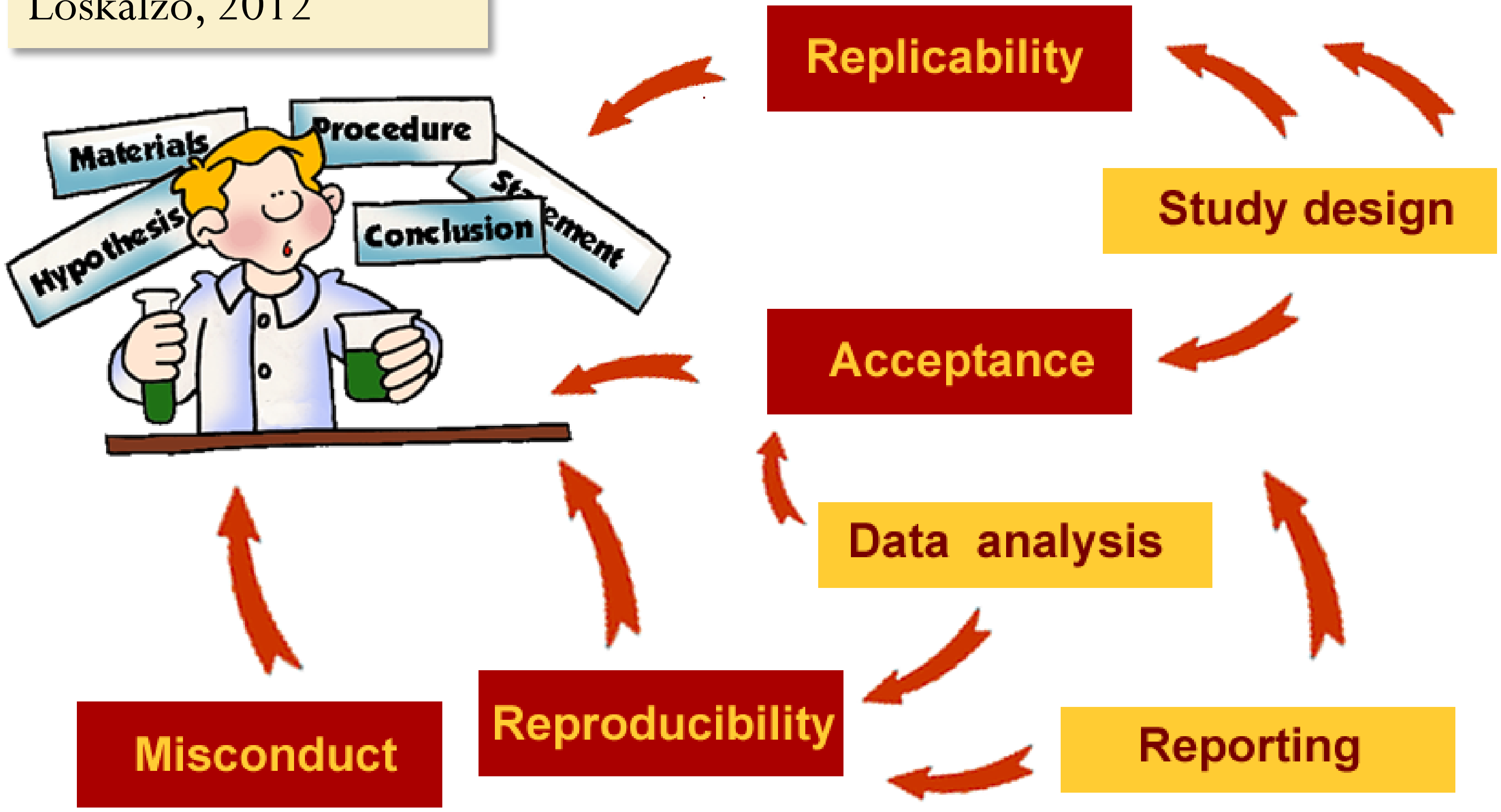
- **85%** of biomedical research is **wasted** (Begley & Ioannidis, 2015)
- US alone **US $28 B/year** spent on research that is not replicable (Freedman, et al. 2015)
- Mostly study design, data analysis & reporting





The problem of doing "good science"
Some critical reviews

Kilkenny, 2009
Loskalzo, 2012



Replicability

Study design

Acceptance

Data analysis

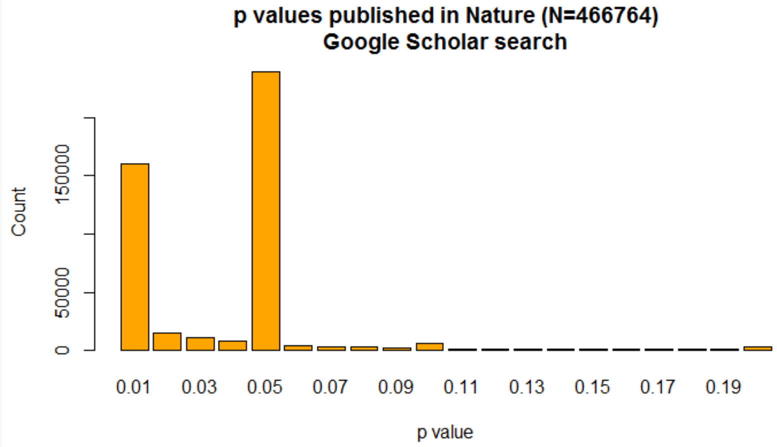
Misconduct

Reproducibility

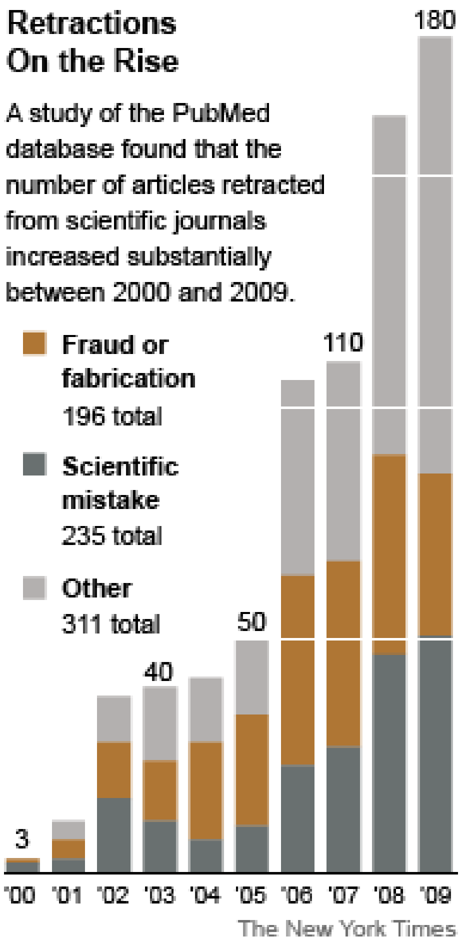
Reporting

The responsibility of scientific journals

- Peer reviewers & editors:
  - ✓ little or no statistical background
  - ✓ do not detect methodological errors
- Publication bias
  - ✓ Focus on statistically significant results of unexpected findings
  - ✓ Not always looking at practical importance
  - ✓ Small studies put replicability in danger (Ioannidis, 2005)
  - ✓ Truth inflation (Reinhart, 2015)



Things get even worse
Retraction rates



Retractions On the Rise

A study of the PubMed database found that the number of articles retracted from scientific journals increased substantially between 2000 and 2009.

The New York Times

- 21% of retractions were due to error
- 67% misconduct, including fraud or suspected fraud, but also Hanlon’s razor

The integrity of our profession is put into question

The New York Times

THE WALL STREET JOURNAL.

Scientists' Elusive Goal: Reproducing Study Results

How Bright Promise in Cancer Testing Fell Apart

Most Science Studies Appear to Be Tainted By Sloppy Analysis

The integrity of our profession is put into question

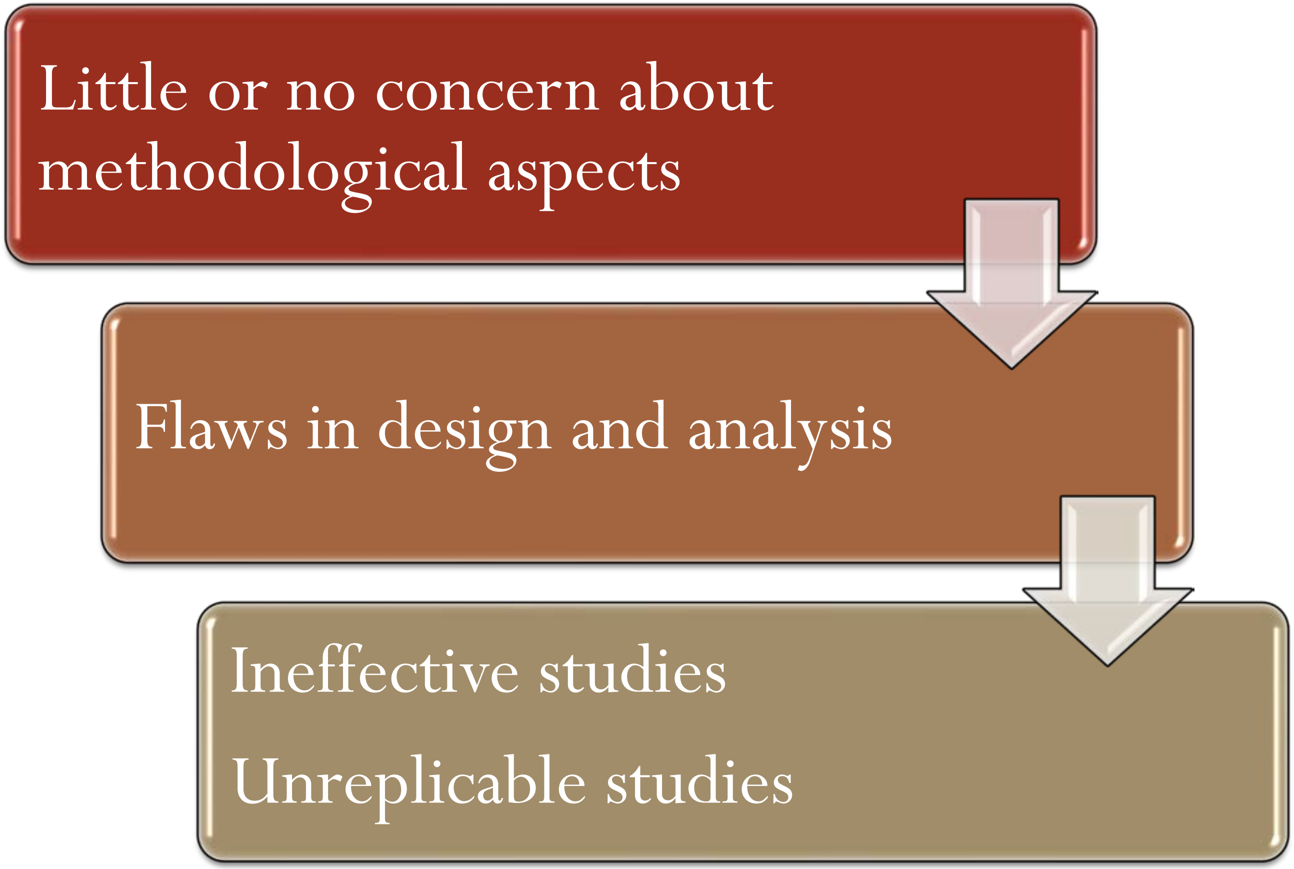
"The way we do our research (with our animals) is stone age"

Ulrich Dirnagl,

Charité University Medicine, Berlin

Science 2013

Summary of problems



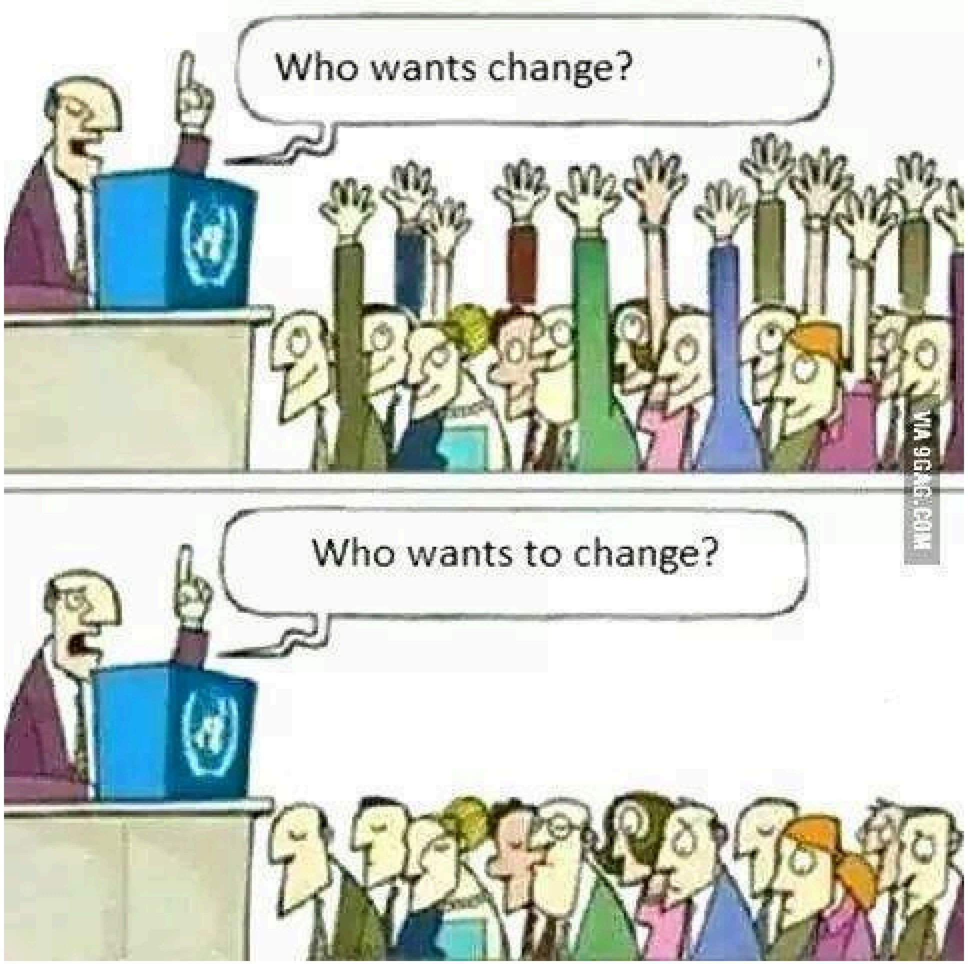
Little or no concern about methodological aspects

Flaws in design and analysis

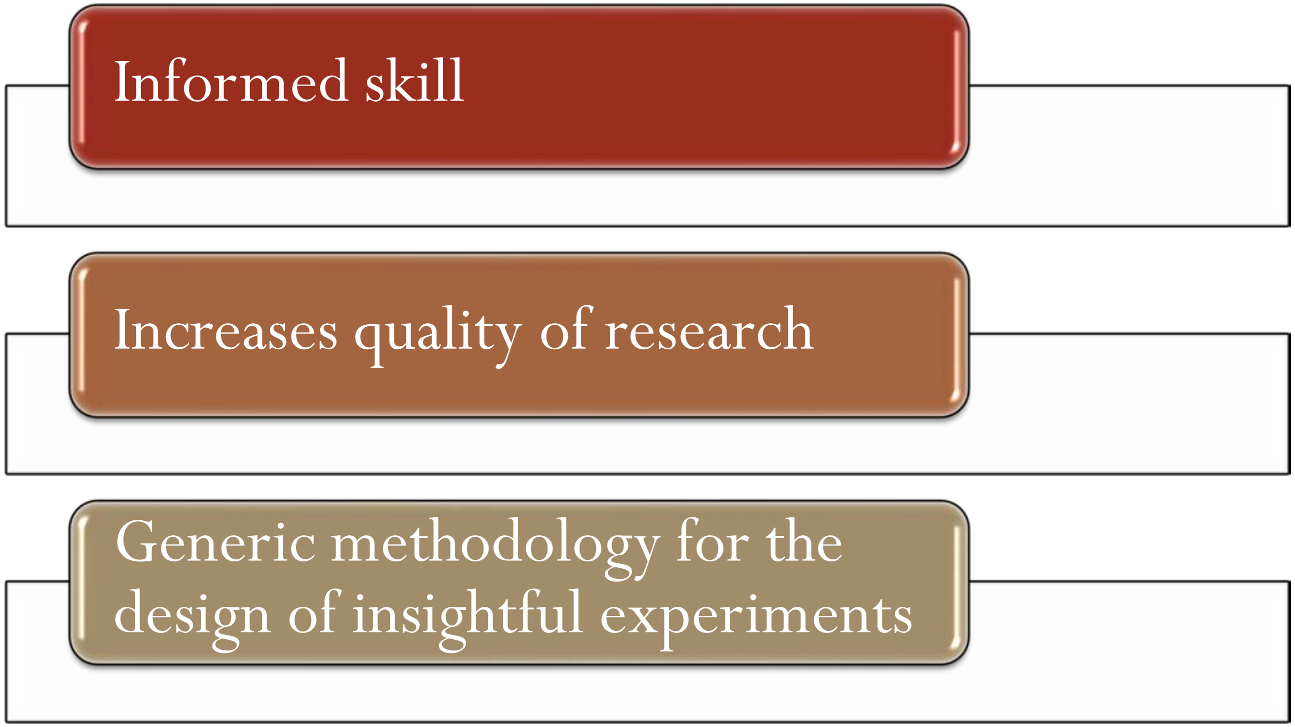
Ineffective studies

Unreplicable studies

Time to transform and improve the research process



Course objective
Statistical Thinking

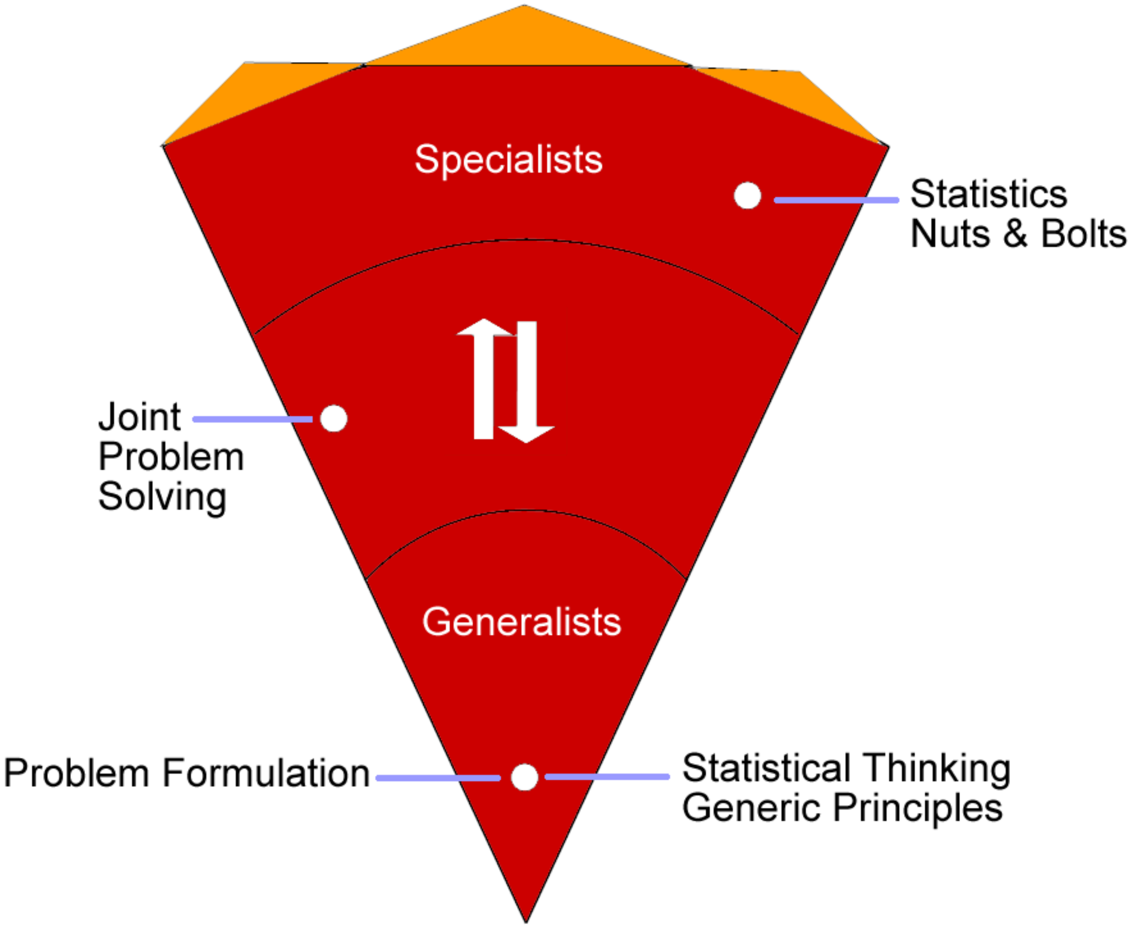


Informed skill

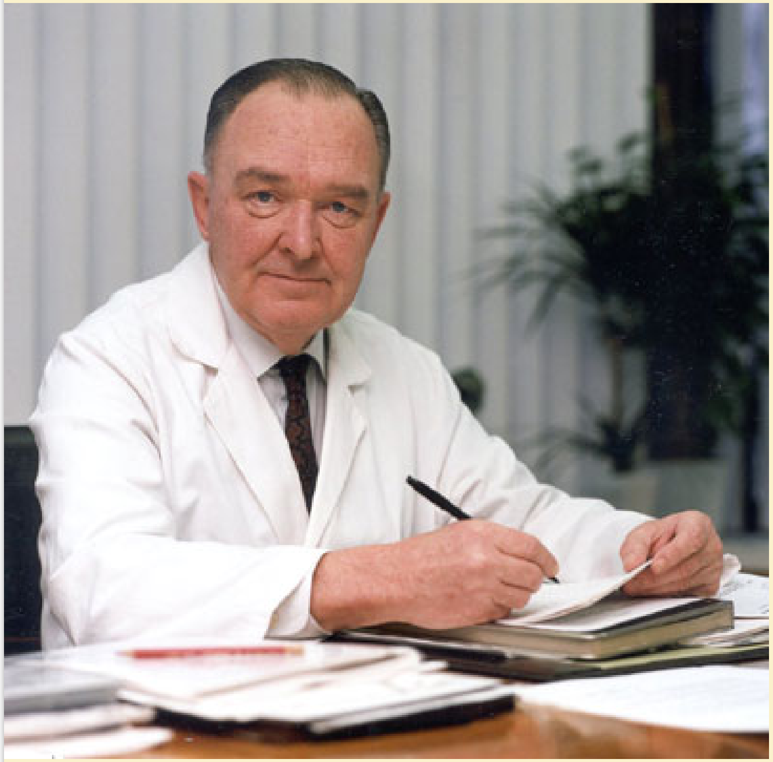
Increases quality of research

Generic methodology for the design of insightful experiments

This course prepares you for a *dialogue*



Statistical thinking permeates the process and leads to highly productive research



- Set of statistical precepts (rules) accepted by his scientists
- Kept research to proceed in orderly and planned fashion
- Open mind for the unexpected
- World record: 77 approved medicines over a period of 40 years
- 200 research papers/year

Software for Designing Experiments

- Software for data analysis SPSS, SAS, JMP, not always suited for designing experiments
- Specialized software expensive
- Occasional use of R
- Freely available, no cost
- Many add-ons (packages) for experimental design, sample size calculations, etc.
- See Appendix B for how to obtain and install R

About R

- Prompt sign is normally “>“
- Assignment by combination of “<“ and “-”
  e.g. x<- 1
- Help is always available ?pwr or help(pwr)
- Cap-sensitive: FrF2 is not the same as fRf2

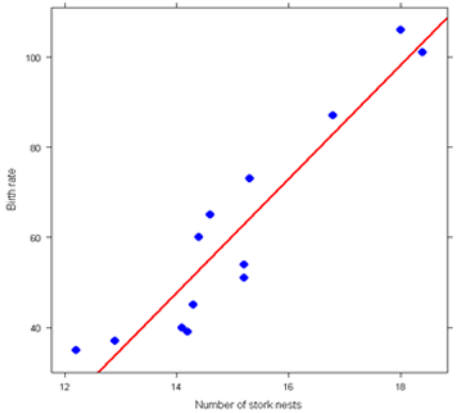
The Smart Design of Animal Experiments

II. Smart Research Design by Statistical Thinking

The architecture of experimental research
The two types of studies

Observational

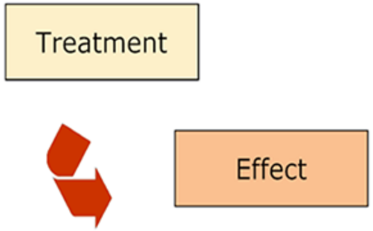
Variation by nature
Confounding (biased)
Birth rate ~ Stork Nests



Confounder = Year

Controlled Experiment

Variation by investigator
Causality



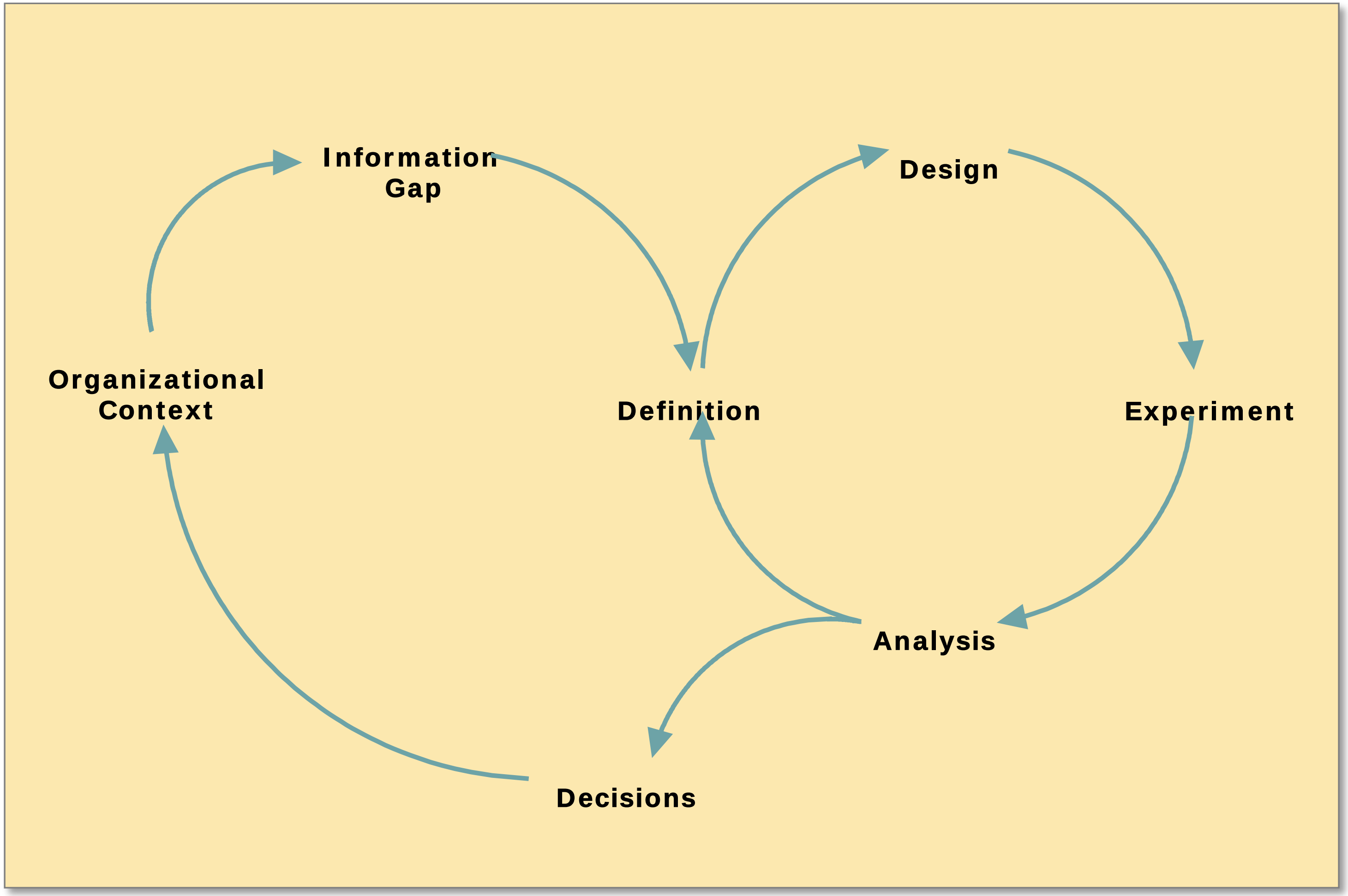
Treatment

Effect

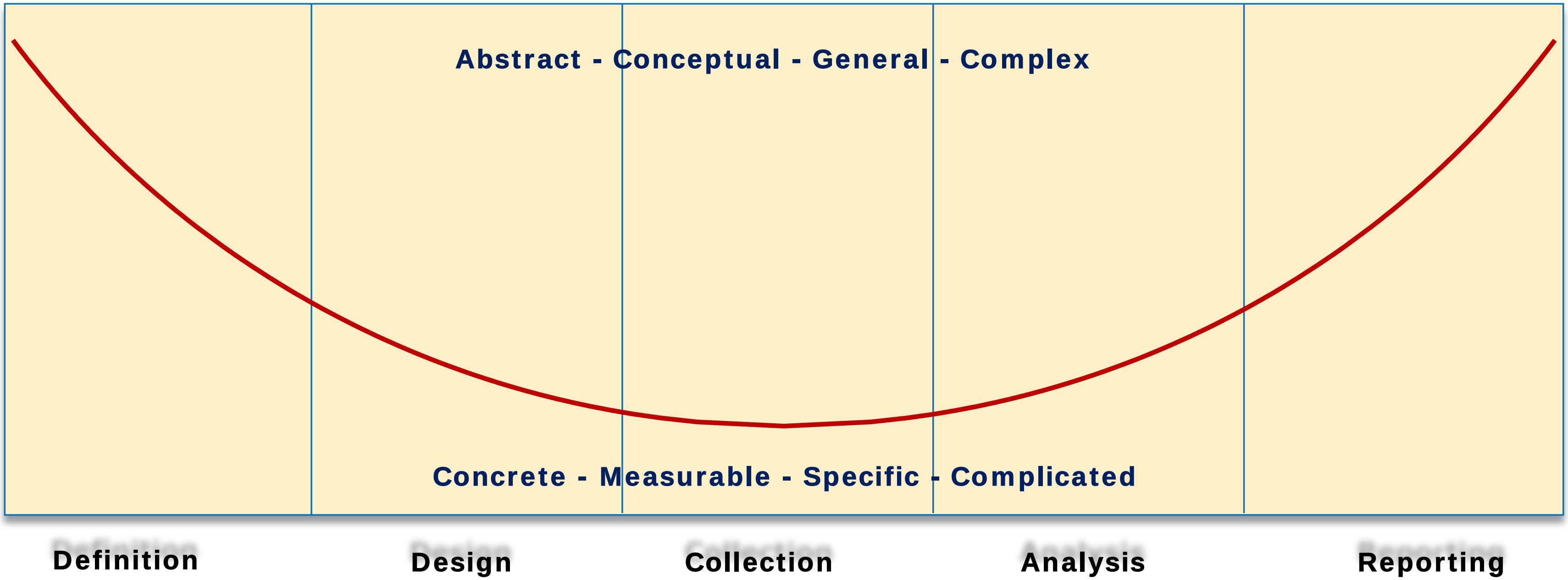
The architecture of experimental research Research is a *phased* process

Phase		*Deliverable*
1. Definition	→	proposal
2. Design	→	protocol
3. Data Collection	→	data set
4. Analysis	→	conclusions
5. Reporting	→	report

The architecture of experimental research Research is an *iterative process*



Modulating between the concrete and the abstract



Abstract - Conceptual - General - Complex

Concrete - Measurable - Specific - Complicated

Definition

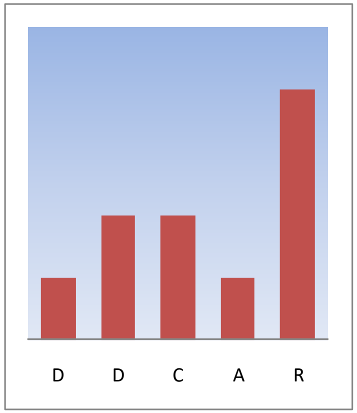
Design

Collection

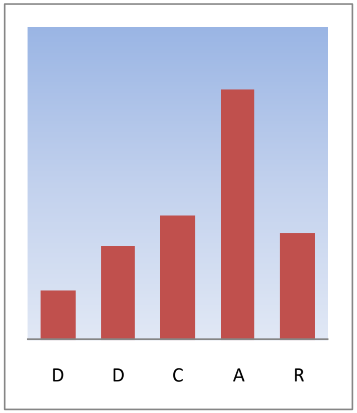
Analysis

Reporting

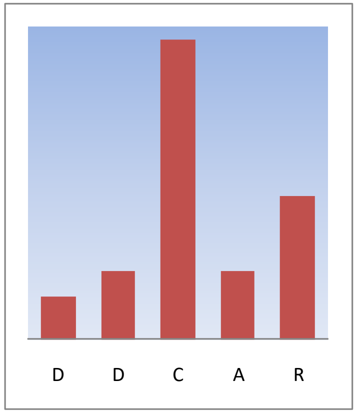
Research styles may vary…



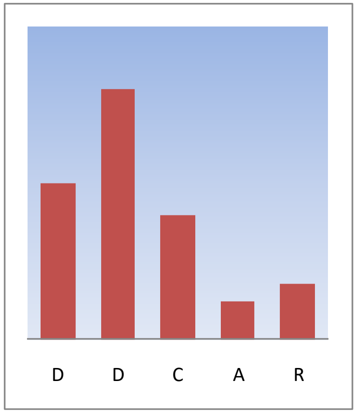
The 'novelist'



The 'data salvager'



The 'lab freak'



The 'smart researcher'

- Definition
- Design
- Data Collection
- Analysis
- Reporting

The smart researcher

- time spent planning and designing an experiment at the outset saves time and money in the long run
- optimises the lab work and reduces the time spent in the lab
- the design of the experiment is most important and governs how data are analysed
- minimises time spent at the data analysis stage
- can look ahead to the reporting phase with peace of mind since early phases of study are carefully planned and formalised (proposal, protocol)

The design of insightful experiments is a process informed by statistical thinking

STATISTICS

- Specialist skill
- Science
- Technology
- Closure, seclusion
- Introvert
- Discrete interventions
- Builds on good thinking

STATISTICAL THINKING

- Generalist skill
- Informed practice
- Principles, patterns
- Ambiguous, dialogue
- Extravert
- Permeates research process
- Valued skill in itself

The power of statistics is founded upon good statistical thinking

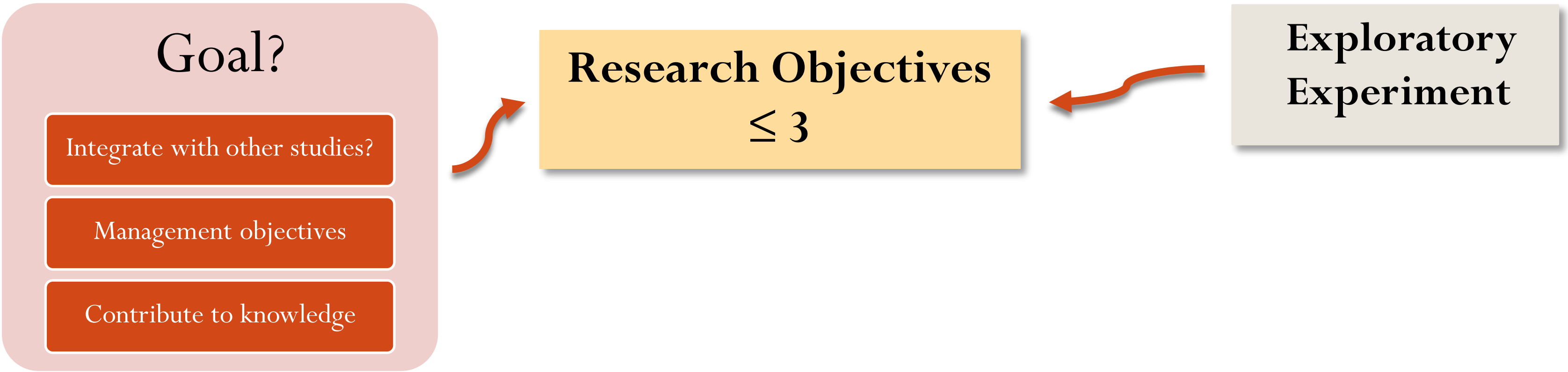
The seven principles of statistical thinking

1. Time spent thinking on the conceptualization and design of an experiment is time wisely spent
2. The design of an experiment reflects the contributions from different sources of variability
3. The design of an experiment balances between its internal validity and external validity
4. Good experimental practice provides the clue to bias minimisation
5. Good experimental design is the clue to the control of variability
6. Experimental design integrates various disciplines
7. A priori consideration of statistical power is an indispensable pillar of an effective experiment

The Smart Design of Animal Experiments

III. Planning the Experiment

The planning process



Goal?

Integrate with other studies?

Management objectives

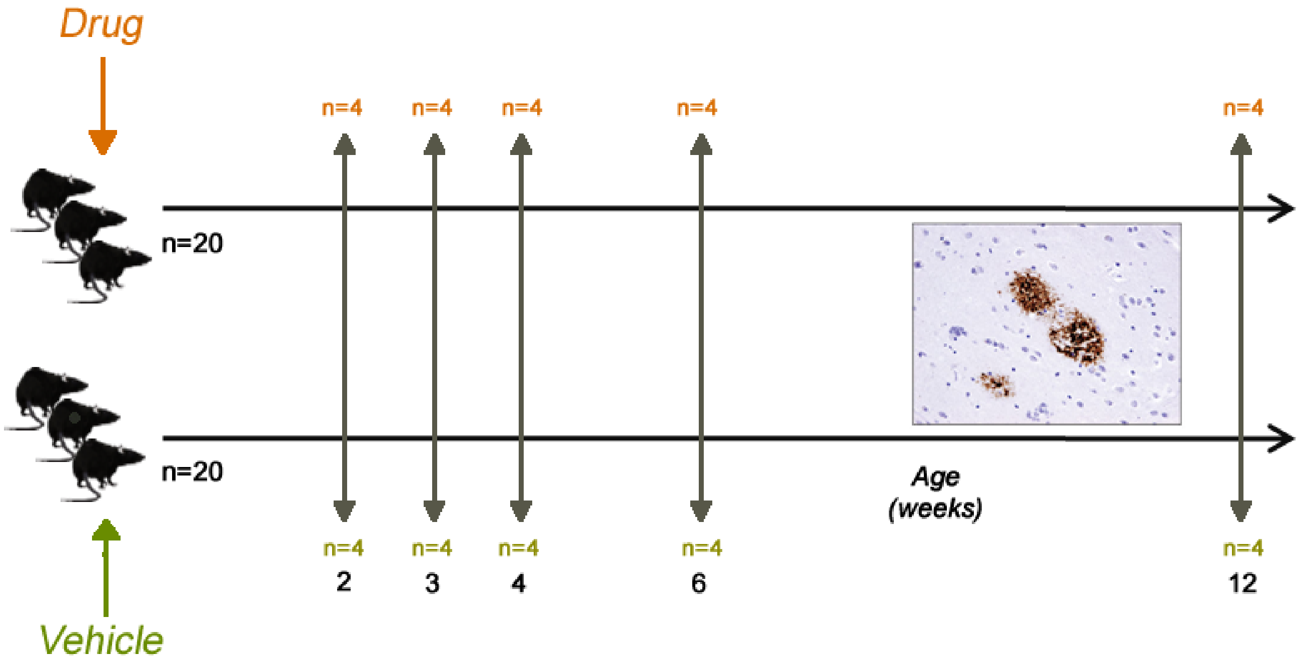
Contribute to knowledge

Research Objectives ≤ 3

Exploratory Experiment

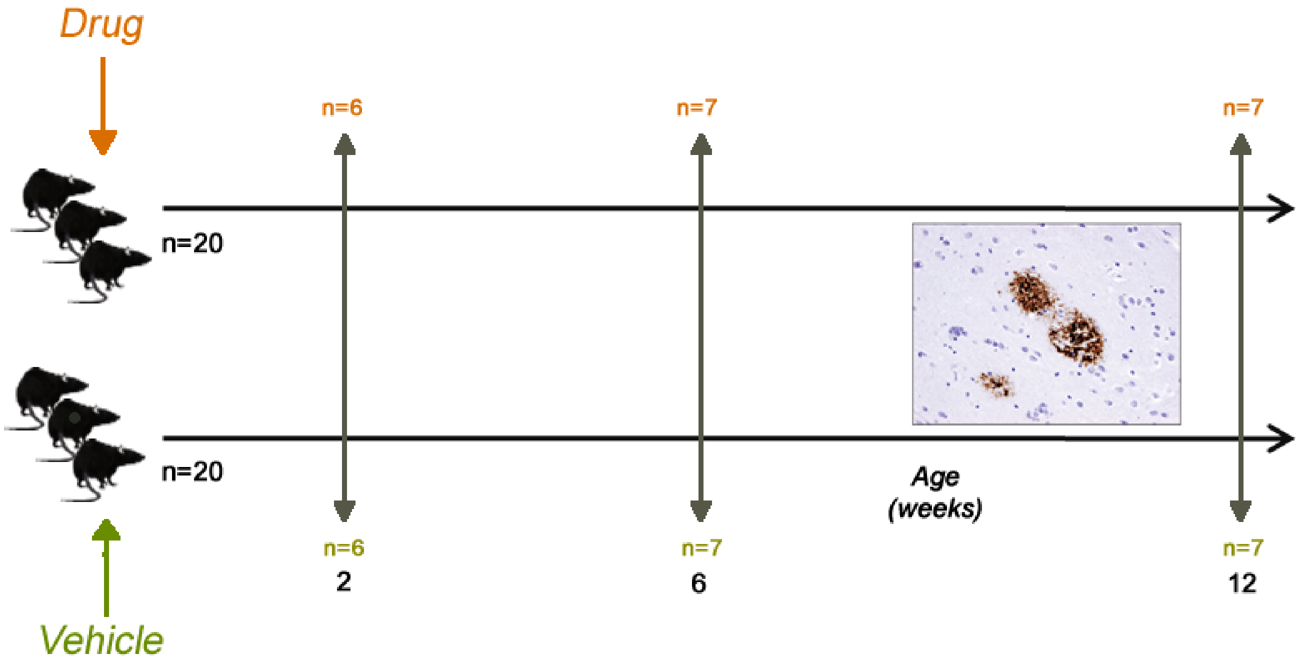
The planning process
The number of objectives

Effect of pharmacological treatment on amyloid plaque disposition in transgenic mice



The planning process
The number of objectives

Effect of pharmacological treatment on amyloid plaque disposition in transgenic mice



The planning process
The number of research objectives

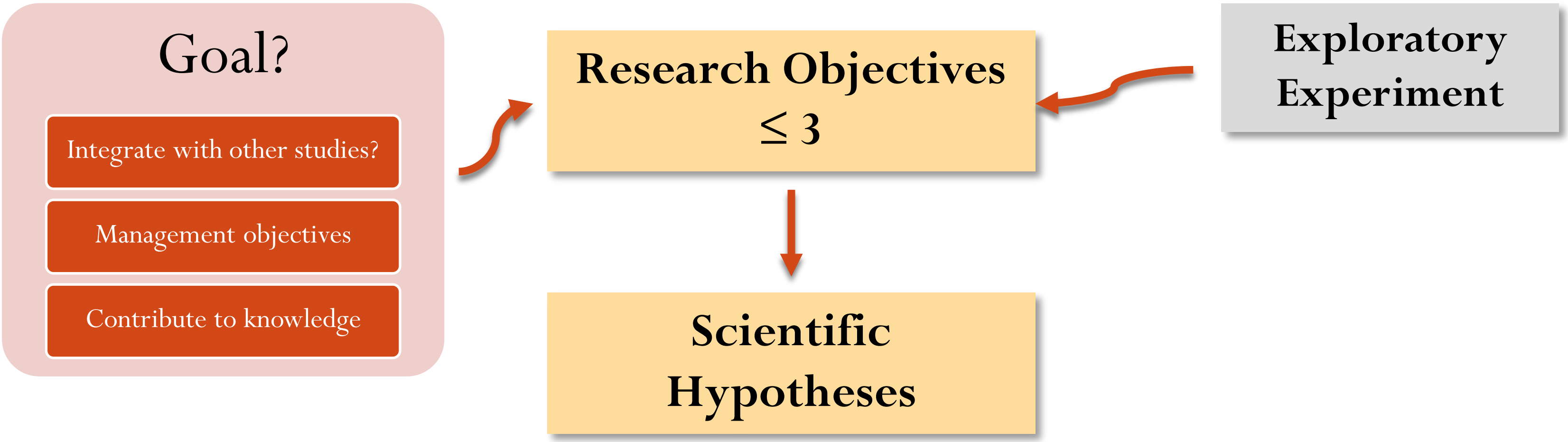
Number of research objectives should be limited

(≤ 3 objectives)

Example Séralini study:

- 10 treatments, females and males = 20 groups
- 200 animals/20 = 10 animals/group
- 50 animals/group "golden standard" in toxicology

The planning process



Goal?

Integrate with other studies?

Management objectives

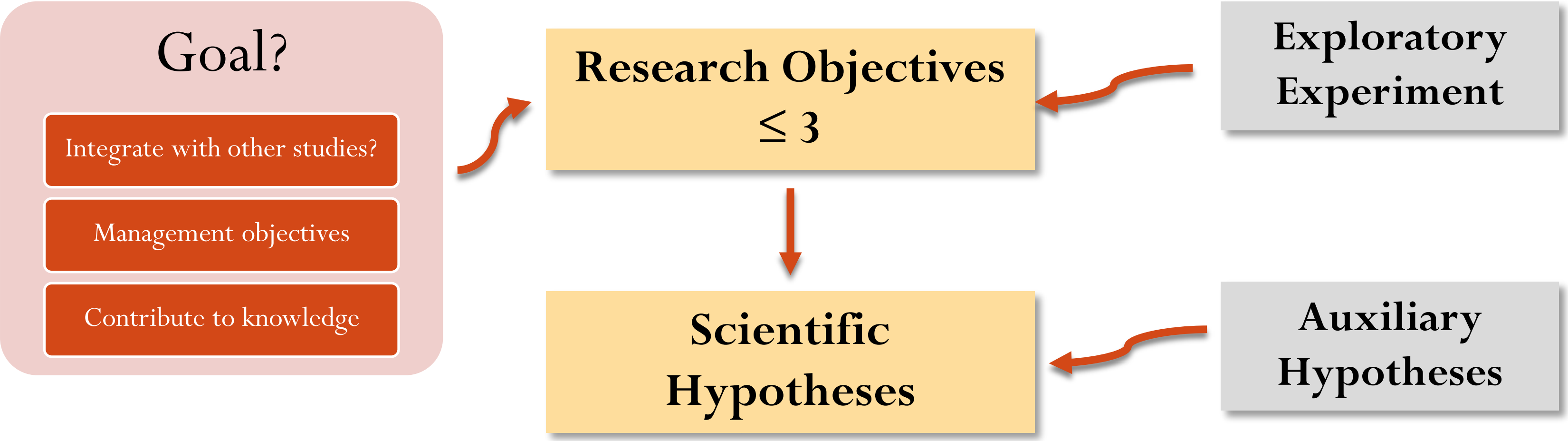
Contribute to knowledge

Research Objectives ≤ 3

Exploratory Experiment

Scientific Hypotheses

The planning process



Goal?

Integrate with other studies?

Management objectives

Contribute to knowledge

Research Objectives
≤ 3

Exploratory Experiment

Scientific Hypotheses

Auxiliary Hypotheses

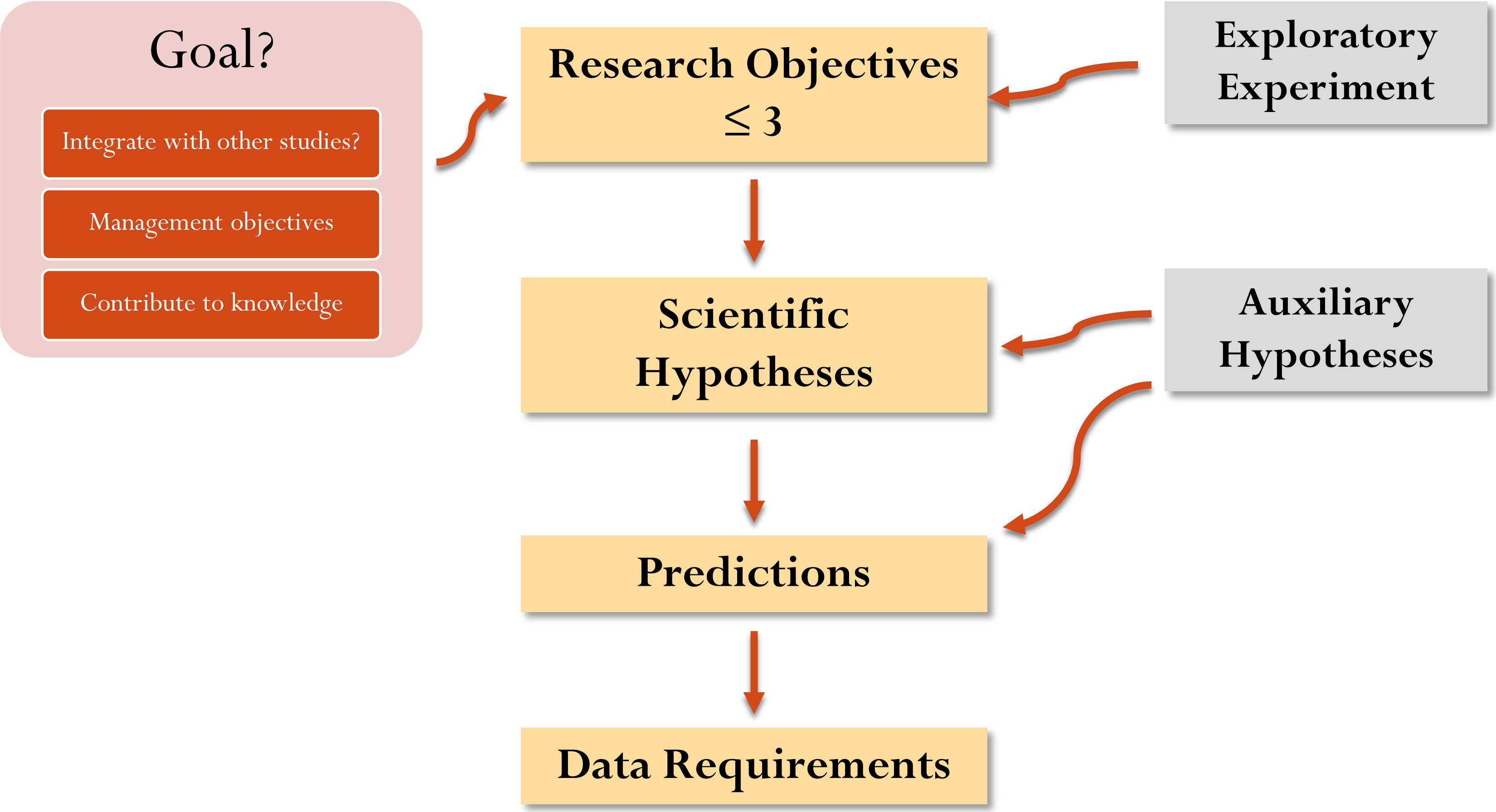
The planning process
Importance of auxiliary hypotheses

Reliance on auxiliary hypotheses is the rule rather than the exception in testing scientific hypotheses (Hempel, 1966)

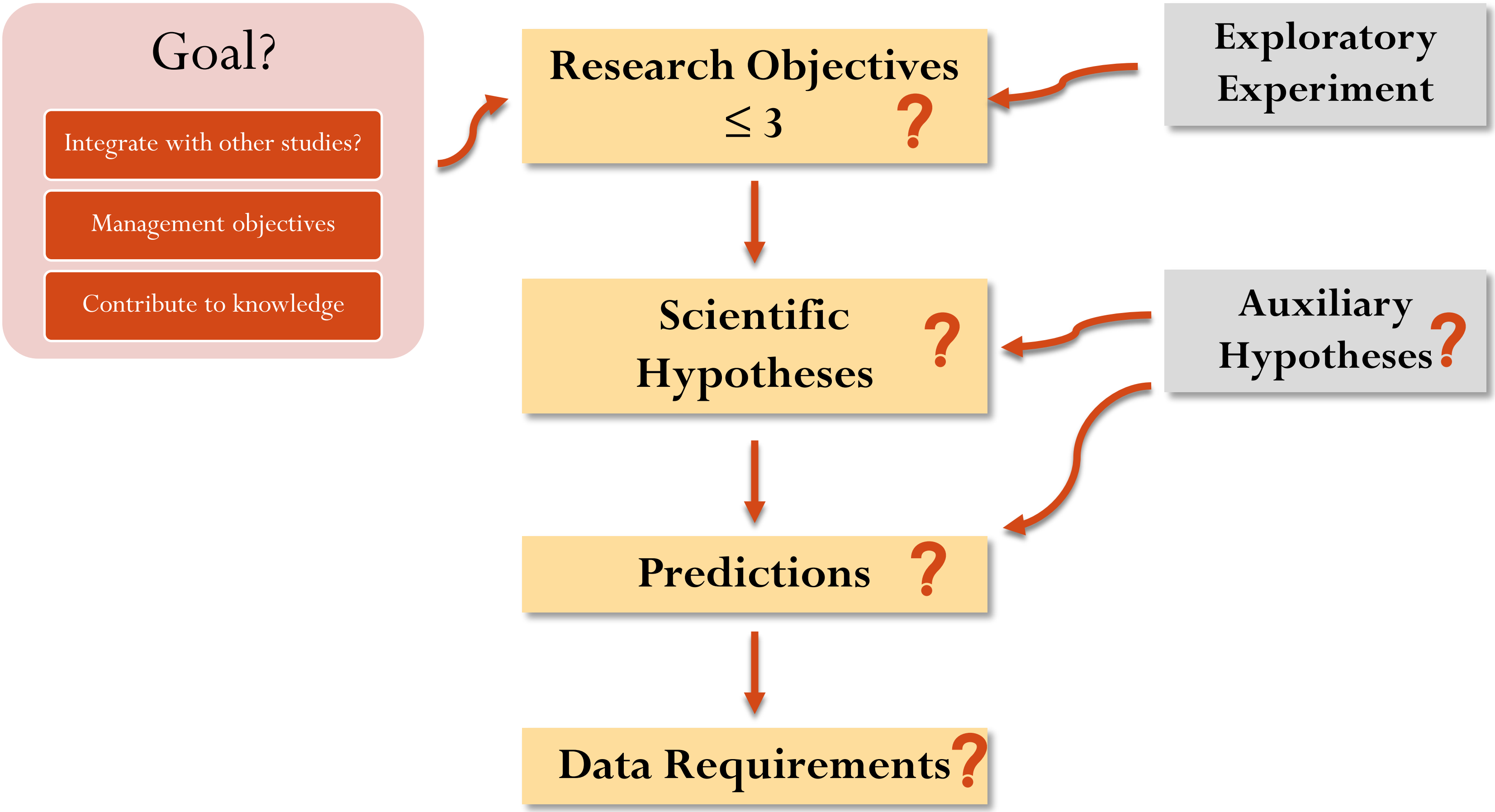
Example Séralini study:

- Use of Sprague-Dawley breed of rats as model for "human" cancer
- Known to have higher mortality in 2 year studies
- Prone to develop spontaneous tumors

The planning process



The planning process



Goal?

Integrate with other studies?

Management objectives

Contribute to knowledge

Research Objectives ≤ 3 ?

Exploratory Experiment

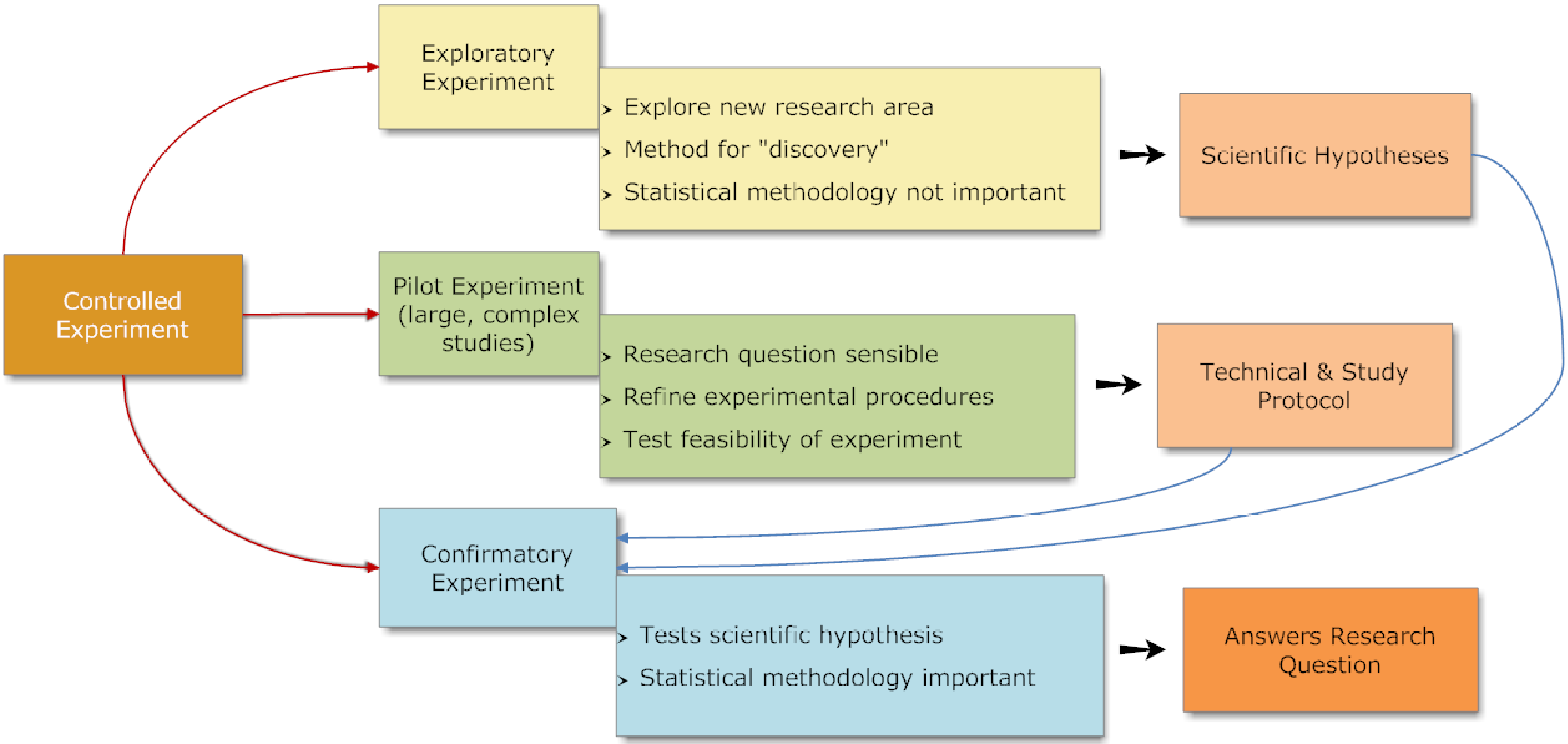
Scientific Hypotheses ?

Auxiliary Hypotheses ?

Predictions ?

Data Requirements ?

Types of experiments



Controlled Experiment

Exploratory Experiment

- Explore new research area
- Method for "discovery"
- Statistical methodology not important

Scientific Hypotheses

Pilot Experiment (large, complex studies)

- Research question sensible
- Refine experimental procedures
- Test feasibility of experiment

Technical & Study Protocol

Confirmatory Experiment

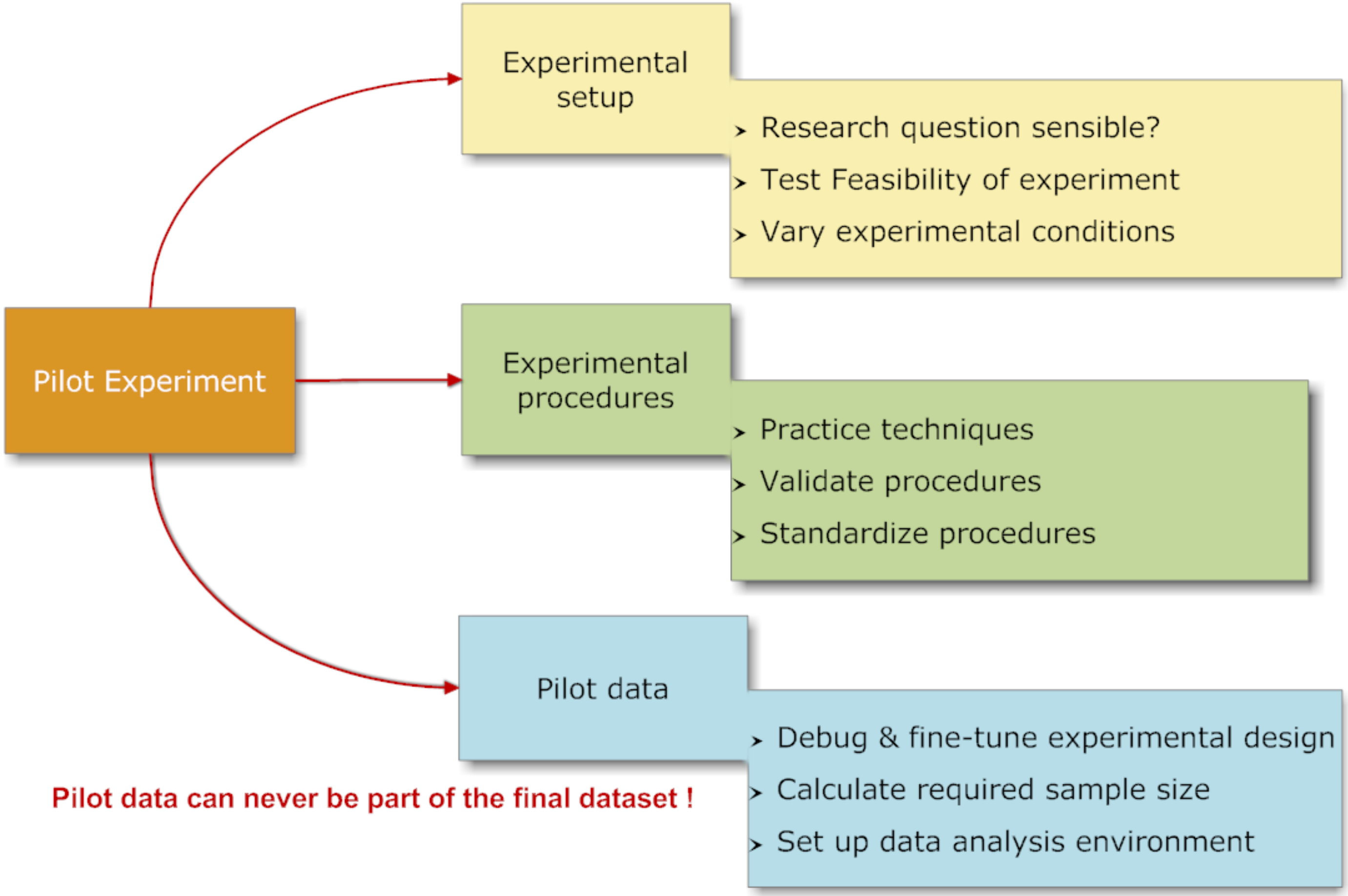
- Tests scientific hypothesis
- Statistical methodology important

Answers Research Question

Abuse of exploratory experiments

- Most published work in biomedical research is exploratory
- Too often published as if confirmatory experiments
- Do not unequivocally answer research question
- Use of same data that generated research hypotheses to prove these hypotheses involves **circular reasoning**
- Contributes to false positive (unreplicable) results
- Exploratory data analysis as opposed to confirmatory tests of hypotheses (e.g. Kimmelman, 2014)
- Inferential statistics should be interpreted and published as "**for exploratory purposes only**"
- Séralini study was actually conceived as exploratory

Role of the pilot experiment



Pilot Experiment

Experimental setup

- Research question sensible?
- Test Feasibility of experiment
- Vary experimental conditions

Experimental procedures

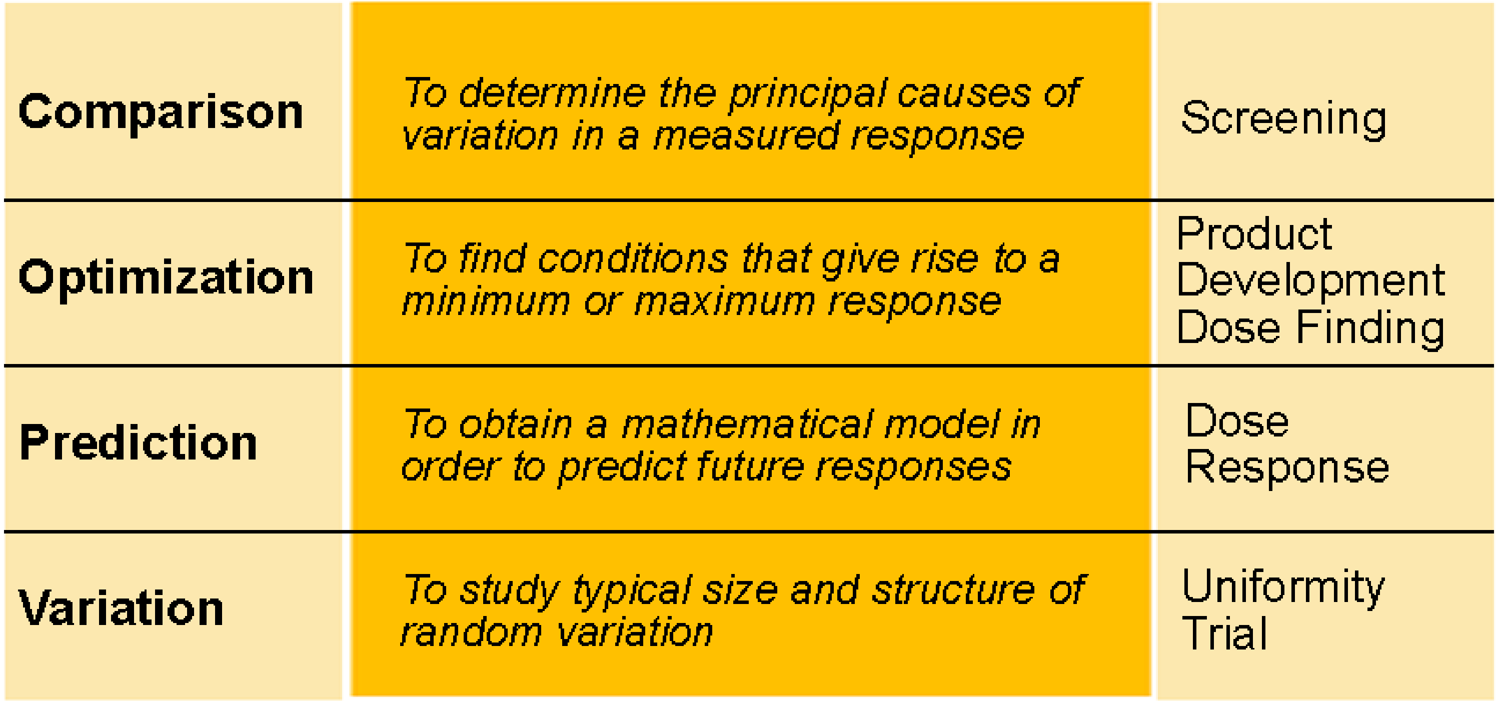
- Practice techniques
- Validate procedures
- Standardize procedures

Pilot data

- Debug & fine-tune experimental design
- Calculate required sample size
- Set up data analysis environment

Pilot data can never be part of the final dataset !

Types of experiments objective - usage



Comparison	*To determine the principal causes of variation in a measured response*	Screening
Optimization	*To find conditions that give rise to a minimum or maximum response*	Product Development Dose Finding
Prediction	*To obtain a mathematical model in order to predict future responses*	Dose Response
Variation	*To study typical size and structure of random variation*	Uniformity Trial

The Smart Design of Animal Experiments

IV. Principles of Statistical Design

Replication
The cornerstone of science

Replicate what?

- intervention-entity pairs (drugs-animals)
- multiple measurements outcome
- several applications in same animal
- experiment at different location

Replication
The cornerstone of science

Replicate what?

- intervention-entity pairs (drugs-animals)
- multiple measurements outcome
- several applications in same animal
- experiment at different location

Types of replication

- **genuine repeat**
  independently repeated data
- **pseudoreplication**
  no evidence for reproducibility of results

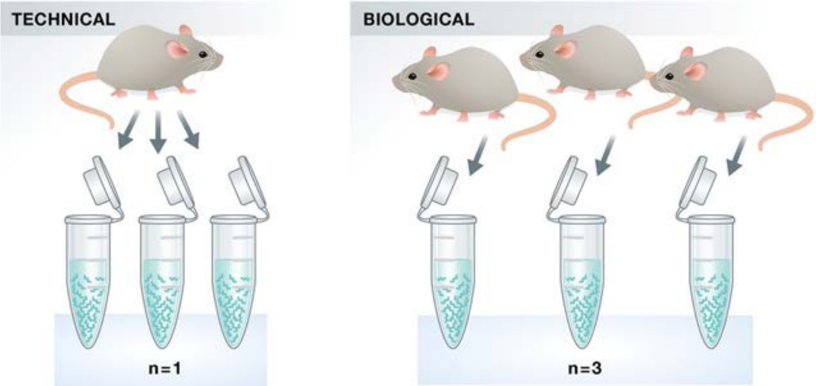
Replication
The cornerstone of science

Replicate what?

- intervention-entity pairs (drugs-animals)
- multiple measurements outcome
- several applications in same animal
- experiment at different location

Types of replication

- **genuine repeat**
  independently repeated data
- **pseudoreplication**
  no evidence for reproducibility of results



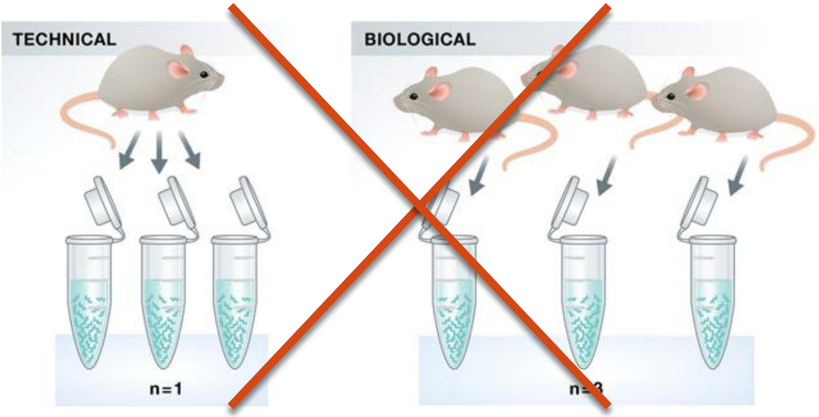
Replication
The cornerstone of science

Replicate what?

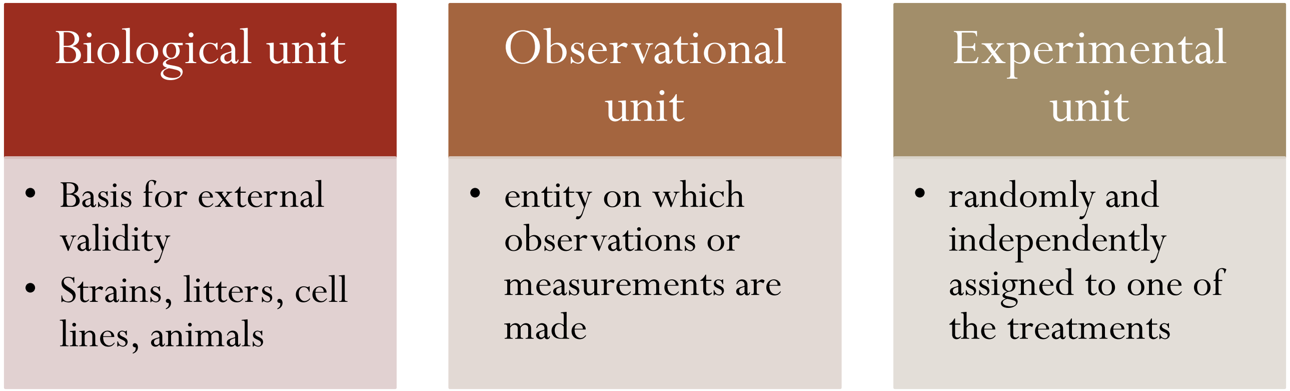
- intervention-entity pairs (drugs-animals)
- multiple measurements outcome
- several applications in same animal
- new experiment at different location

Types of replication

- **genuine repeat**
  independently repeated data
- **pseudoreplication**
  no evidence for reproducibility of results



The importance of proper replication
Levels of replication



Biological unit

- Basis for external validity
- Strains, litters, cell lines, animals

Observational unit

- entity on which observations or measurements are made

Experimental unit

- randomly and independently assigned to one of the treatments

The importance of proper replication
The experimental unit

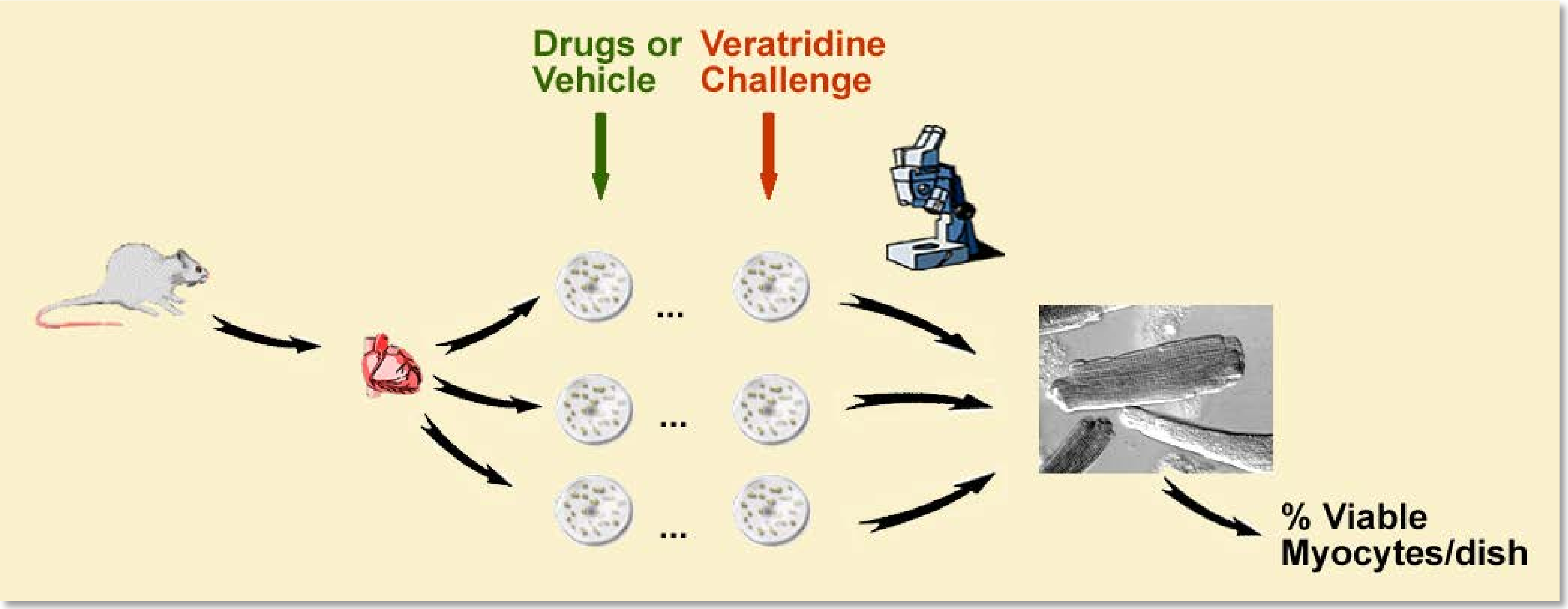
Smallest division of experimental material to which a treatment can be applied

- 2 units = 2 different treatment
- **independence of units**
- wrong choice most serious mistake in design and analysis
- most frequent error in biomedical studies

- a biological unit of interest;
- groups of biological units;
- parts of a biological unit;
- a sequence of observations or measurements on a biological unit

Genuine repeats = Replication of experimental units

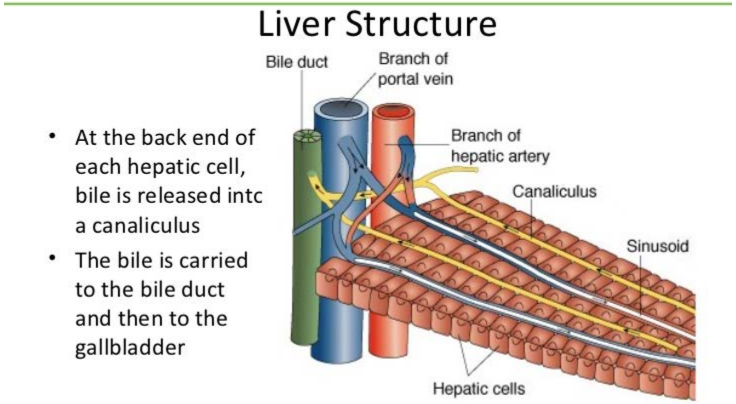
The importance of proper replication
Cardiomyocyte experiment



- Biological unit: rat (1)
- Experimental unit: Petri dish
- Observational unit: individual cell

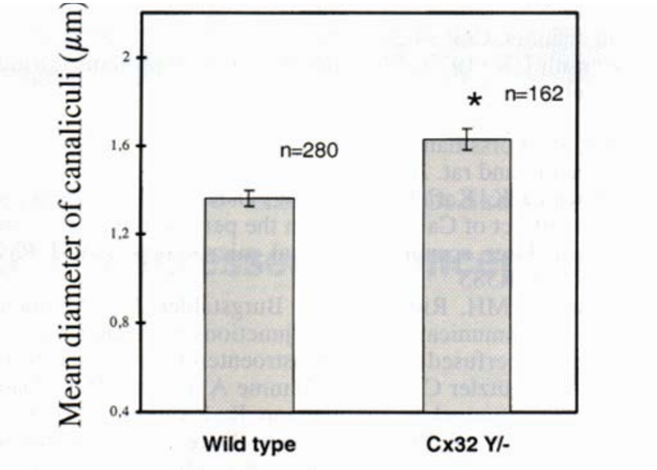
The importance of proper replication
Pseudoreplication

- Temme *et al.* (2001)
- Comparison of two genetic strains of mice, wild-type and connexin32-deficient
- 3 animals/genotype
- Diameters of bile caniliculi in livers
- Statistically significant difference between two strains (P < 0.005 = 1/200)
- Experimental unit = Observational unit
- Is this correct?



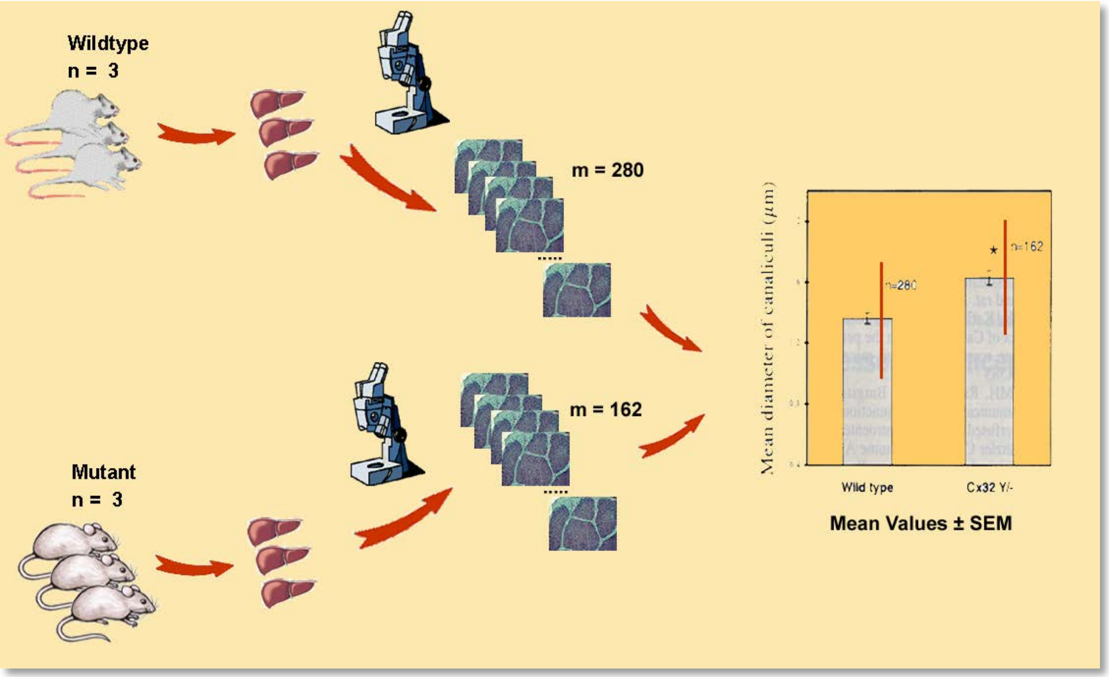
Liver Structure

- At the back end of each hepatic cell, bile is released intc a canaliculus
- The bile is carried to the bile duct and then to the gallbladder



Means ± SEM from 3 livers

The importance of proper replication
Pseudoreplication



- Experimental unit = animal
- Results not conclusive at all
- Wrong choice made out of ignorance, or out of **convenience**? Hanlon's razor?
- This was published in **peer reviewed** journal !!!

The importance of proper replication
Cage-wise treatment application

- Rivenson et al. (1988) Toxicity of N-nitrosamines
- Rats housed 3/cage
- Treatment supplied in drinking water
- **Biological unit = rat**
- Impossible to treat any 2 rats differently
- Rats within a cage not independent
- Rat **not** experimental unit
- **Experimental unit = cage**

- Same remarks for Séralini study
- Rats housed 2/cage, treatment in food
- Rats in same cage are not independent of one another
- Number of experimental units is not 10/treatment, but 5/treatment

The choice of the experimental unit
Animals housed together

Group housing of gregarious animals is requested by authorities (Council of Europe, 2006)

Animals interact, e.g.:

- socially dominant animal prevents others from eating/drinking;
- aggression in male mice
- “barbering” of rats and mice
- reduced surface area and behavioural thermoregulation in mice;
- cross-contamination by excrements
- …

The choice of the experimental unit
Animals housed together

Group housing of gregarious animals is requested by authorities (Council of Europe, 2006)

Animals interact, e.g.:

- socially dominant animal prevents others from eating/drinking;
- aggression in male mice
- "barbering" of rats and mice
- reduced surface area and behavioral thermoregulation in mice;
- cross-contamination by excrements
- …

- **Experimental unit = Cage**
- Required number of animals larger
- Var cages < Var animals
- # cages < # individual animals
- 10 animals/group housed individually = 12/group, housed as 4/cage

The choice of the experimental unit
Group housing - Example

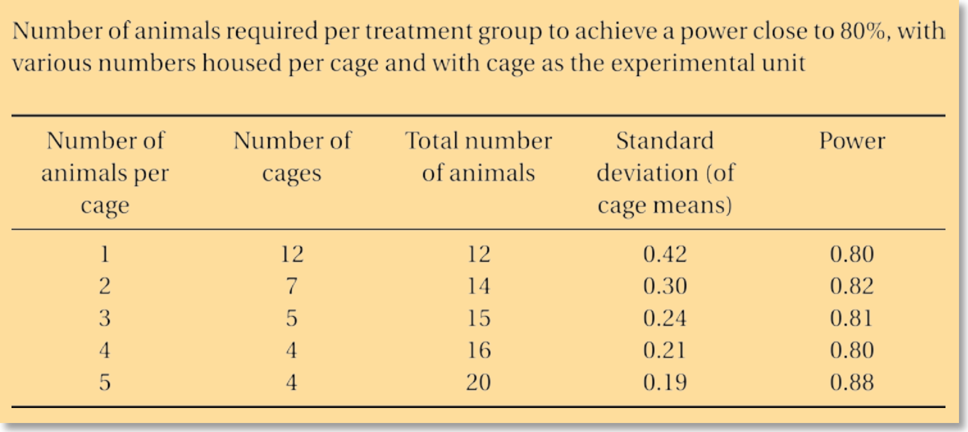
- Temme study
- 25 measurements/animal
- Outcome = average canaliculi diameter/animal
- SD = 0.42 µm
- 12 animals/treatment group to detect 0.5 µm difference

The choice of the experimental unit
Group housing - Example

- Temme study
- 25 measurements/animal
- Outcome = average canaliculi diameter/animal
- SD = 0.42 µm
- 12 animals/treatment group to detect 0.5 µm difference

Group housing of animals:

- *m* animals/cage
- outcome = mean value of cage
- SD reduced by $\sqrt{m}$
- 2 – 3 mice/cage optimum



Number of animals required per treatment group to achieve a power close to 80%, with various numbers housed per cage and with cage as the experimental unit

Number of animals per cage	Number of cages	Total number of animals	Standard deviation (of cage means)	Power
1	12	12	0.42	0.80
2	7	14	0.30	0.82
3	5	15	0.24	0.81
4	4	16	0.21	0.80
5	4	20	0.19	0.88

Group housing makes animals more content and thereby reduces variability (Fry, 2014)

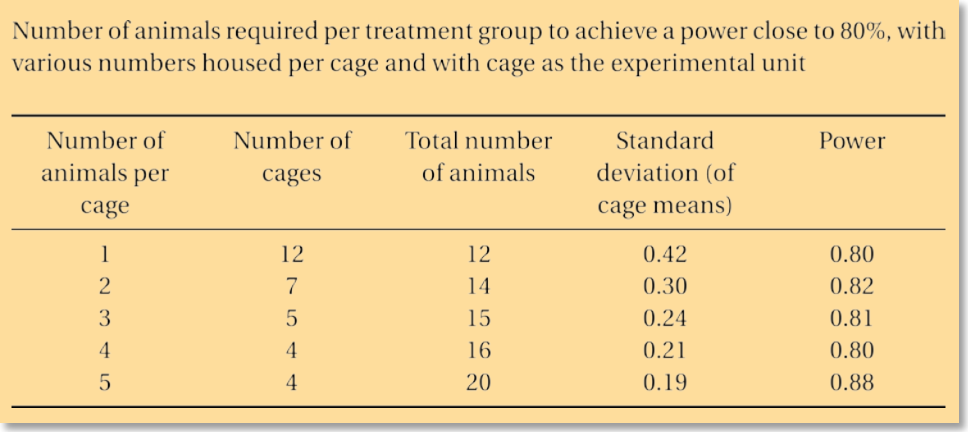
The choice of the experimental unit
Group housing - Example

- Temme study
- 25 measurements/animal
- Outcome = average canaliculi diameter/animal
- SD = 0.42 µm
- 12 animals/treatment group to detect 0.5 µm difference

- Experimental unit: cage
- Biological unit: animal (mouse)
- Observational unit: histological section

Group housing of animals:

- *m* animals/cage
- outcome = mean value of cage
- SD reduced by $\sqrt{m}$
- 2 – 3 mice/cage optimum

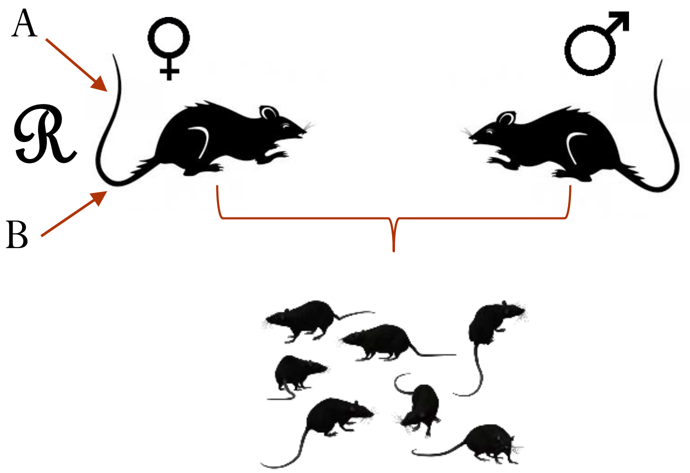


Number of animals required per treatment group to achieve a power close to 80%, with various numbers housed per cage and with cage as the experimental unit

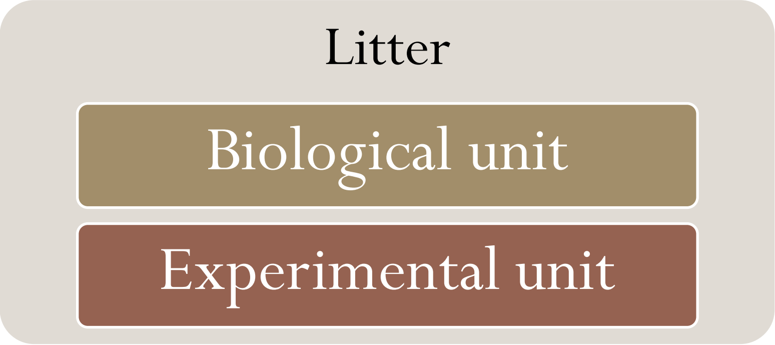
Number of animals per cage	Number of cages	Total number of animals	Standard deviation (of cage means)	Power
1	12	12	0.42	0.80
2	7	14	0.30	0.82
3	5	15	0.24	0.81
4	4	16	0.21	0.80
5	4	20	0.19	0.88

Group housing makes animals more content and thereby reduces variability (Fry, 2014)

The choice of the experimental unit
Reproductive studies



- Homozygous mutant female rats randomly assigned to drug-treatment or control
- Mated with homozygous mutant males, producing homozygous mutant offspring



Litter

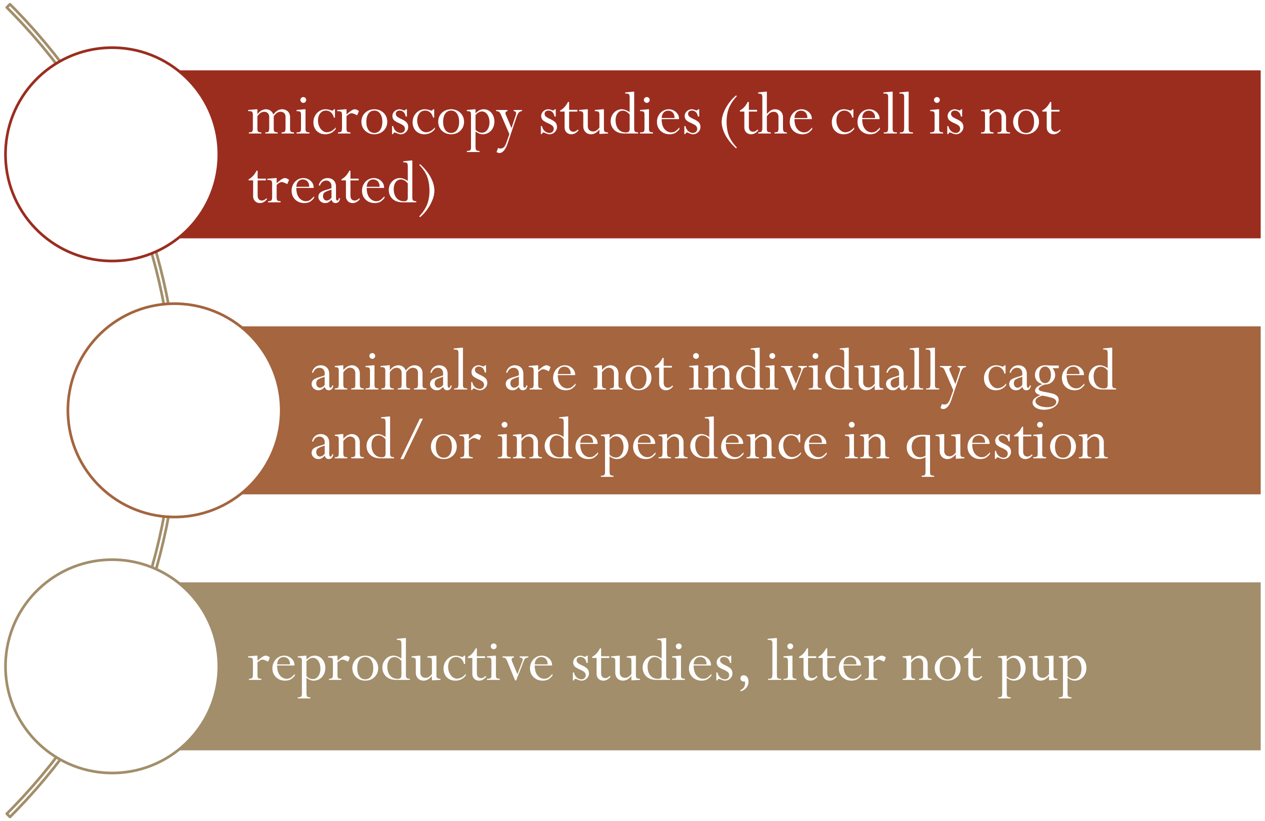
Biological unit

Experimental unit

The choice of the experimental unit
The reverse case

- Regrowth of epithelium across a wound
- 4 treatment conditions
- 12 small wounds in back of pig
- Wounds far enough apart for independence
- Each treatment condition randomly assigned to 3 different wounds
- Experimental unit = wound, not pig
- 3 E.U./ treatment condition

The choice of the experimental unit Summary - Pseudoreplication



microscopy studies (the cell is not treated)

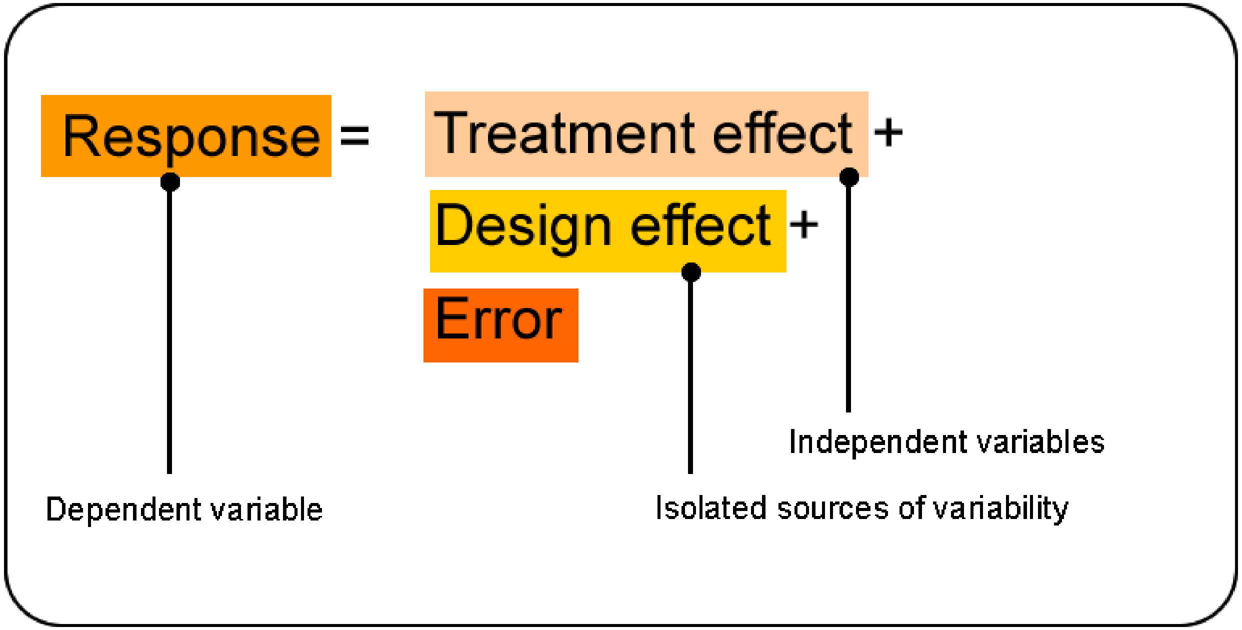
animals are not individually caged and/or independence in question

reproductive studies, litter not pup

Some terminology

- **Factor:**
  the condition that we manipulate in the experiment (e.g. concentration of a drug, temperature)
- **Factor level:**
  the value of that condition (e.g. 1.25 $mg.kg^{-1}$, female, intravenous)
- **Treatment:**
  combination of factor levels (e.g. 1.25 $mg.kg^{-1}$ given intravenously in females)
- **Response or dependent variable:**
  characteristic that is measured
- See Appendix B, Glossary of Statistical Terms

The structure of the response



Response = Treatment effect + Design effect + Error

Dependent variable

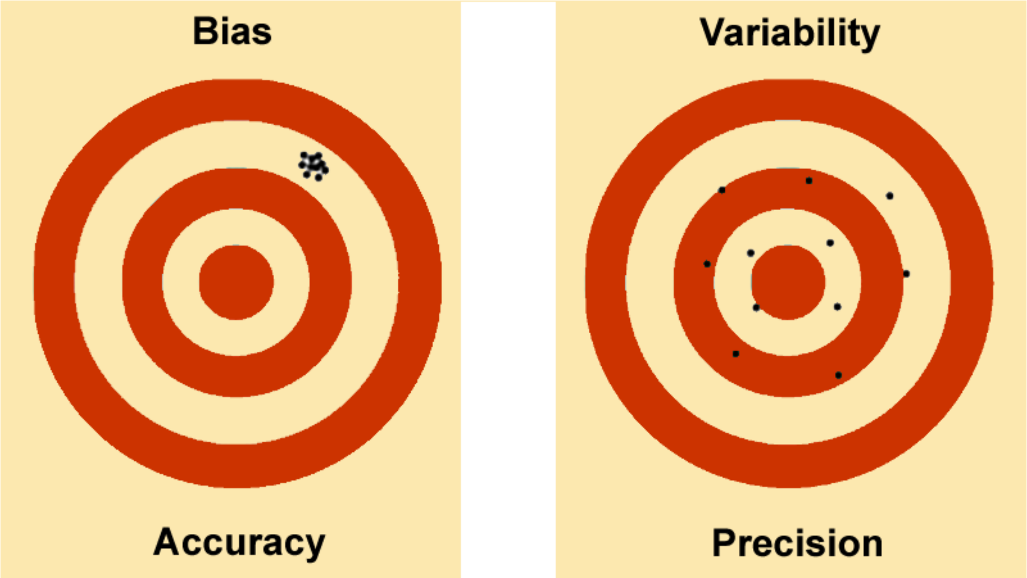
Independent variables

Isolated sources of variability

Assumptions

- Additive model
- Treatment effects constant
- Treatment effects in one unit do not affect other units

Bias and variability
Failure to minimize bias ruins an experiment



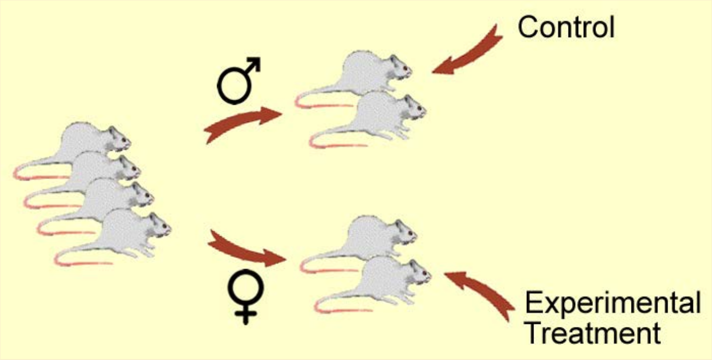
Failure to minimize **bias** results in lost experiments

Failure to control **variability** can sometimes be remediated after the facts

Allocation to treatment is most important source of **bias**

Failure to minimize bias ruins an experiment

- Bias enters by the way experimental units are allocated to treatment groups



- Gender = **Confounding Bias**

- Cage effect:
  all animals of treatment in same cage, effect of treatment confounded with that of cage

- Rack location/shelf level:
  effect on body temperature, food consumption, body weight, neoplasms (Gore & Stanley 2004; Greenman 1983, 1984)

Types of confounding Bias

- Selection bias
  - Bias caused by non-random allocation of animals to treatment groups
    e.g. healthy animals assigned to high dosage group

- Performance bias
  - Differences in level of husbandry care given to animals across treatment groups
    e.g. sick animals in control group are given the benefit of doubt and kept longer alive

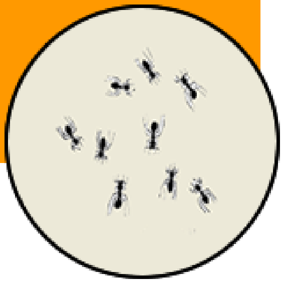
- Detection bias (observer bias)
  - Researcher assessing the effect of the treatment knows which treatment the animal received
    e.g. subjective evaluation of histologic material

- Attrition bias
  - Unequal occurrence and handling of deviations from the protocol and loss to follow-up between treatment groups
    e.g. many animals that were excluded from the high-dose group

Why engineers don't care very much about statistics

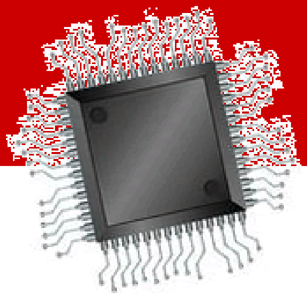
- Comparison
- Variation

Natural and
Social Systems



Uncontrolled variation
large
compared to effects

High Technology



- Optimization
- Prediction

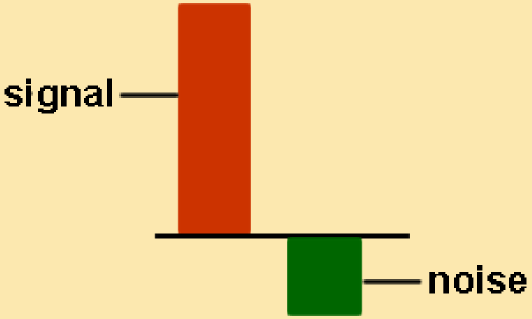
Uncontrolled variation
small
compared to effects

The basic dilemma:
balancing internal and external validity

INTERNAL VALIDITY

⬍

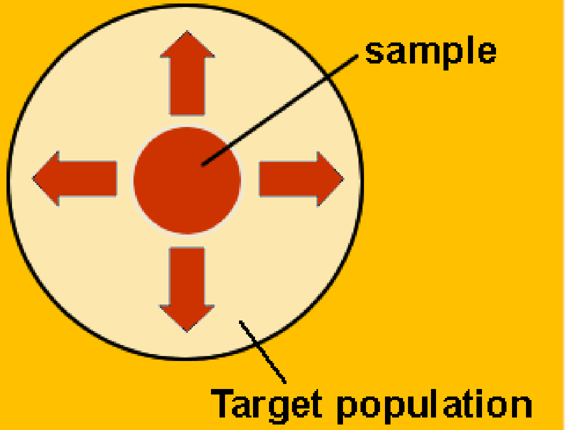
Maximize S/N



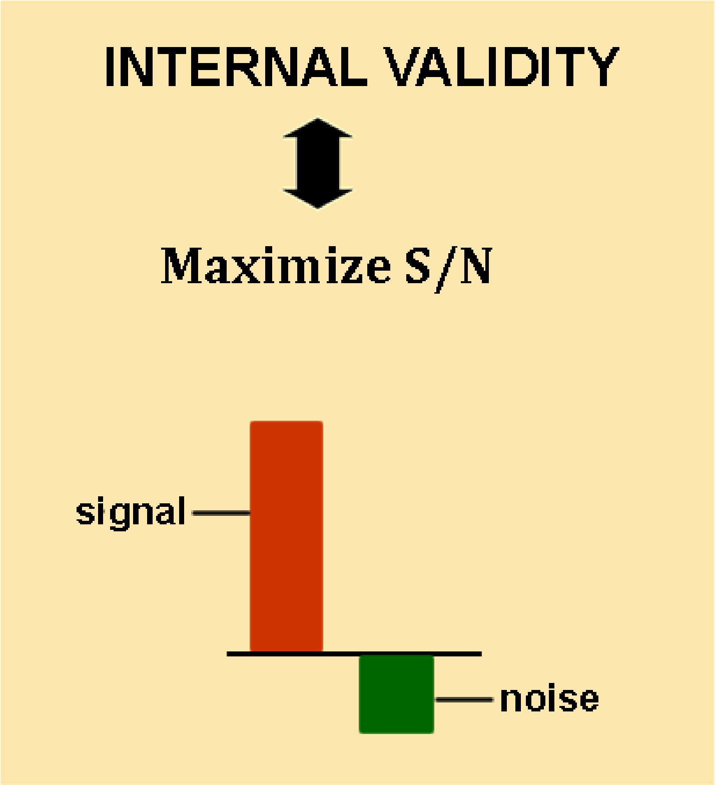
EXTERNAL VALIDITY

⬍

Ensure generalizability



The basic dilemma:
balancing internal and external validity



INTERNAL VALIDITY

Maximize S/N

Increase signal

- more sensitive material (use of pilot experiment)

Reduce noise

- more experimental units
- experimental material as much alike as possible

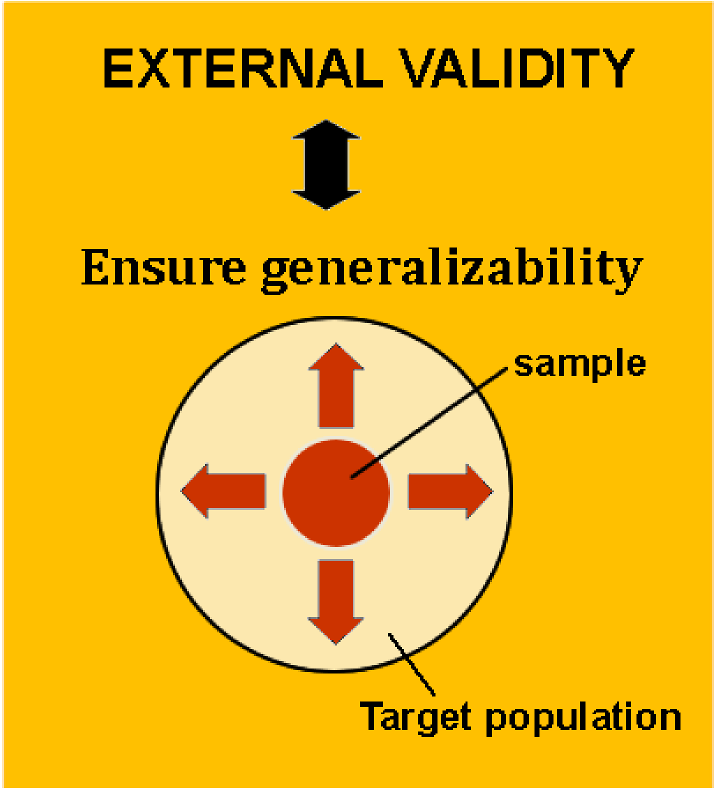
The basic dilemma: balancing internal and external validity

Choice of target population

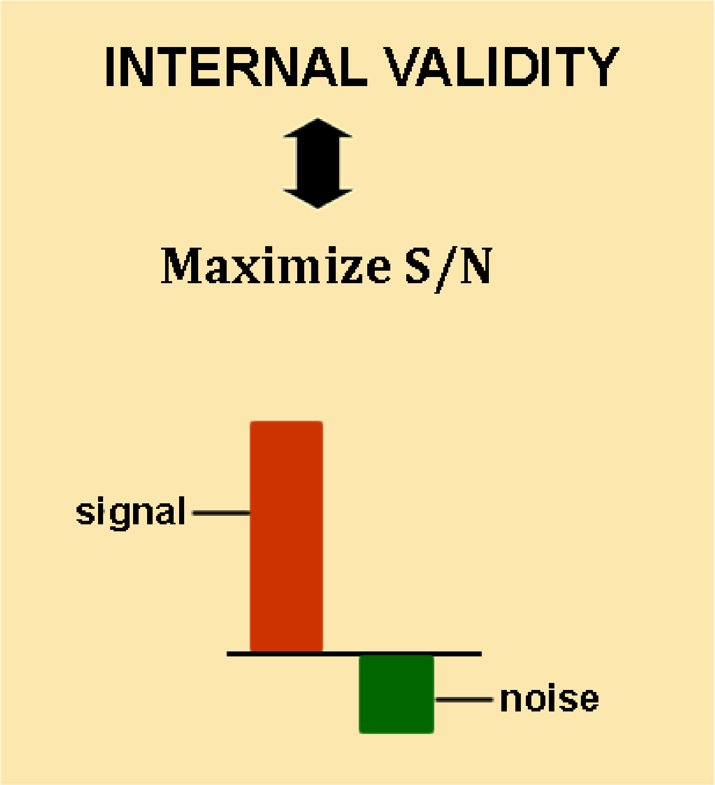
Sampling from target population

Experimental procedures

Replication of biological unit basis for generalisability

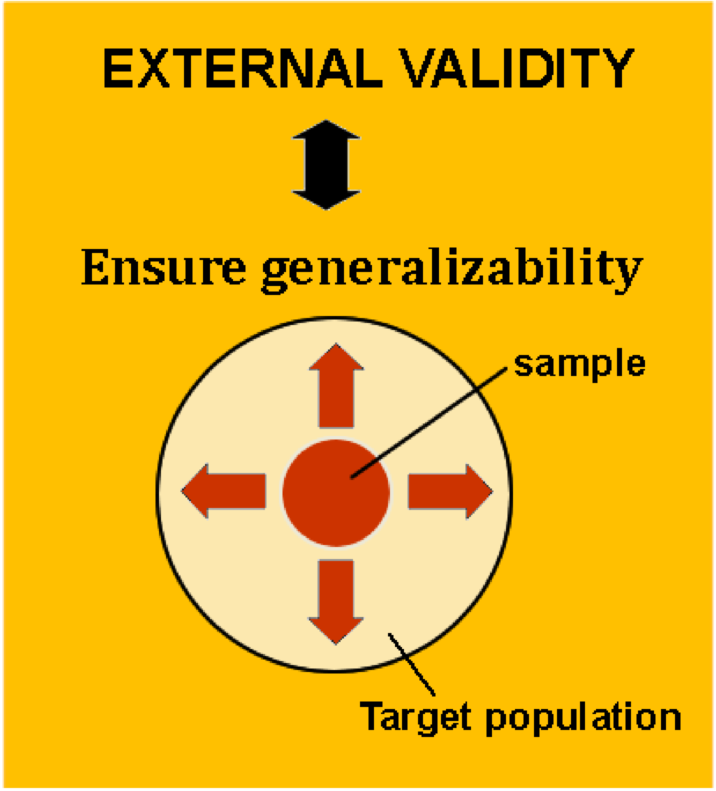


The basic dilemma: balancing internal and external validity



INTERNAL VALIDITY

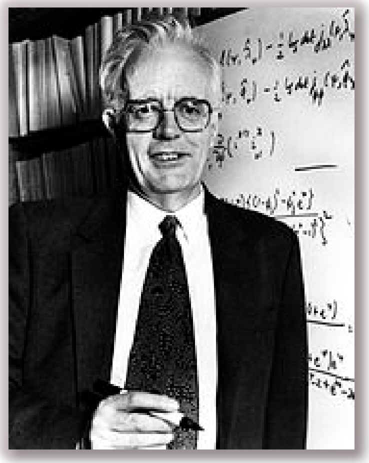
Maximize S/N



EXTERNAL VALIDITY

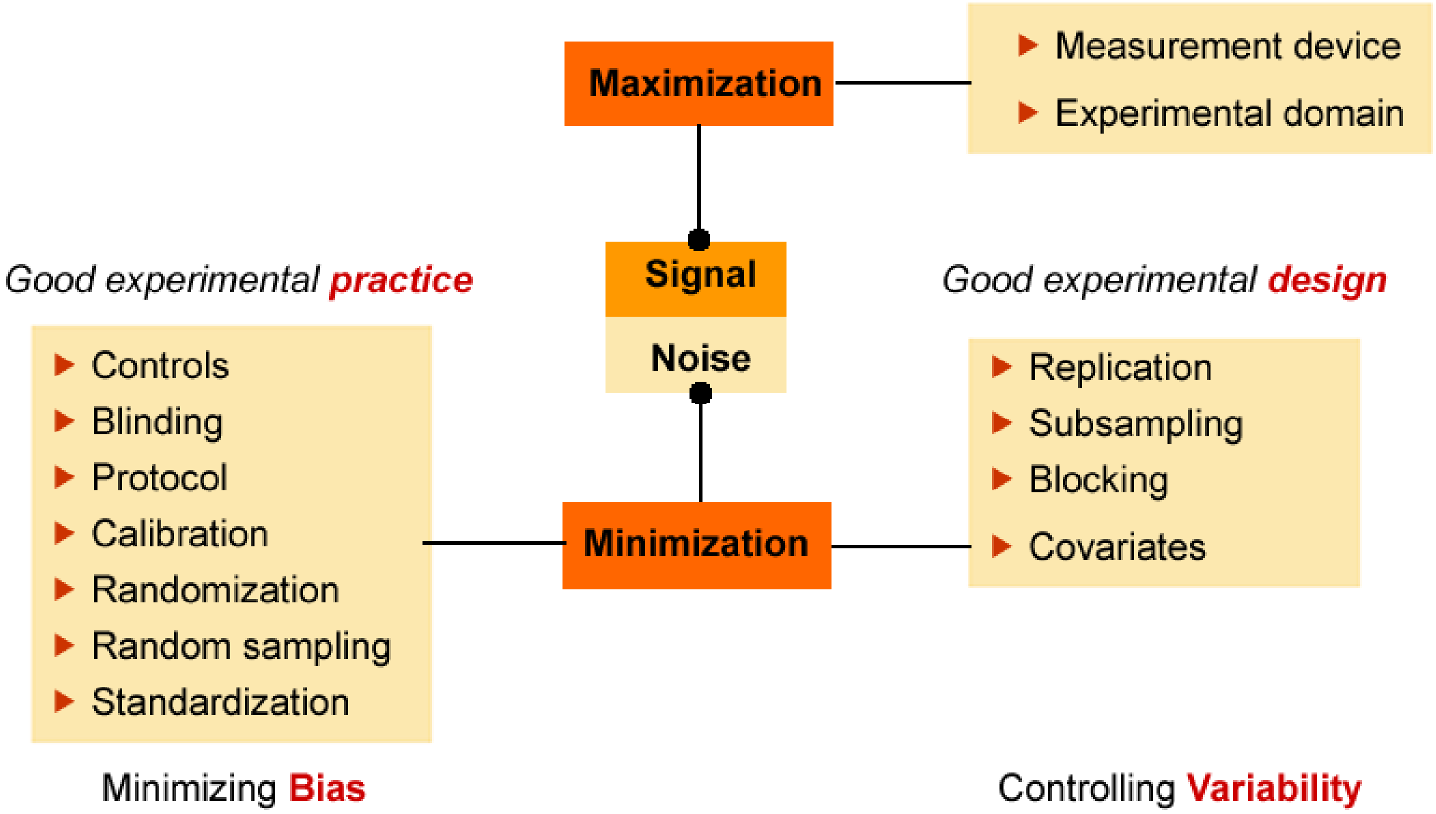
Ensure generalizability

Requirements for a good experiment



- Treatment comparisons free of systematic error
- Comparisons sufficiently precise (high S/N ratio)
- Wide range of external validity
- Experimental setup as simple as possible
- Uncertainty (error) must be assessable

Basic strategies for maximising Signal/Noise ratio



Maximization

- Measurement device
- Experimental domain

Signal

Noise

Minimization

Good experimental ***practice***

- Controls
- Blinding
- Protocol
- Calibration
- Randomization
- Random sampling
- Standardization

Minimizing **Bias**

Good experimental ***design***

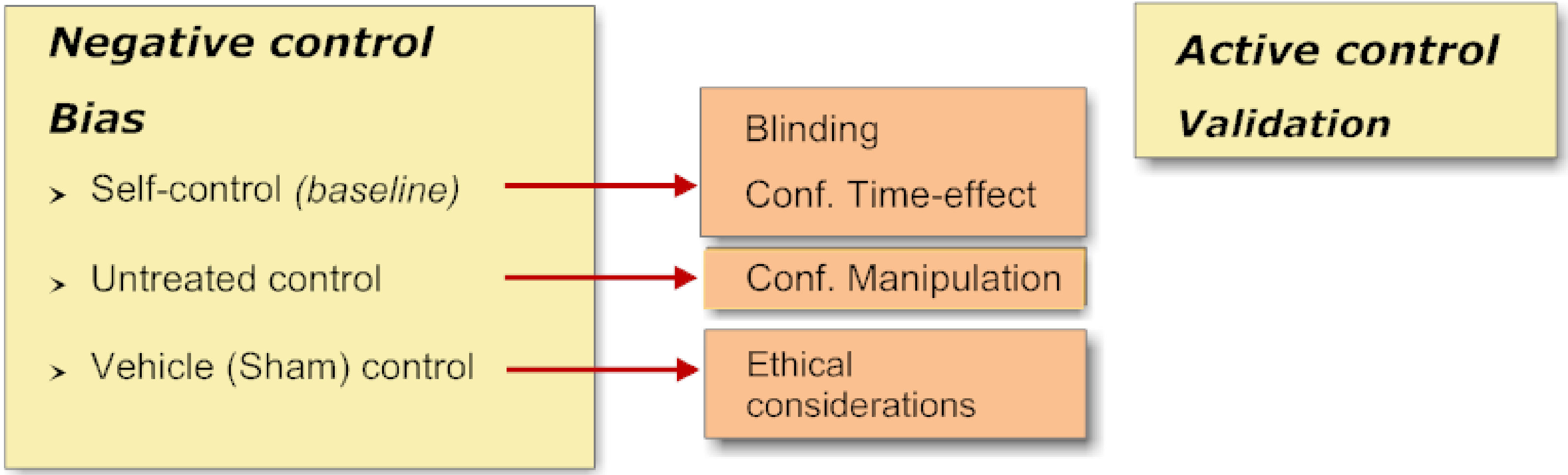
- Replication
- Subsampling
- Blocking
- Covariates

Controlling **Variability**

Minimising Bias

CONTROLS

A **control** is a standard treatment condition against which all others may be compared



Negative control

Bias

- Self-control *(baseline)* → Blinding; Conf. Time-effect
- Untreated control → Conf. Manipulation
- Vehicle (Sham) control → Ethical considerations

Active control

Validation

Minimising Bias

BLINDING

- **Single blinding** – the condition under which investigators are uninformed as to the treatment condition of experimental units

- **Double blinding** – the condition under which both *experimenter and observer* are uninformed as to the treatment condition of experimental units

- Neutralizes **investigator bias**

- Neutralizes **investigator bias** and bias in the observer's **scoring system**

In toxicological histopathology, both blinded and unblinded evaluation are recommended

Minimising Bias BLINDING

- ➢ Systematic Review (van Luijk et al. 2014):
  - Only 24% animal studies in nonclinical research use blinded assessment
  - 15% blinding of caretaker/investigator
- ➢ Study by Holman et al. (2015):
  - Effect size inflated by 27%
  - More statistically significant studies than their blinded counterpart
- ➢ **Blinding should always be considered when the response variable is subjectively evaluated**

Minimising Bias

BLINDING METHODS

- **Group blinding**

- Entire treatment groups are coded: A, B, C

- **Individual blinding**

- Each experimental unit is coded individually
- Codes are kept in *randomization list*
- Involves independent person that maintains list and prepares treatments

Minimising Bias

PROTOCOL

- Describes practical actions
- Guidelines for lab technicians
  - ✓ Manipulation of experimental units (animals, etc.)
  - ✓ Materials used
  - ✓ Logistics
  - ✓ Definitions of measurements and scoring methods
  - ✓ Data collection & processing
  - ✓ Personal responsibilities

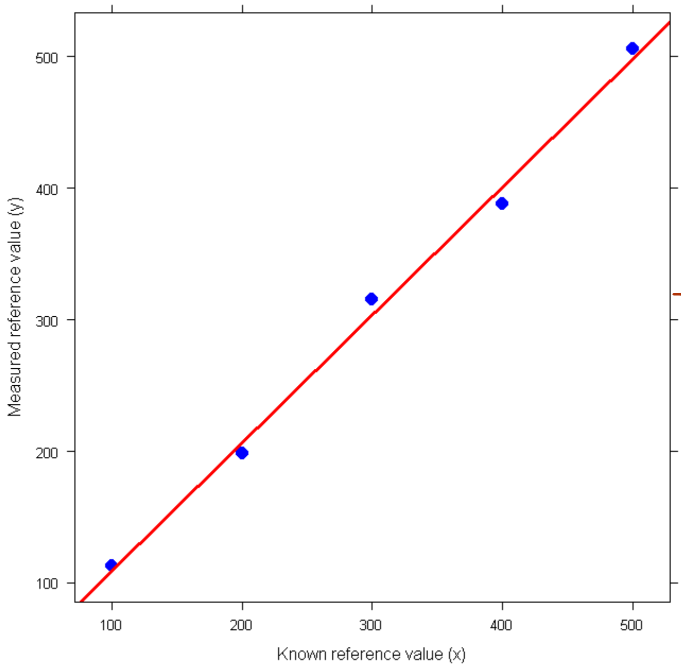
A detailed protocol is **imperative** to minimise potential bias

Minimising Bias

CALIBRATION

Calibration is an operation that compares the *output* of a measurement device to standards of known value, leading to correction of the values indicated by the measurement device

Neutralises **bias** in the investigator's **measurement system**



'Calibration curve'

$y=a+bx+\varepsilon$

Minimising Bias

RANDOMISATION

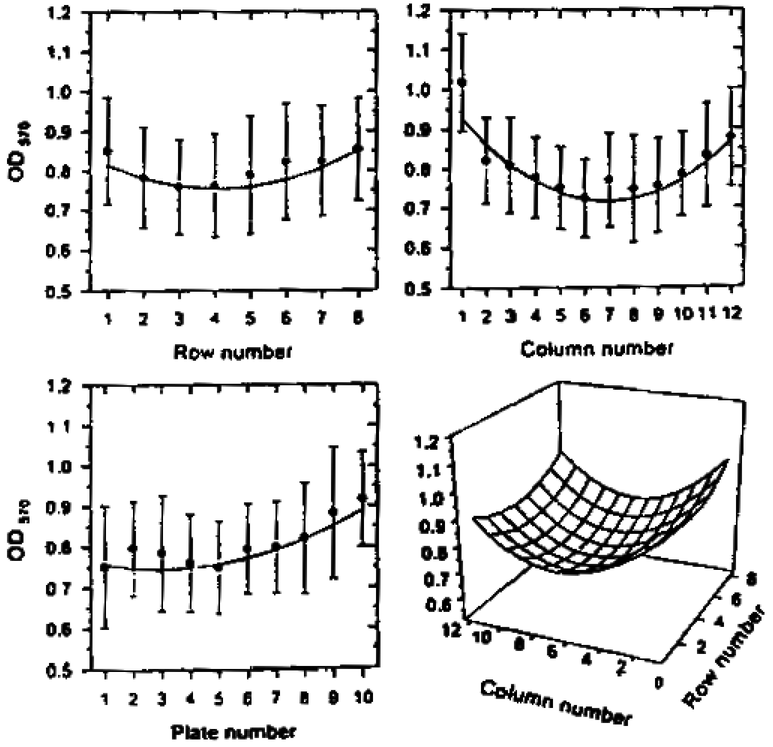
- Randomisation ensures that the ***effect of uncontrolled sources of variability*** has equal probability in all treatment groups
- Randomisation provides a **rigorous** method to assess the **uncertainty** (standard error) in treatment comparisons
- Formal randomisation involves a **chance dependent mechanism**, e.g. flip of a coin, by computer (Appendix D)
- Formal randomisation is ***not haphazard*** allocation
- Moment of randomisation such that the **randomisation covers all substantial sources of variation**,

  i.e. immediately before treatment administration
- Additional randomisation may be required at different stages

"to omit randomisation because one cannot see clearly how bias could occur is like trusting that glassware in chemistry is clean because it does not look dirty" (Mainland 1954)

Randomisation turns lethal bias into noise
Plate location effect in 96-well plates

Bias due to plate location effects in microtiter plates (Burrows (1984), causes parabolic patterns:



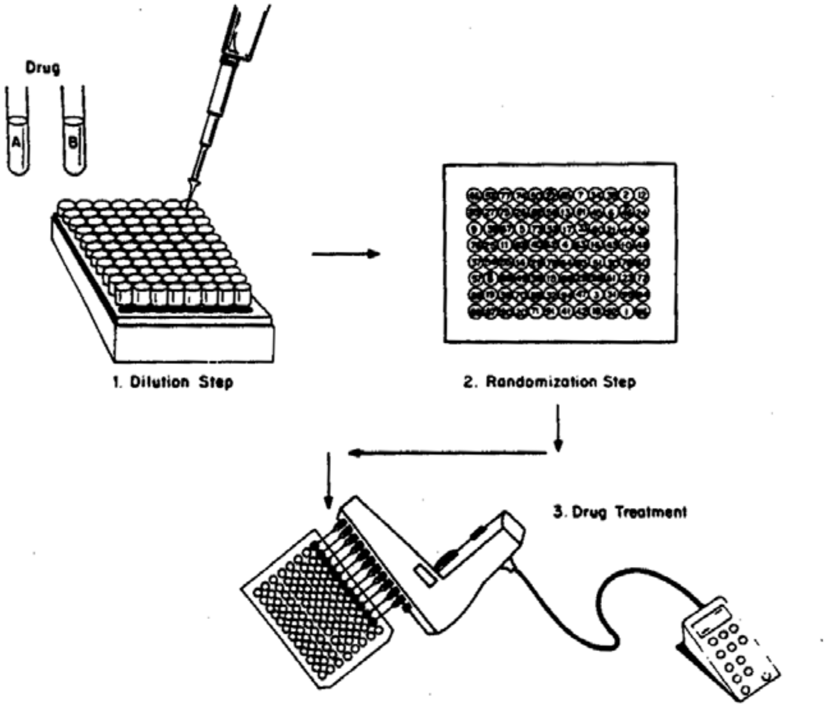
Underlying causes unkown

Solution by Faessel (1999):

- **Random** allocation of treatments to the wells
- Bias of plate location is now introduced as random variation

Randomisation can be used to transform lethal bias into noise

A convenient way to randomise microtiter plates



- dilutions made in tube rack 1
- tubes are numbered
- MS Excel used to make randomisation map
- map taped on top of rack 2
- tubes from rack 1 are pushed to corresponding number in rack 2
- multipipette used to apply drugs to 96-well plate
- results entered in MS Excel and sorted

Alternative methods:

- design (Burrows et al. 1984, Schlain et al. 2001)
- statistical model (Schlain et al. 2001, Straetemans et al. 2005)

Minimising Bias

RANDOMISATION

Randomised versus Systematic Allocation

- First unit treatment A, second treatment B, etc. yielding sequence AB, AB, AB
- Other sequences as AB BA, BA AB, etc. possible

Random Allocation

R.A. Fisher

⟺

Systematic Allocation

W.S. Gosset

Minimising Bias RANDOMISATION

Randomised versus Systematic Allocation

- First unit treatment A, second treatment B, etc. yielding sequence AB, AB, AB
- Other sequences as AB BA, BA AB, etc. possible

Random Allocation

R.A. Fisher

⟺

Systematic Allocation

W.S. Gosset

Minimising Bias

RANDOMISATION

Randomised versus Systematic Allocation

- First unit treatment A, second treatment B, etc. yielding sequence AB, AB, AB
- Other sequences as AB BA, BA AB, etc. possible

.

Random Allocation

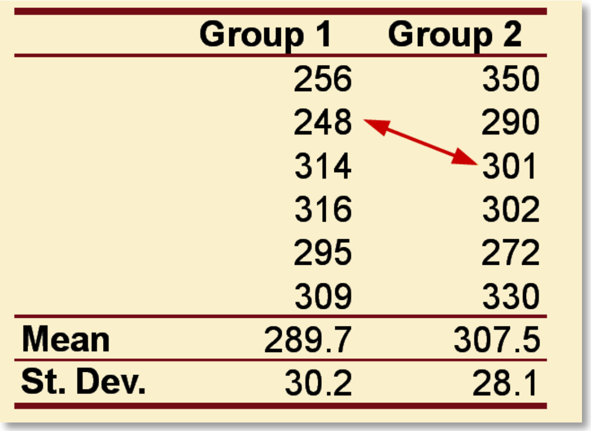
R.A. Fisher

- any systematic arrangement might coincide with a specific pattern in the variability, yielding a biased estimate of the treatment effect
- randomisation provides unbiased estimate of the error
- randomisation is a necessary condition for a valid statistical analysis

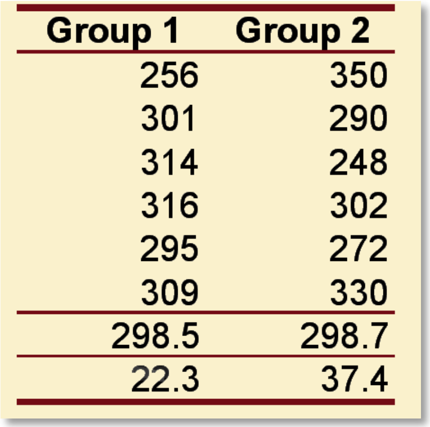
Minimising Bias

RANDOMISATION

Improvements of Randomisation Schemes



	Group 1	Group 2
	256	350
	248	290
	314	301
	316	302
	295	272
	309	330
Mean	289.7	307.5
St. Dev.	30.2	28.1



Group 1	Group 2
256	350
301	290
314	248
316	302
295	272
309	330
298.5	298.7
22.3	37.4

- Balancing means causes imbalance in variability
- Invalidates subsequent statistical analysis

Minimising Bias RANDOMISATION

Haphazard Allocation versus Formal Randomisation

- Effect of 2 diets on body weight of rats
- 12 animals arrive in 1 cage
- technician takes animals out of cage
- first 6 animals ➔diet A, remaining 6 ➔diet B

➔ Isn’t this *Random* ?

Minimising Bias RANDOMISATION

Haphazard Allocation versus Formal Randomisation

- Effect of 2 diets on body weight of rats
- 12 animals arrive in 1 cage
- technician takes animals out of cage
- first 6 animals ➔diet A, remaining 6 ➔diet B

Isn’t this *Random* ?

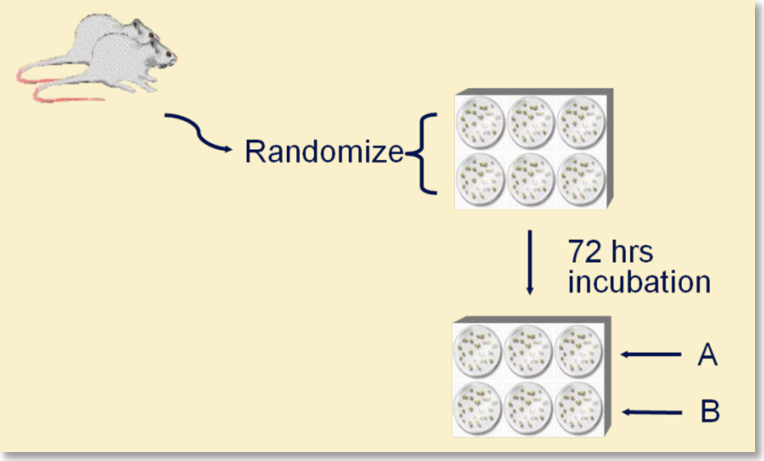
- heavy animals react slower and are easier to catch than smaller animals
- first 6 animals weigh more than remaining 6

- ➢ Haphazard treatment allocation is **NOT** the same as formal randomisation
- ➢ Haphazard allocation is **subjective** and can introduce bias
- ➢ Proper randomisation requires a physical “randomising device”

Minimising Bias

RANDOMISATION

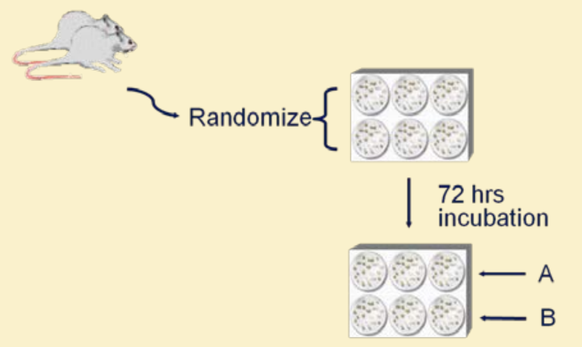
Moment of Randomisation



- Rat brain cells harvested and seeded on Petri-dishes (1 animal = 1 dish)
- Dishes randomly divided in two groups
- Two sets of Petri dishes placed in incubator
- After 72 hours group 1 treated with drug A and group 2 with drug B

Minimising Bias RANDOMISATION

Moment of Randomisation



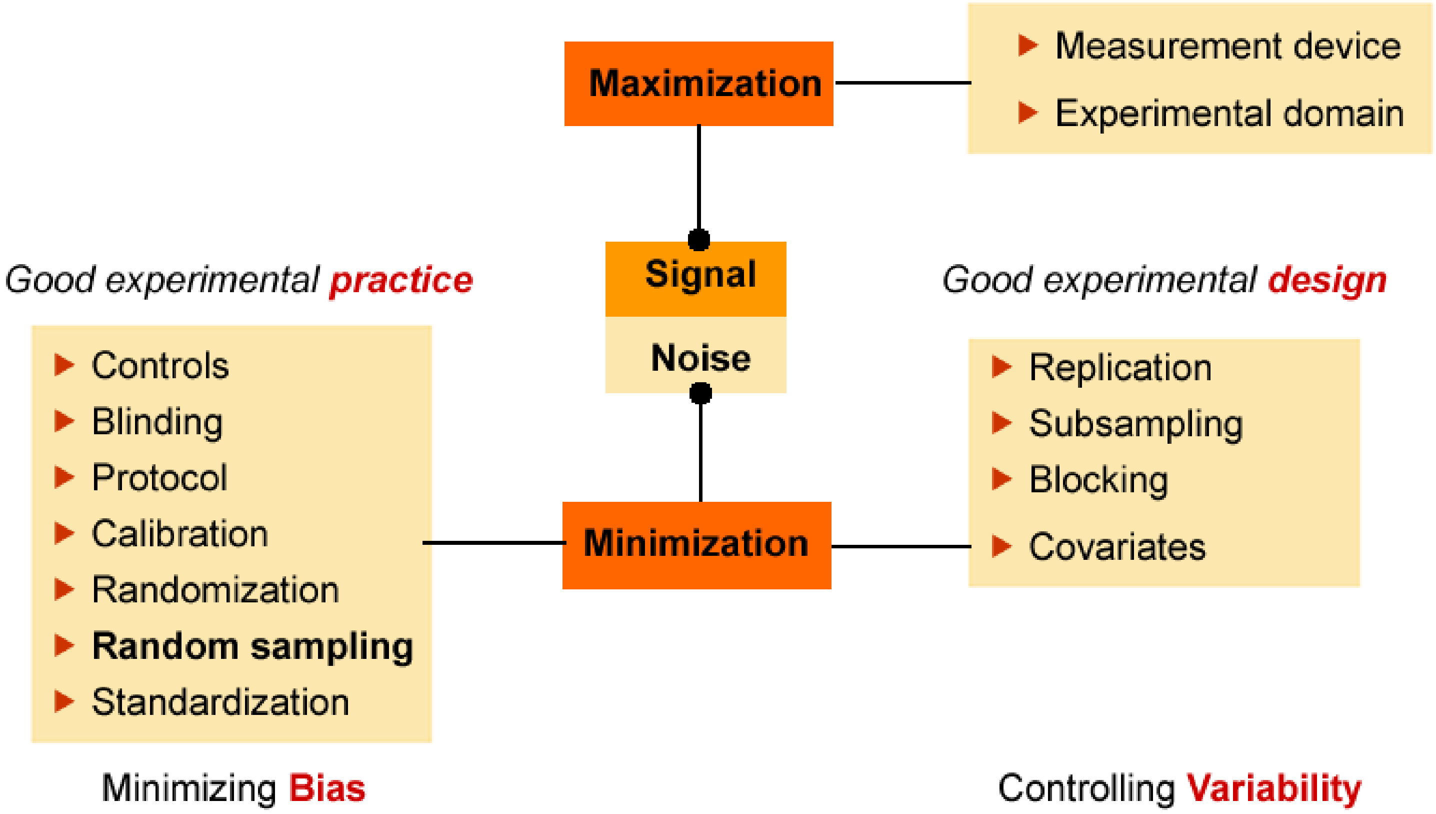
- Rat brain cells harvested and seeded on Petri-dishes (1 animal = 1 dish)
- Dishes randomly divided in two groups
- Two sets of Petri dishes placed in incubator
- After 72 hours group 1 treated with drug A and group 2 with drug B

- ➢ Effect of treatment confounded with systematic errors due to incubation
- ➢ Randomisation should be done as late as possible just before treatment application
- ➢ Randomisation sequence should be maintained throughout the experiment

Randomisation should apply to each stage of the experiment (Fry, 2014):

- ➢ allocation of independent experimental units to treatment groups
- ➢ order of exposure to test alteration within an environment
- ➢ order of measurement

Basic strategies for maximising Signal/Noise ratio



Maximization

- Measurement device
- Experimental domain

Signal

Noise

Minimization

*Good experimental **practice***

- Controls
- Blinding
- Protocol
- Calibration
- Randomization
- **Random sampling**
- Standardization

Minimizing **Bias**

*Good experimental **design***

- Replication
- Subsampling
- Blocking
- Covariates

Controlling **Variability**

Minimising Bias

RANDOM SAMPLING

Simple random sample
a selection of units from a defined target population such that each unit has an equal chance of being selected

- Neutralises **sampling bias**
- Increases **External Validity**
- Provides the foundation for the **population model of inference**
- Particularly important in **genetic research** (Nahon & Shoemaker)

A random sample is a stringent requirement which is very difficult to fulfill in biomedical research

→

Switch to **randomisation model** of statistical inference in which inference is restricted to the actual experiment only

Minimising Bias

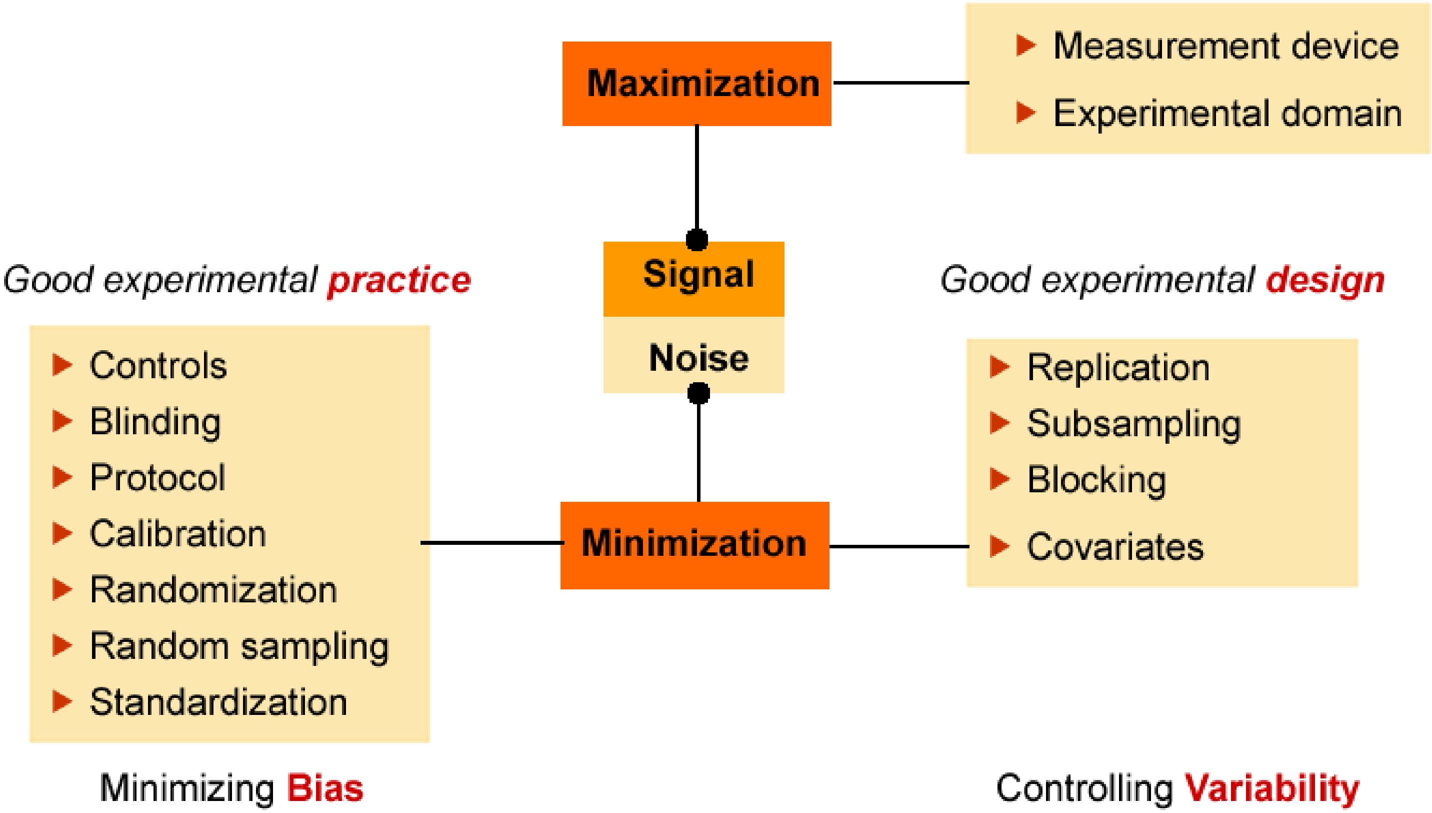
STANDARDISATION

Reducing the **intrinsic variability and bias** by:

- use of genetically uniform animals
- use of phenotypically uniform animals
- environmental control
- nutritional control
- acclimatization
- measurement system

❗ Beware of **external validity**

Basic strategies for maximising Signal/Noise ratio



Maximization

- Measurement device
- Experimental domain

Signal

Noise

Minimization

*Good experimental **practice***

- Controls
- Blinding
- Protocol
- Calibration
- Randomization
- Random sampling
- Standardization

Minimizing **Bias**

*Good experimental **design***

- Replication
- Subsampling
- Blocking
- Covariates

Controlling **Variability**

Controlling Variability REPLICATION

Replication can be an **effective** strategy to control variability (precision)

Replication is an **expensive** strategy to control variability

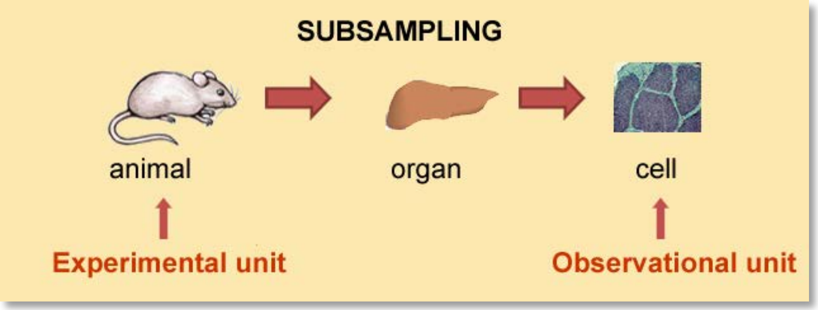
Precision of experiment is quantified by standard error of difference between two treatment means

$$SD \times \sqrt{2/no.\ exp.\, units\ per\ treatment}$$

Replication is also needed to estimate SD

Controlling Variability

SUBSAMPLING



n experimental units, standard deviation σ_n
m subsamples, standard deviation of subsamples σ_m

$$SD = \sqrt{\sigma_n^2 + \frac{\sigma_m^2}{m}}$$

Standard error of difference: $\sqrt{\frac{2}{n}\left(\sigma_n^2 + \frac{\sigma_m^2}{m}\right)}$

subsampling:

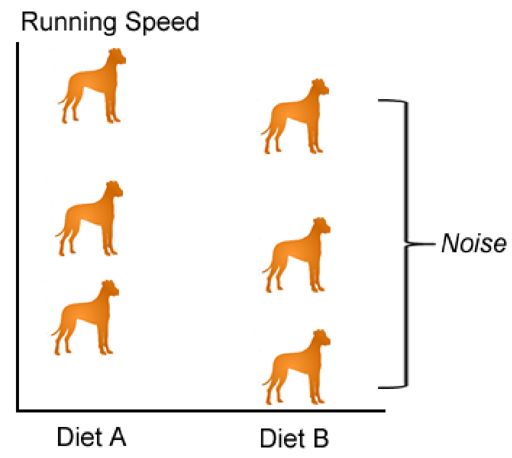
- multiple observations over time
- duplicate, triplicate measurement
- animals in a cage (cage is EU)
- different levels of subsampling

- Replication is in most cases only effective at the level of the **true experimental unit**
- Replication at the subsample level (observational unit) makes only sense when the variability at this level is substantial as compared to the variability between the experimental units

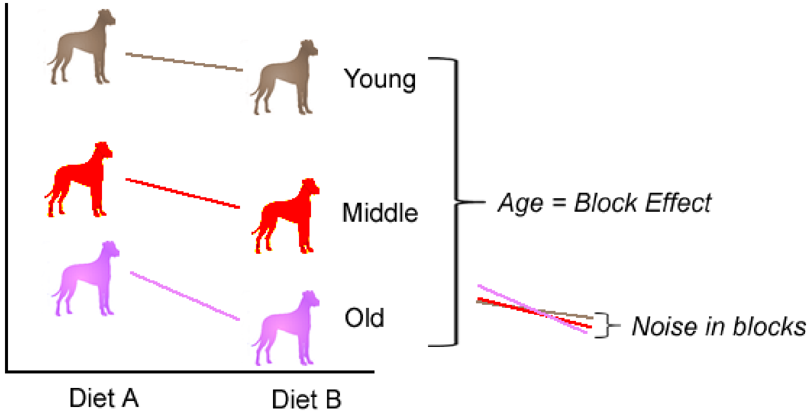
Controlling Variability

BLOCKING

A **blocking factor** is a known and unavoidable **source of variability** that adds or subtracts an offset to every experimental unit in the block



Total Variability = Treatment Effect + Noise



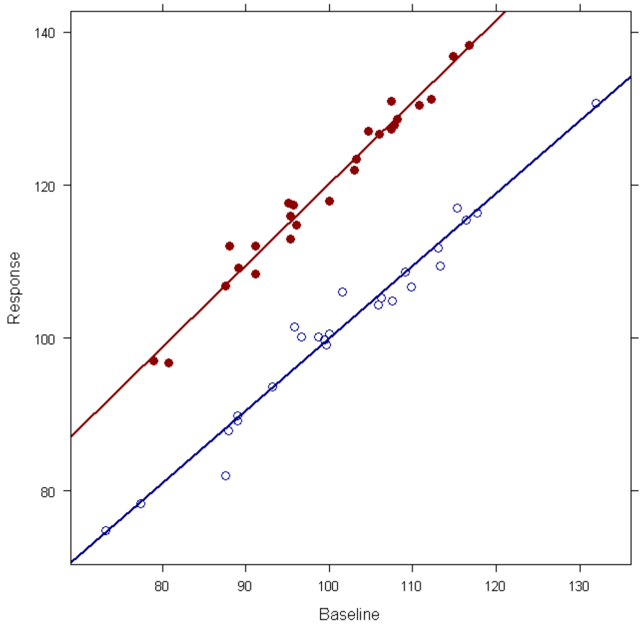
Total Variability = Treatment Effect + Block Effect + Within block Noise

Typical blocking factors

- age group
- microtiter plate
- run of assay
- batch of material
- subject (animal)
- date
- operator
- laboratory

Controlling Variability COVARIATES

A **covariate** is an uncontrollable but measurable attribute of the experimental unit or its environment that is **unaffected** by the treatments but may have an influence on the measured response



Response = Treatment effect + β.Covariate effect + Error

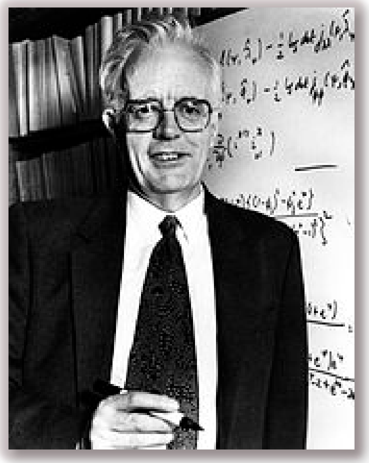
Typical covariates

- baseline measurement
- weight
- age
- ambient temperature

Covariates *filter out* one particular source of variation. Rather than blocking it represents a quantifiable (by coefficient β) attribute of the system studied.

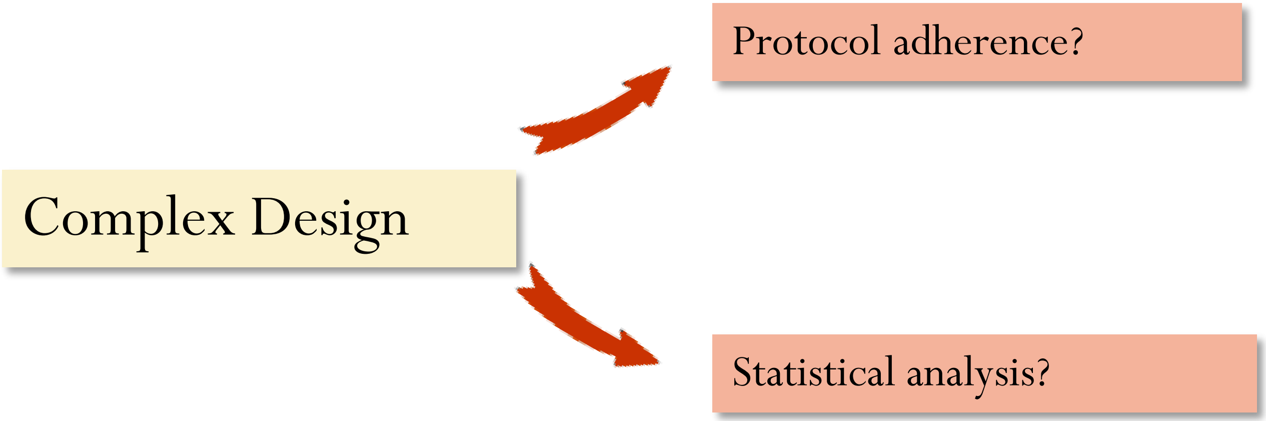
Covariate adjustment requires **slopes to be parallel**

Requirements for a good experiment



- Treatment comparisons free of systematic error
- Comparisons sufficiently precise (high S/N ratio)
- Wide range of external validity
- Experimental setup as simple as possible
- Uncertainty (error) must be assessable

Simplicity of design



Complex Design

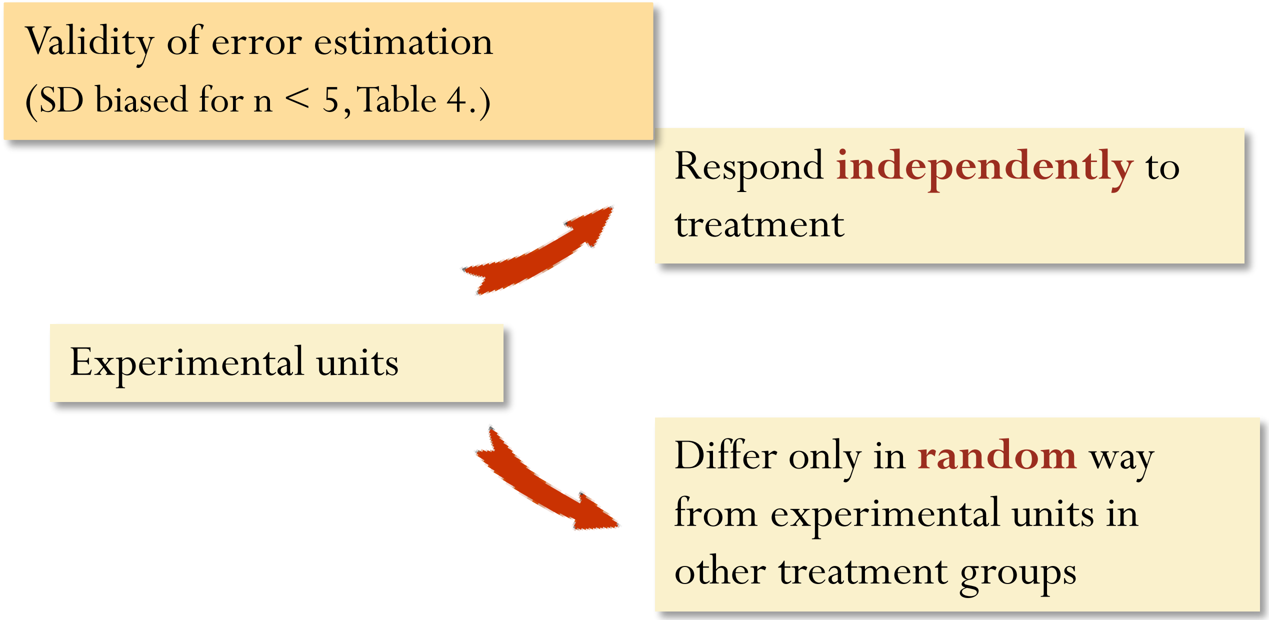
Protocol adherence?

Statistical analysis?

Simple designs require a simple analysis

The calculation of uncertainty (error)

"It is possible and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available" (Fisher, 1935)



Validity of error estimation
(SD biased for n < 5, Table 4.)

Experimental units

Respond **independently** to treatment

Differ only in **random** way from experimental units in other treatment groups

Without a valid estimate of error, a valid statistical analysis is impossible

The Smart Design of Animal Experiments

V. Common Designs in Biological Experimentation

The jungle of experimental designs

Latin Square Design

Split Plot

Completely Randomized Design

Greco-Latin Square Design

Randomized Complete Block Design

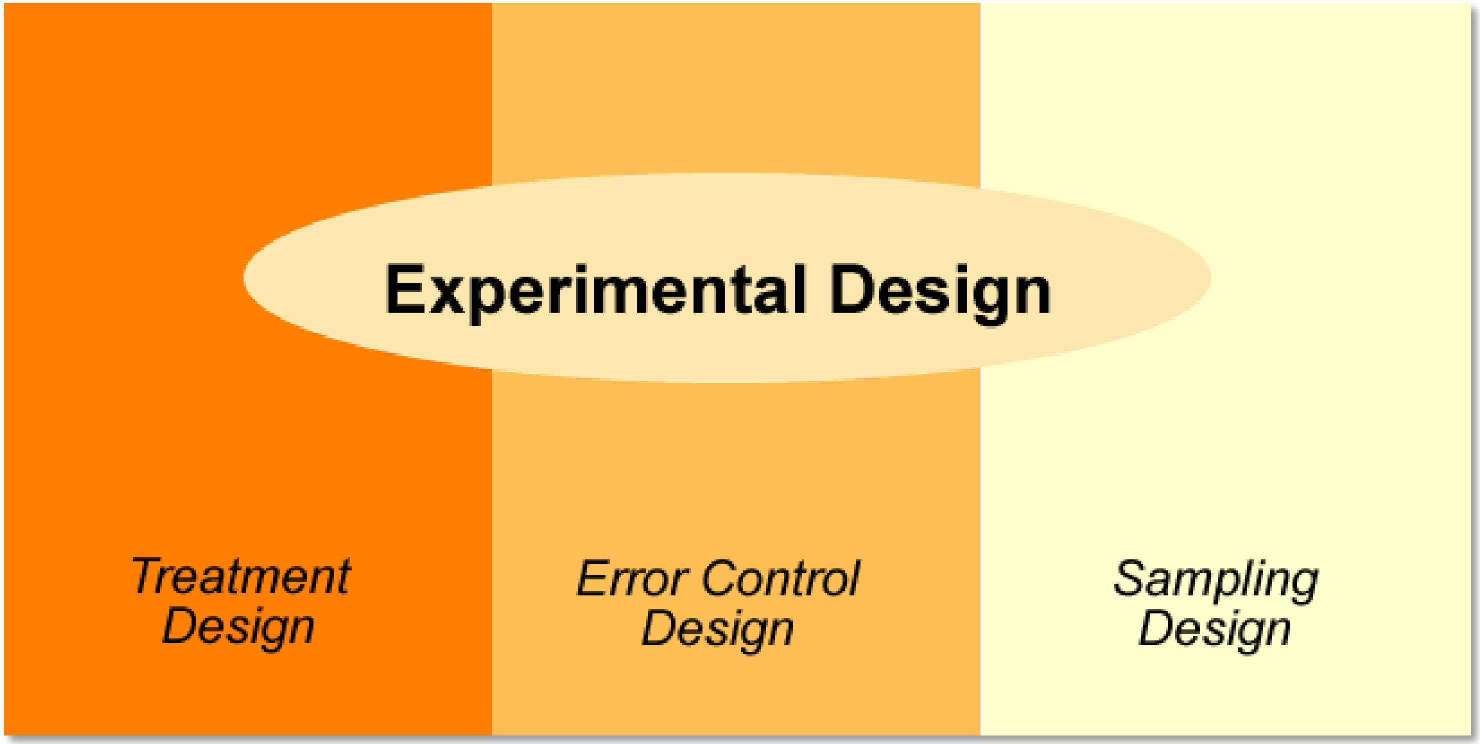
Youden Square Design

Plackett-Burman Design

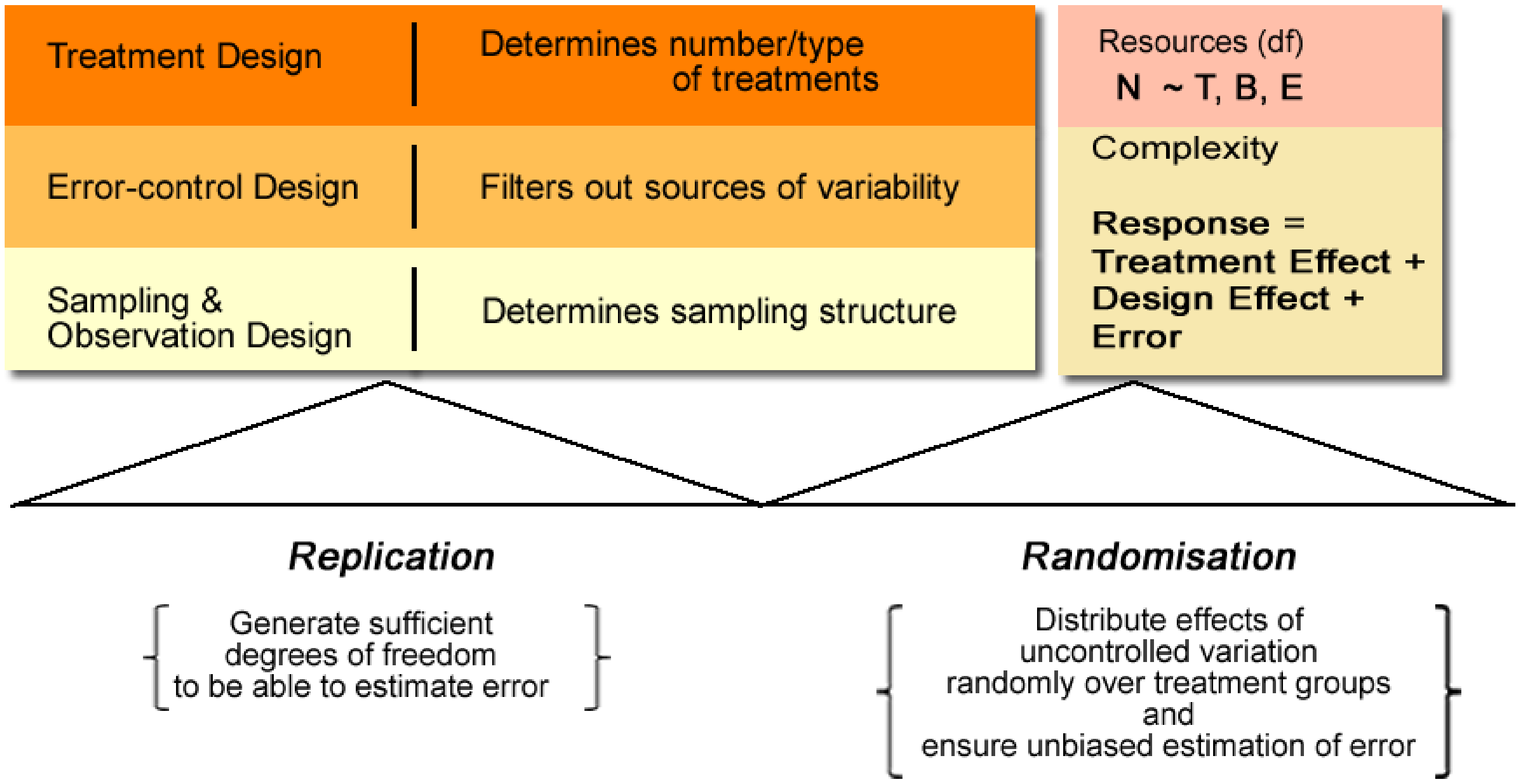
Factorial Design

Experimental design is an integrative concept

An experimental design is a synthetic approach to minimise bias and control variability



Three aspects of experimental design determine complexity and resources



Treatment Design	Determines number/type of treatments
Error-control Design	Filters out sources of variability
Sampling & Observation Design	Determines sampling structure

Resources (df)

N ~ T, B, E

Complexity

Response = Treatment Effect + Design Effect + Error

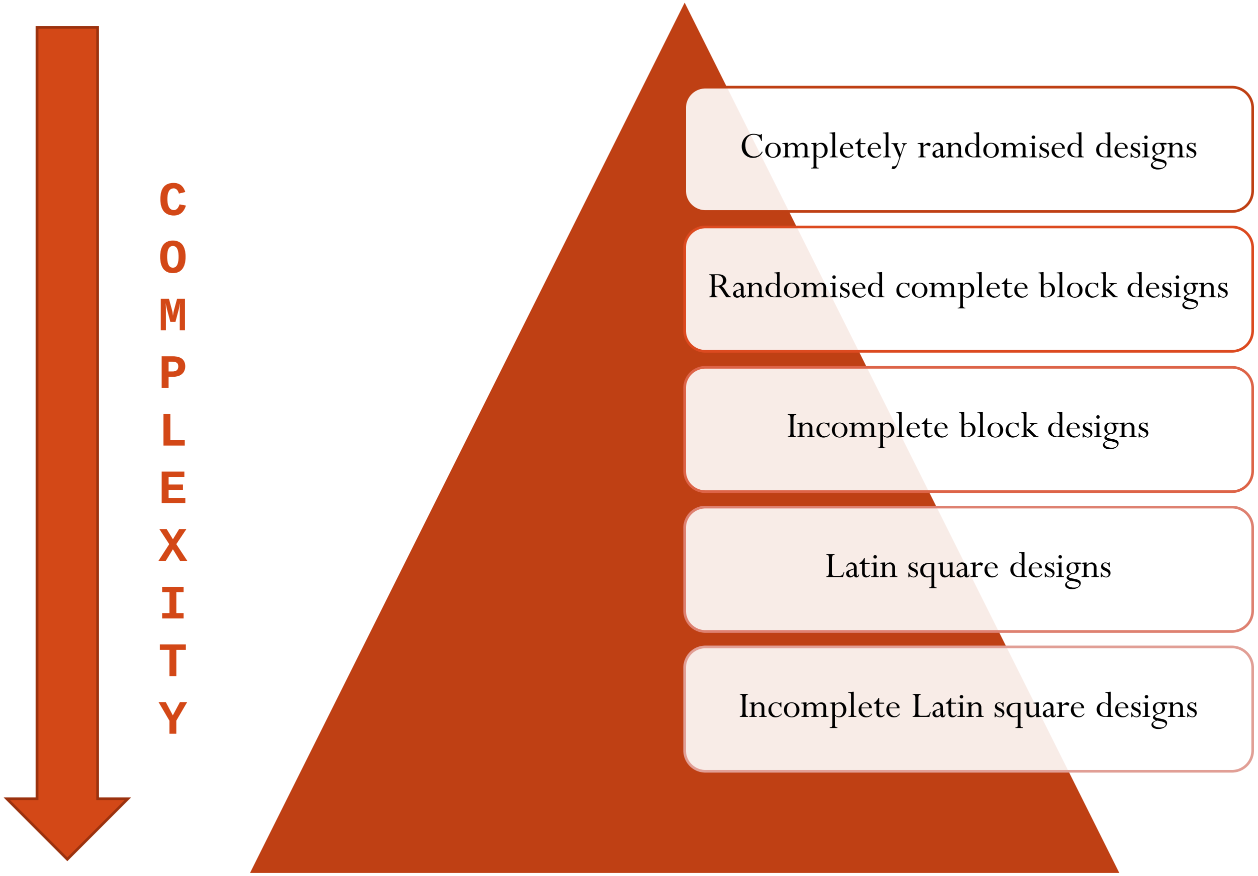
Replication

Generate sufficient degrees of freedom to be able to estimate error

Randomisation

Distribute effects of uncontrolled variation randomly over treatment groups and ensure unbiased estimation of error

Error-control designs



C
O
M
P
L
E
X
I
T
Y

Completely randomised designs

Randomised complete block designs

Incomplete block designs

Latin square designs

Incomplete Latin square designs

Error-control designs
Completely Randomised Design

- Completely randomised **error–control design**
- Random assignment of experimental units to treatment conditions
- Most common design (default design), simple, easy to implement

Lack of precision in comparisons is major drawback

Completely Randomised Design Example 1

- Effect of treatment on proliferation of gastric epithelial cells in rats
- 2 drugs & 2 controls (vehicle) = 4 experimental groups
- 10 animals/group, individual housing in cages
- Cages numbered and treatments assigned randomly to cage numbers
- Cages sequentially put in rack
- Blinding of laboratory staff and of histological evaluation,
- Treatment code = cage number (individual blinding)

Error-control designs
Randomised Complete Block Design

- ➢ Treatment conditions are compared in the presence of a single **isolated** extraneous source of variability (blocks)
- ➢ Random assignment of experimental units **within** blocks to treatment conditions
- ➢ One way block **error-control design**

- Increase of precision
- Eliminates possible imbalance in blocking factor

Complete:
Each block contains all treatments

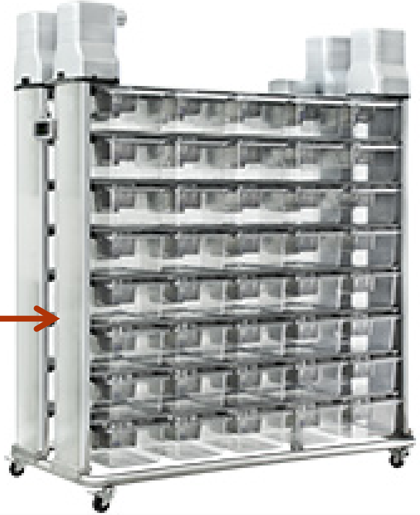
- Assumes treatment effect the same among all blocks
- Blocking characteristic must be related to response or else loss of efficiency due to loss in df

Randomised Complete Block Design Example

- Effect of treatment on proliferation of gastric epithelial cells in rats
- 2 drugs & 2 controls (vehicle) = 4 experimental groups
- 10 animals/group, individual housing in cages
- Cages numbered and treatments assigned randomly to cage numbers
- Blinding of laboratory staff

Shelf height will probably influence the results

- 1 shelf = 1 block
- 8 animals (cages) / shelf, 5 shelves
- Randomize treatments per shelf
- 2 animals / shelf / treatment



Not restricted to a single factor such as shelf height, other characteristics can also be included (e.g. body weight)

Randomised Complete Block Design Example – Aggregate Blocking Variable

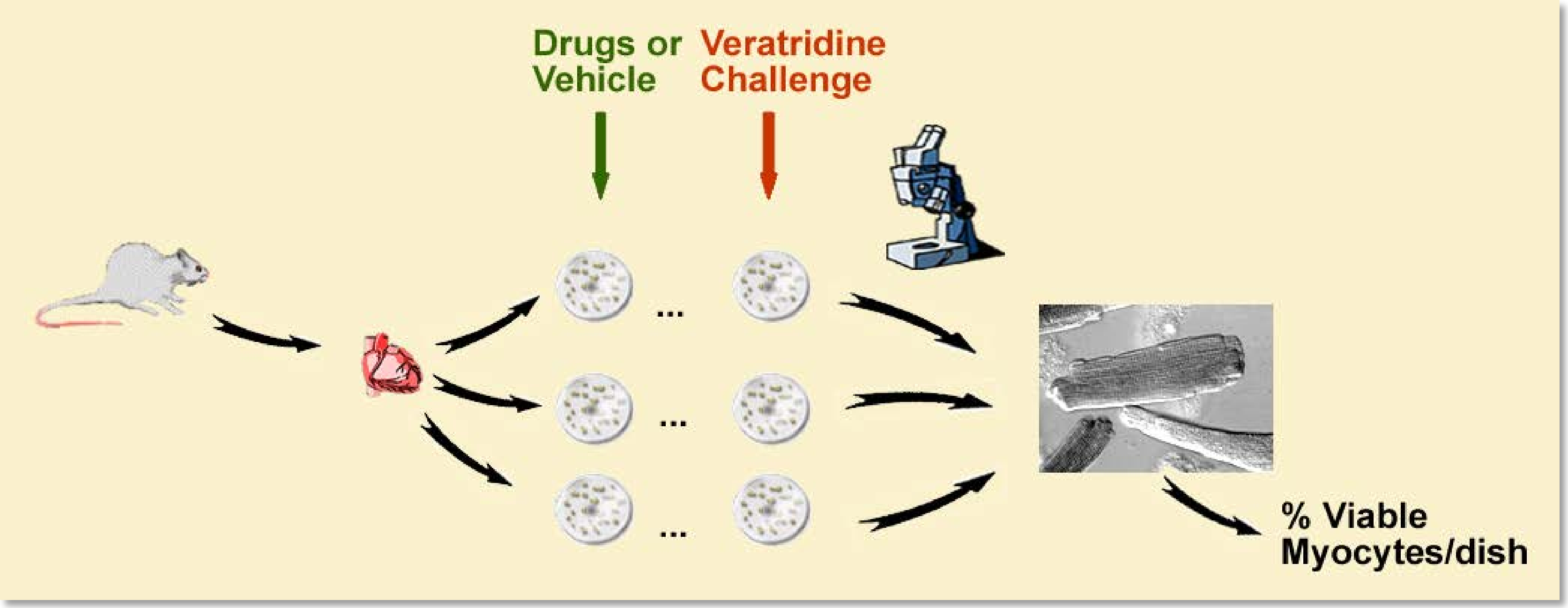
Shelf height and rat body weight will probably influence the results

- Order animals according to body weight
- Put first 8 rats on top row, next 8 on row 2, etc.
- 1 shelf = 1 block
- 8 animals (cages) / shelf, 5 shelves
- Randomize treatments per shelf
- 2 animals / shelf / treatment

Design controls for body weight and shelf height

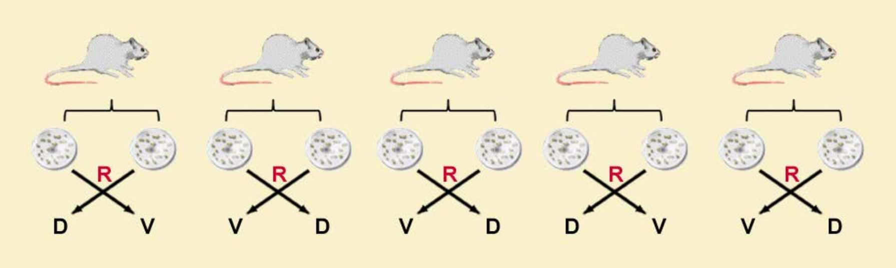
Randomised Complete Block Design
Paired Design – Example 1

- Special case of RCB design with only 2 treatments and 2 EU/block
- Example Cardiomyocyte experiment

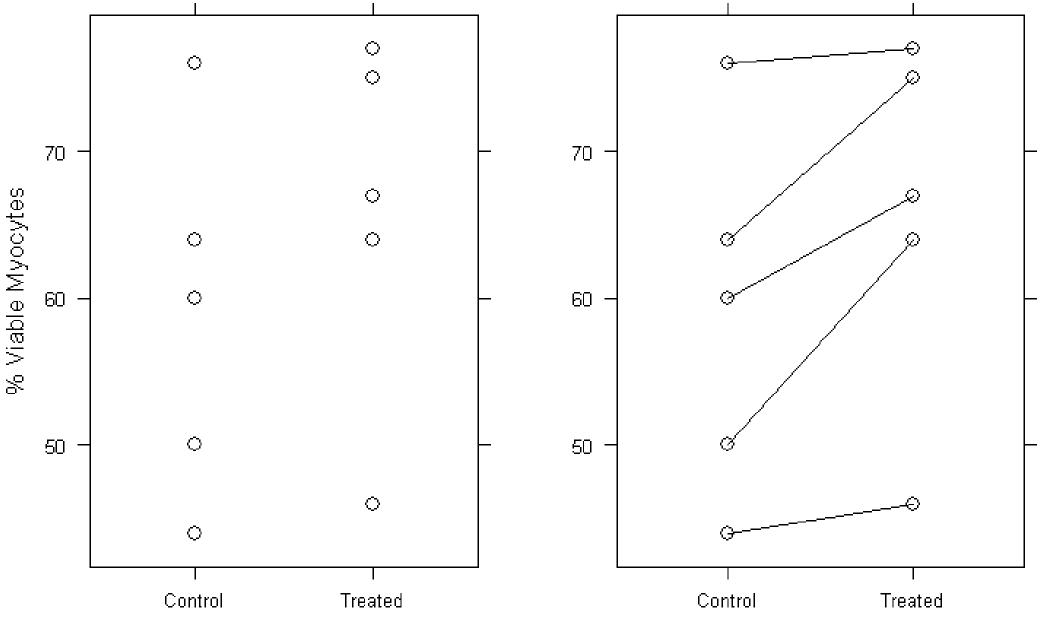


Randomised Complete Block Design
Paired Design – Example 1

Experimental Design



Results



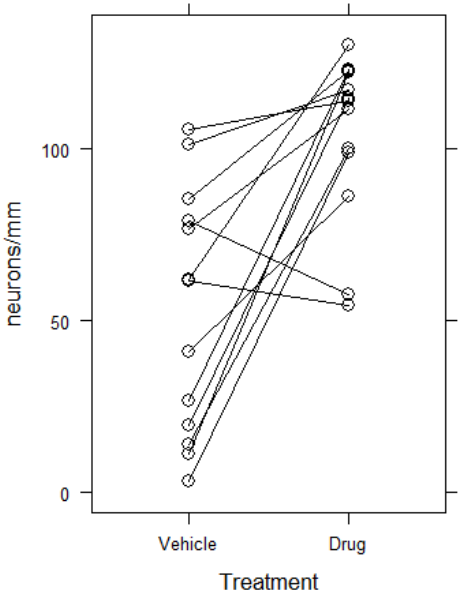
- Left panel
  - ignores pairing
  - groups overlap,
  - standard error drug-vehicle = 7.83
- Right panel
  - pairs are connected by lines
  - consistent increase in drug treated dishes
  - standard error drug-vehicle = 2.51

The CRD requires 8 times more EU's then the paired design

Randomised Complete Block Design
Paired Design – Example 2

Neuronal Protection in Rats

Results



Experimental Design
Pairing by date of experiment

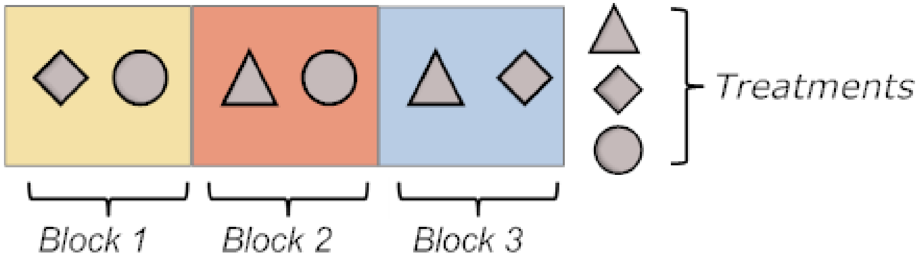
Day	Animal 1	Animal 2
1	Drug	Vehicle
2	Drug	Vehicle
3	Vehicle	Drug
…	….	….
12	Vehicle	Drug
13	Drug	Vehicle

- Pairing criterium not successful
- Mean difference 51.2 neurons/mm
- Standard error drug-vehicle = **12.3**
- As CRD, standard error = **12.0**
- Degrees of freedom, paired = 12
- Degrees of freedom, CRD = 24

Blocking is only effective when the within-block variation is much less than the between-block variation. Otherwise, a paired design is less efficient by loss of degrees of freedom

Error-control designs
Incomplete Complete Block Design

- Block size smaller than number of treatments
- Not all treatments present in each block
- One way block **error-control design**



Include comparisons of specific interest in each of the blocks, e.g. comparison with control

Balanced incomplete block (BIB) designs allow all pairwise comparisons with equal precision

- each block same number of EU's
- each treatment same number of times in design
- every pair of treatments occurs together in same number of blocks

R-package *agricolae*

BIB's exist only for certain combinations of number of treatment and block size

Balanced incomplete block design
Example 1

- assess effect of Vitamin A and protein supplement on weight gain of lambs
- 4 treatment conditions *A*, *B*, *C*, and *D*
- 3 replicates/treatment condition
- 6 pairs of sibling lambs

Sibling lamb pair	**First lamb**	**Second lamb**
1	A	B
2	A	C
3	A	D
4	B	C
5	B	D
6	C	D

Sometimes we have to add or omit a treatment to find a suitable BIB-design

Balanced incomplete block design

-package *agricolae*

```
> library(agricolae) # load library
> trt<-LETTERS[1:4]  #trt contains 4 treatments labelled A, B, C, D
> design.bib(trt,2,seed=543)$sketch  > # Blocksize = 2, Change seed for other randomization
Parameters BIB
===============
Lambda : 1
treatmeans : 4
Block size : 2
Blocks : 6
Replication: 3
Efficiency factor 0.6666667
<<< Book >>>
[,1] [,2]
[1,] "D" "A"
[2,] "B" "D"
[3,] "A" "B"
[4,] "C" "B"
[5,] "A" "C"
[6,] "C" "D"
```

Balanced incomplete block design
Example 2 Biggers et al. (1981)

- Effects of intrauterine injections of 6 prostaglandin antagonists on fertility of mice
- Uterine horns not connected
- 2 treatments/female
- All possible pairwise comparisons: $\binom{6}{2}$= 15 females
- 5 replicates per treatment

Assumes effect of a treatment in one uterine horn is local and has no effect on the contralateral horn

```
> # Label 6 treatments A, B, C, D, E, F
> trt<-LETTERS[1:6]
> # Blocksize = 2, 2 uterine horns
> # Change seed for other randomization
> design.bib(trt,2,seed=4338)$sketch
Parameters BIB
===============
Lambda : 1
treatmeans : 6
Block size : 2
Blocks : 15
Replication: 5
Efficiency factor 0.6
<<< Book >>>
[,1] [,2]
[1,] "D" "B"
[2,] "B" "C"
[3,] "E" "B“
.....
```

Balanced incomplete block design
Final Remarks

- BIB designs not always possible
- Most cases basic design has to be repeated (use different seeds, or use argument **r**)
- See Dean & Voss for more details on how to compute required number of replicates

A BIB design comparing 6 treatments with 6 repeats each does not exist:

```
# Label 6 treatments A, B, C, D, E, F
> trt<-LETTERS[1:6]
> # Block size 2,
> # 6 replicates of each treatment
> out<-design.bib(trt,2,6,seed=4338)$sketch
```

Change r by 5, 10, 15, 20 …

Let's do what the program asks us and take 10 repeats per treatment:

```
> out<-design.bib(trt,2,10,seed=4338)$sketch
Parameters BIB
==============
Lambda : 2
treatmeans : 6
Block size : 2
Blocks : 30
Replication: 10
Efficiency factor 0.6
<<< Book >>>
```

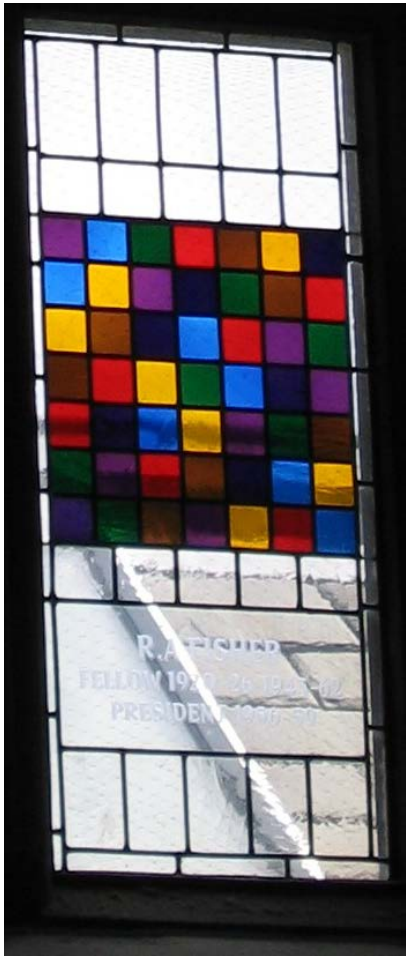
Error-control designs
Latin Square Design

- Compares treatment conditions in presence of **two** isolated, extraneous sources of variability (blocks)
- Two-way block **error-control design**

- In 1 *k x k* Latin square only *k* EU's/treatment
- More EU's by stacking Latin squares upon each other or next to each other
- **Randomising the rows (columns) of the stacked *LS* design can be advantageous**
- Use R-packages ***magic*** or ***agricolae***

Latin squares to eliminate row-column effects in 96-well microtiter plates, 1 plate = 6 x (4 x 4 LS)
also of interest for experiments in green houses and growth chambers

Latin Square Design

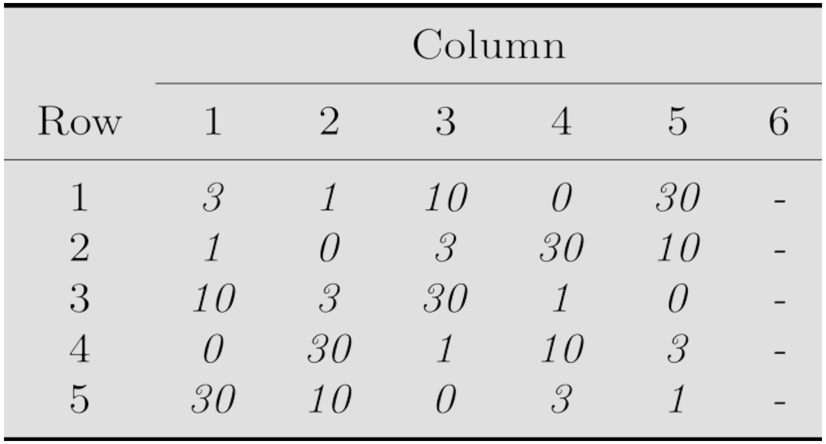


Latin square designs were first introduced by R.A. Fisher in agricultural experimentation to as designs for blocking in two directions.

This stained glass window in the dining hall of Caius College, in Cambridge, commemorates Fisher for his contributions to experimental science.

Latin Square Design
Example – Weight Gain Study

- Weight gain study in female CD-1 mice (Gore and Stanley, 2005)
- Control & 4 doses (1, 3, 10 mg/kg) test compound
- Mice housed singly in cages across 3 racks
- Rack = 5 shelves of 6 cages (columns), column 6 left empty
- Independent LS /rack, each treatment group in each row and in each column of the rack



	Column					
Row	1	2	3	4	5	6
1	*3*	*1*	*10*	*0*	*30*	-
2	*1*	*0*	*3*	*30*	*10*	-
3	*10*	*3*	*30*	*1*	*0*	-
4	*0*	*30*	*1*	*10*	*3*	-
5	*30*	*10*	*0*	*3*	*1*	-

Rack influenced water intake, body temperature depended on shelf height

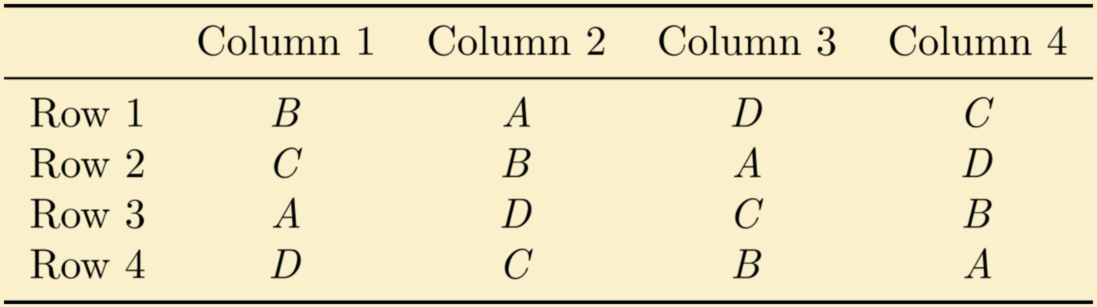
Latin Square Design
Generate Latin Squares

- R-package *agricolae*

```
> library(agricolae) # load package agricolae
> trt<-c("0","1","3","10","30") # 4 doses (1, 3, 10, 30) and control (= 0)
> # Latin square design
> # use seed for different randomization
> design.lsd(trt, seed=3489)$sketch
[,1] [,2] [,3] [,4] [,5]
[1,] "1" "30" "10" "3" "0"
[2,] "0" "10" "3" "1" "30"
[3,] "10" "1" "0" "30" "3"
[4,] "3" "0" "30" "10" "1"
[5,] "30" "3" "1" "0" "10"
```

Latin Square Design
Example 96-well plate

- *In order to minimize measurement error due to a spatial gradient in binding efficiency within plate, microplate wells were grouped into blocks of 16 (4 x 4) wells and each set of 3 samples (A,B,C) was placed in the block using a version of Latin square design along with standards (D)* (Aoki, et al. 2014).



	Column 1	Column 2	Column 3	Column 4
Row 1	B	A	D	C
Row 2	C	B	A	D
Row 3	A	D	C	B
Row 4	D	C	B	A

- 96-wells = 3 column-wise x 2 row-wise = 6 replicates of 4x4-Latin Squares
- *By placement patterns were changed across blocks* = randomisation of rows and blocks

Incomplete Latin squares

More treatments than *k* in
k x k Latin square

Balanced lattice squares

Youden squares

Incomplete Latin Square Designs
Balanced lattice Design

- Example Burrows, et al. (1984) 96-well plate, lattice square design

	1	2	3	4	5	6	7	8	9	10	11	12
A	11	7	3	15	11	10	12	9	15	5	12	2
B	12	8	4	16	6	7	5	8	9	3	14	8
C	9	5	1	13	1	4	2	3	4	10	7	13
D	10	6	2	14	16	13	15	14	6	16	1	11
E	12	14	7	1	5	11	4	14				
F	3	5	16	10	1	15	8	10				
G	13	11	2	8	9	7	16	2				
H	6	4	9	15	13	3	12	6				

- Number of treatments is full square (4, 9, 16, 25,…)
- 5 x 4x4-balanced lattice design, each treatment is tested once in each block
- Each pair of treatments occurs once for each column and each row

Incomplete Latin Square Designs
Youden squares

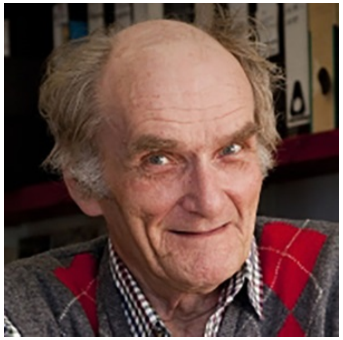
Youden (1937) r x c squares (rectangles):

- r = t
  c < r
- Every treatment in every column, not in every row

Example (Colquhoun, 1963):

- Gastrin assay in rats
- 2 doses standard preparation
- 2 doses unknown potency
- # treatment applications per animal limited to 3
- 4 x 3 Youden square

```
> library(agricolae) # load package
> trts<-c("A","B","C","D") # 4 doses of 2 drugs
> admins<-3 # administrations per animal
> outdesign <-design.youden(trts,admins,seed=3273)
> outdesign$sketch
[,1] [,2] [,3]
[1,] "A" "B" "C"
[2,] "D" "C" "B"
[3,] "B" "D" "A"
[4,] "C" "A" "D"
```



Prof. David Colquhoun
(2016, age 80)

Randomised block designs and laboratory animal experiments

Block Designs

- More powerful
- Less bias
- Higher external validity
- More repeatable results

Completely Randomised Design

- Less powerful
- More bias
- Limited external validity
- Less repeatable results

Festing (2014)

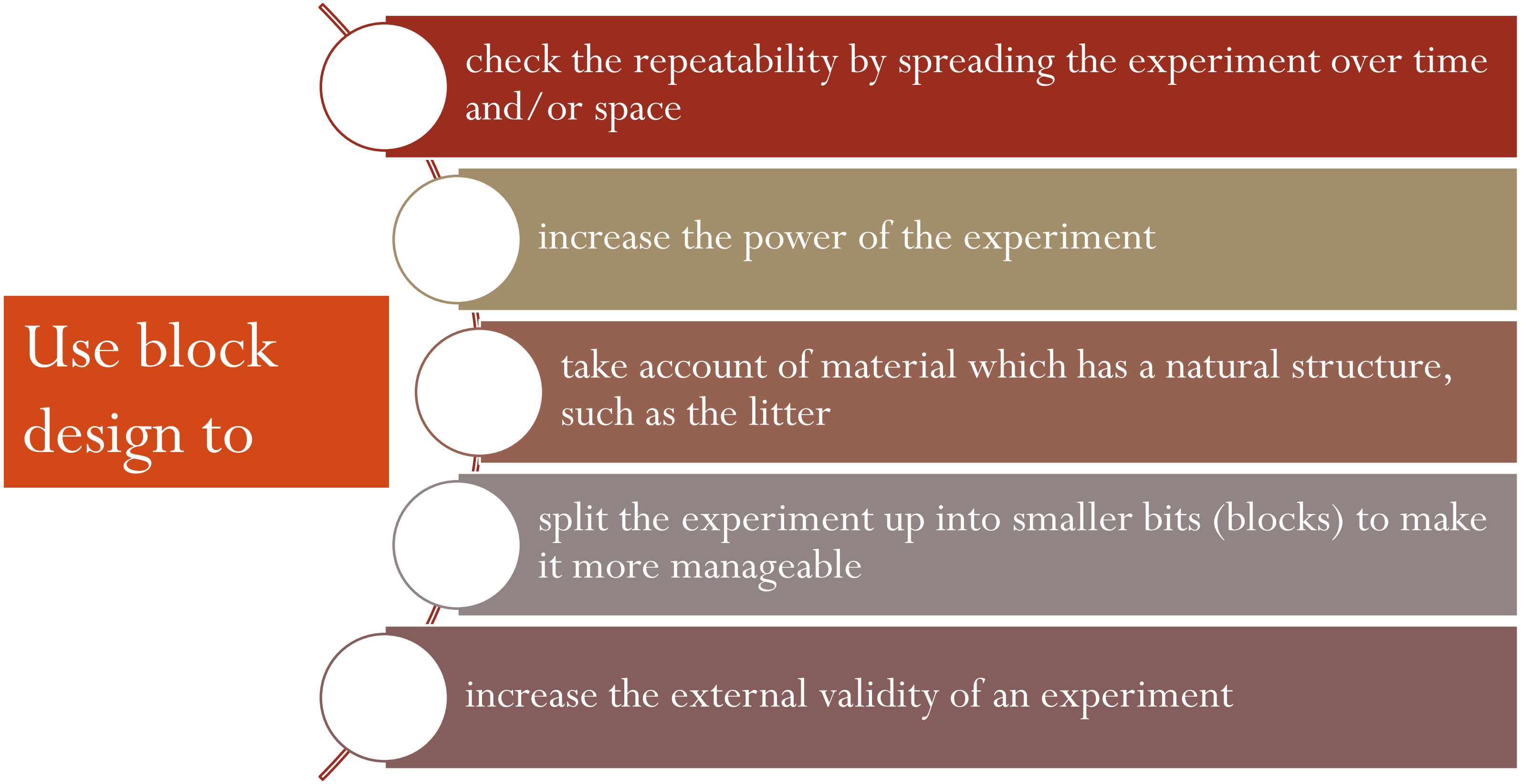
Randomised block designs and laboratory animal experiments

Block designs rarely used in animal experiment

- Waste of animals, money, and scientific resources
- Slowed down scientific progress

Festing (2014)

Use of block designs in laboratory animal experiments



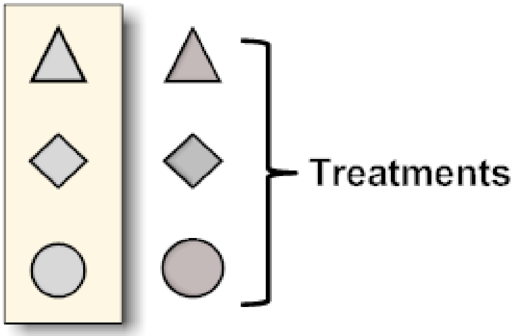
Use block design to

- check the repeatability by spreading the experiment over time and/or space
- increase the power of the experiment
- take account of material which has a natural structure, such as the litter
- split the experiment up into smaller bits (blocks) to make it more manageable
- increase the external validity of an experiment

Festing (2014)

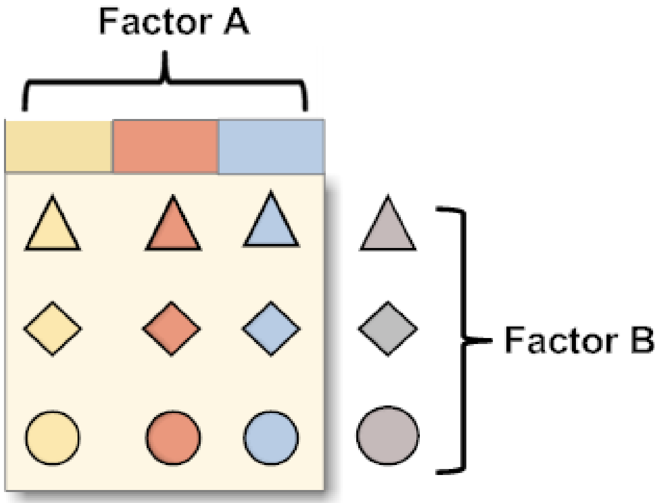
Treatment designs

One-way layout



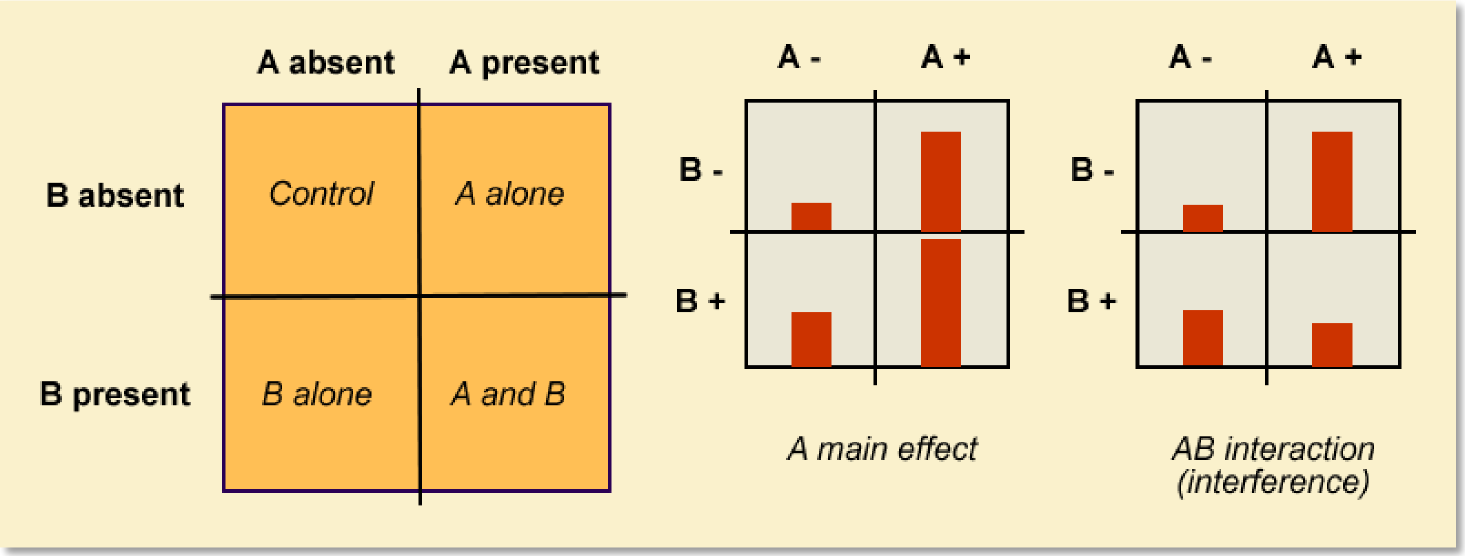
Studies the effect of a single factor

Factorial designs



Simultaneously studies the joint effect of several factors

Factorial Design (full factorial)
Main effects & Interactions



	A absent	A present
B absent	*Control*	*A alone*
B present	*B alone*	*A and B*

A main effect

AB interaction (interference)

Factorial design example

Effect of Western diet atherosclerotic lesions as compared to normal in two different strains of mice C3H apoE$^{-/-}$ and C57BL apoE$^{-/-}$

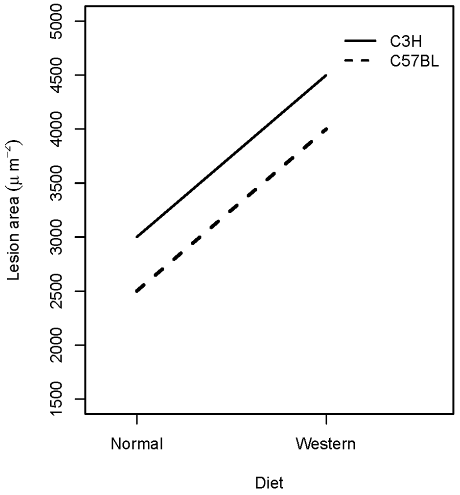
Factor levels (5 animals/group):

1. C3H apoE$^{-/-}$ + normal diet
2. C3H apoE$^{-/-}$ + Western diet
3. C57BL apoE$^{-/-}$ + normal diet
4. C57BL apoE$^{-/-}$ + Western diet

Results – No Interaction:

- difference between diets same for both strains
- overall effect of diet, irrespective of strain, average out over strain (4 x 5 animals)
- overall effect strain irrespective of diet, average out over diet (4 x 5 animals)

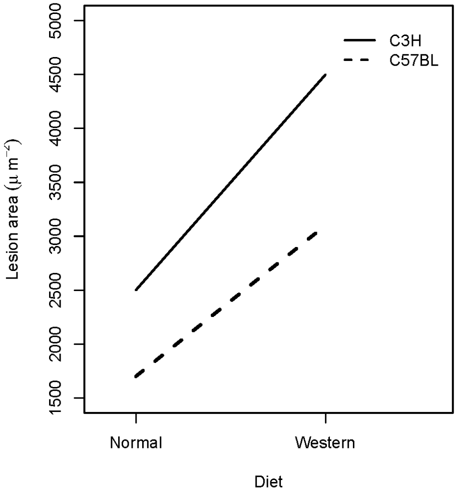
Factorial designs are highly efficient designs when there is no interaction.
All the experimental units are used to test simultaneously 2 hypotheses



Factorial design example (cont.)

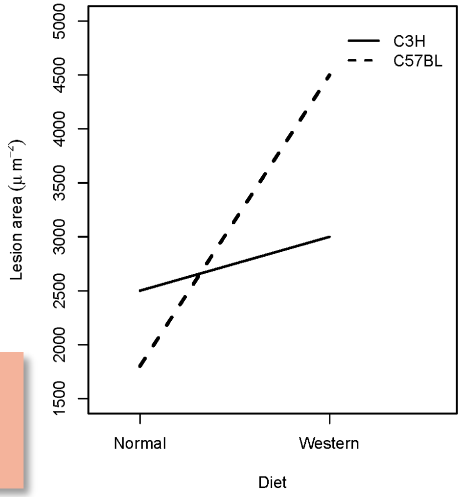
Results – Moderate Interaction:

- **Direction** of effect of diet is same regardless of the strain
- **Size** of effect of diet varies with strain
- C3H strain is more sensitive to diet



Results – Strong Interaction:

- Effect of diet depends entirely on the strain
- Feeding Western diet has almost no effect on C3H strain
- Effect of diet on C57BL is substantial



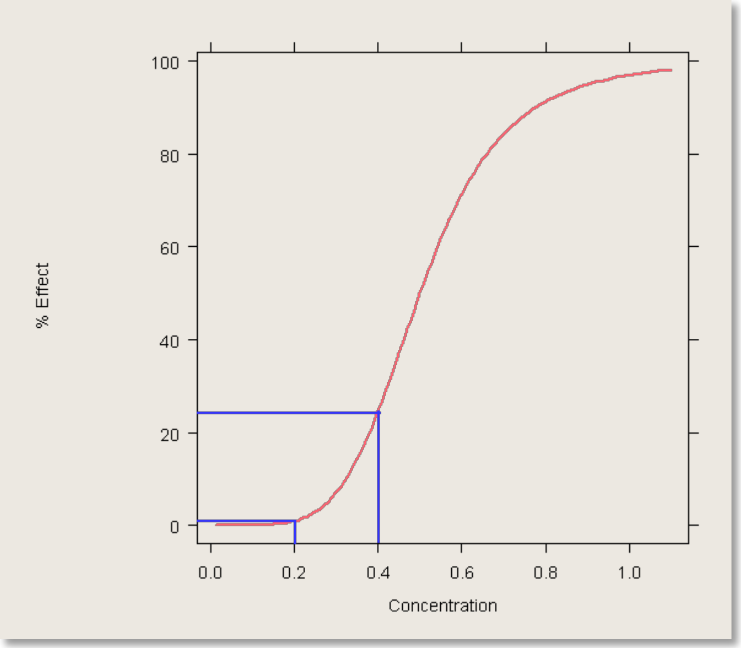
Factorial design is only and also most efficient way to study interaction of different factors

Factorial Designs
Interaction and Drug Synergy

A hypothetical factorial experiment of a drug combined with itself at 0 and 0.2 mg/l

	0	0.2 mg/L
0	0%	1%
0.2 mg/L	1%	22%

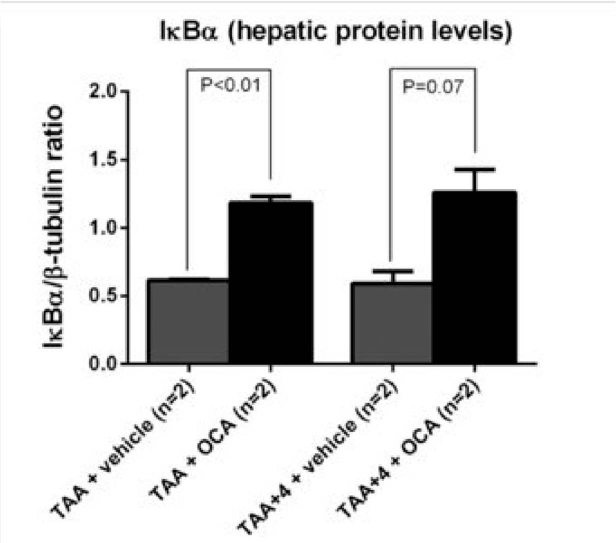
Synergy of a drug with itself?



The pharmacological concept of drug synergy requires more than just statistical interaction

Factorial design
Further examples

- Critser et al. (J Reprod Fert, **64**, 79-83, 1982) 3x4 and 3x5 designs
  - ✓ Reproductive Status:
    pregnant, pseudopregnant, hysteroctomized
  - ✓ Day after mating:
    e.g. 6, 8, 10, 12
- Factorial structure not always recognized



Verbeke, L. et al. *Scientific Reports* 6, Article number: 33453 (2016)

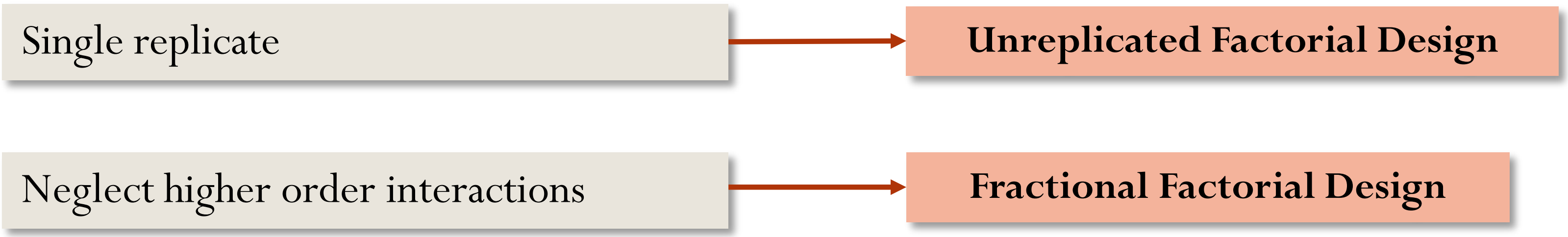
- ➢ The need for a factorial experiment should be **recognized at the design phase**
- ➢ The analysis should make use of the proper methods
  i.e. two-way ANOVA

Factorial design
Higher dimensional designs

- 3 x 3 x 3 design = 27 treatments , 2 replicates = 54 experimental units

  is this manageable?

- Interpretation of higher (3-way) interactions?



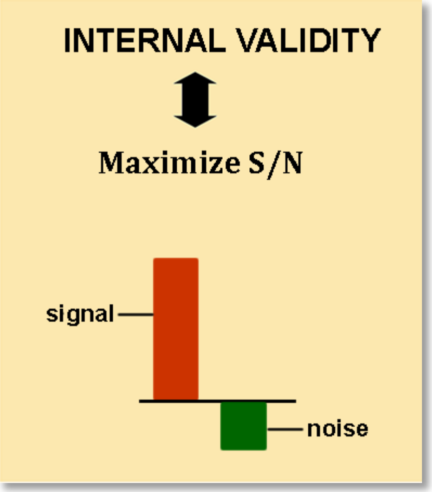
Single replicate → **Unreplicated Factorial Design**

Neglect higher order interactions → **Fractional Factorial Design**

Exploratory studies to identify possible sources of variation

Optimising animal experiments by factorial designs

- Find conditions for which signal/noise ratio is maximised
- Maximum signal/noise ratio minimises required number of animals
- Signal:
  Known positive control – vehicle control
- Factors affecting noise:
  - Animal related:
    sex, strain, age, diet, health status, body weight, etc.
  - Environmental:
    cage and group size, bedding material, etc.
  - Protocol specific:
    dose level, timing of administration and observation, route of administration



INTERNAL VALIDITY

⬍

Maximize S/N

Optimising animal experiments by factorial designs (Shaw et al. 2002)

One factor at a time

Vary each factor one at a time keeping all other factors at a fixed level.

- Each group of animals will contribute to understanding the effects of a single factor
- Does not look at interplay of factors
- Leads to incremental changes and multiple studies over time

Factorial design

Simultaneously vary all factors of interest

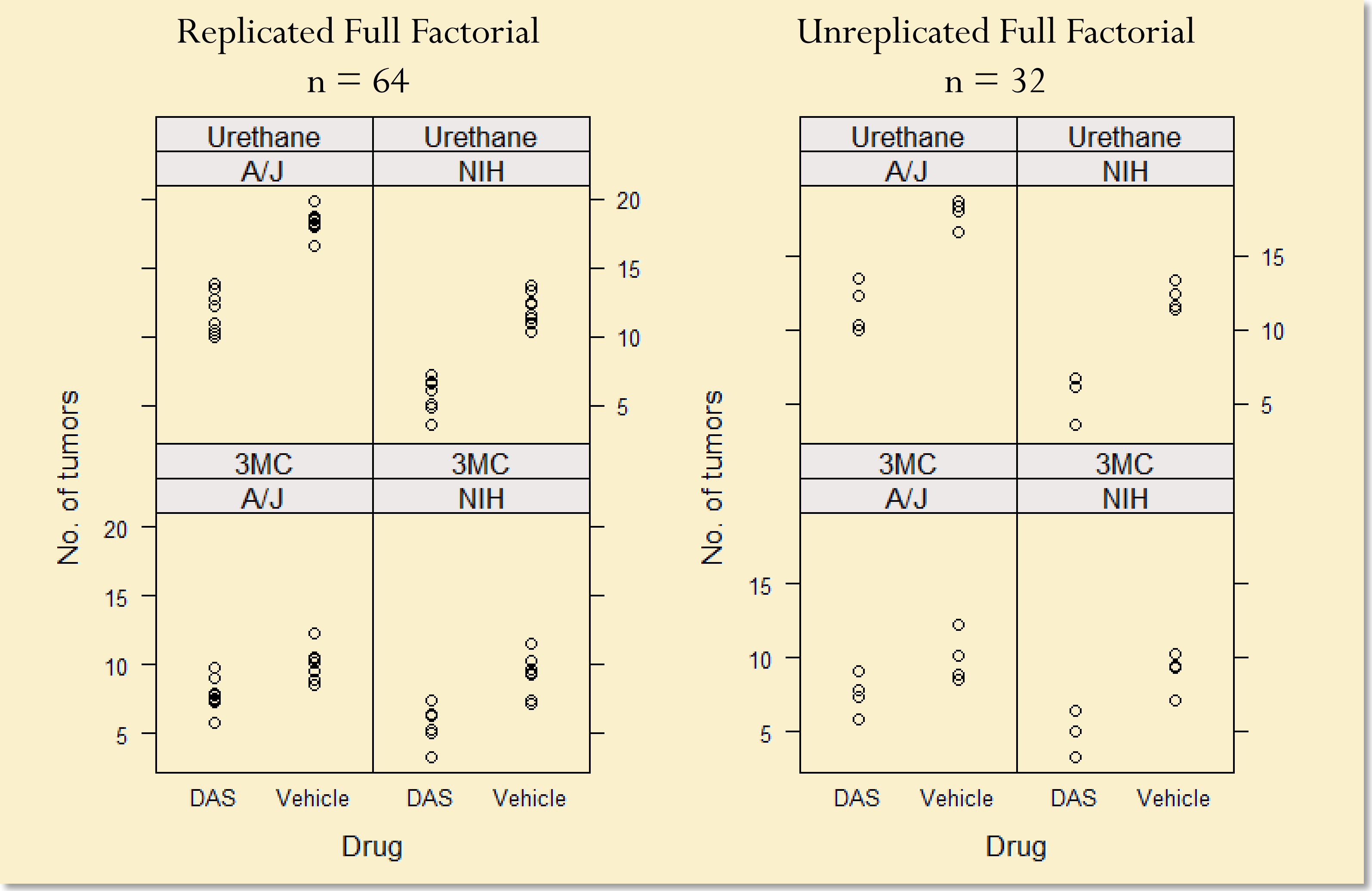
- Each animal contributes to the understanding of the effect of all factors under exploration
- Effect of one factor can depend on the level of another
- All potential factors are considered at the study outset

Optimising animal experiments by factorial designs - Example (Shaw et al. 2002)

- Animal model of lung cancer
- Factor affecting signal:

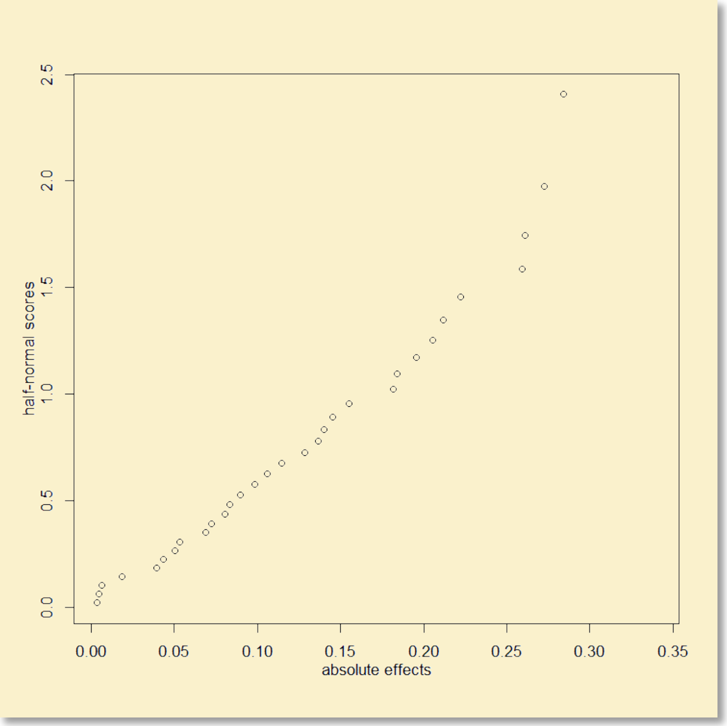
  Pharmacological treatment: Diallyl sulfide (garlic, *DAS*) – Vehicle
- Factors affecting noise:
  - ✓ Strain of mice: *A/J* and *NIH*
  - ✓ Gender
  - ✓ Diet: *RM1* and *RM3*
  - ✓ Carcinogen: urethane – 3-methylcholanthrene (*3MC*)
- Total of 5 factors ➡ $2^5 = 32$ factor combinations
- Testing one factor at a time with 6 animals/group = **60** mice
- Factorial experiment allows detection of interactions
- Full factorial with 2 animals/treatment = **64** animals

Optimising animal experiments by factorial designs - Example (Shaw et al. 2002)

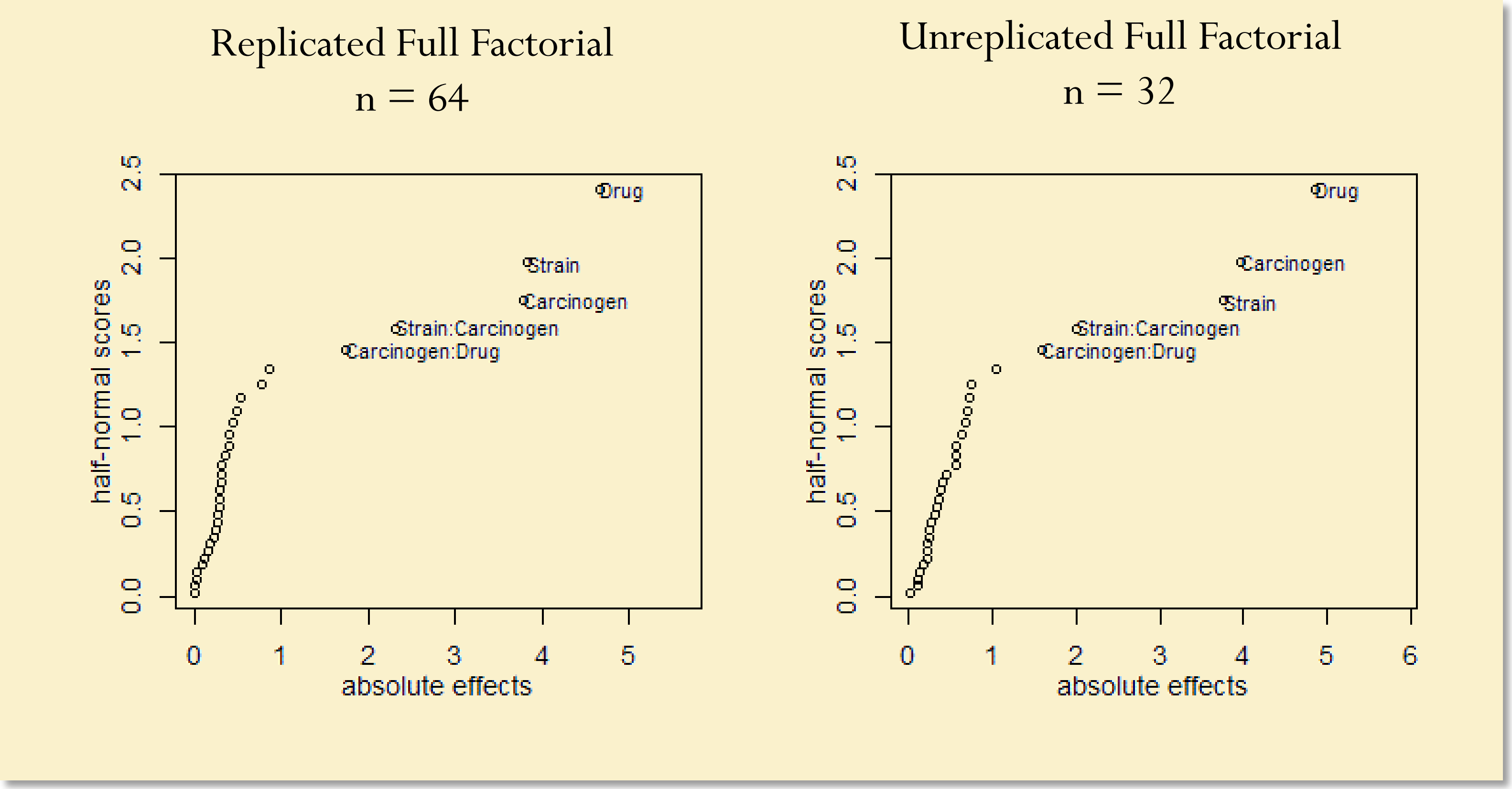


Looking at effects in large factorial designs Half-normal probability plot

- Identify most important factors that influence response
- Treatment effects from ANOVA model
- Take absolute value of a_i effects
- Plot value expected from a normal distribution against the ordered values of a_i
- When no factors are important estimated effects behave like random samples from normal distribution
- Plot is straight line, deviations from line indicate important effects
- *DanielPlot* from package *FrF2*



Optimising animal experiments by factorial designs - Example (Shaw et al. 2002)



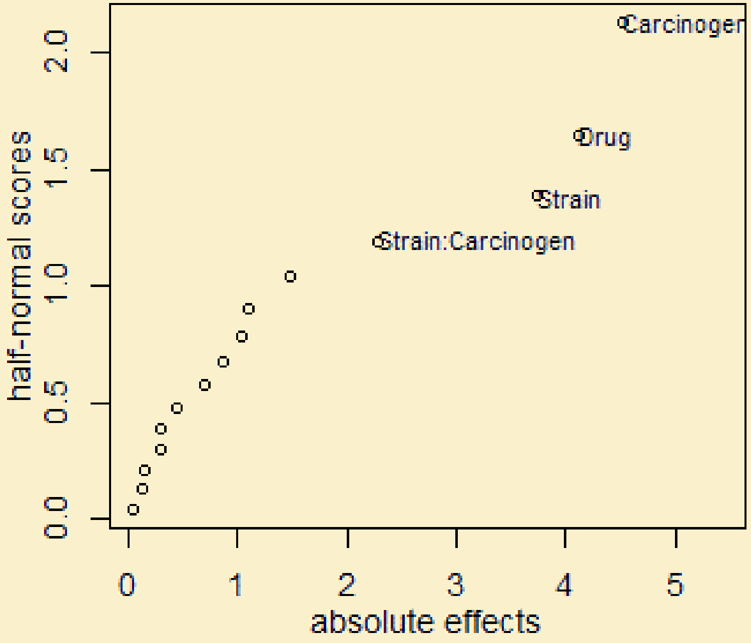
The Replicated Full Factorial Design and the Unreplicated Full Factorial Design lead essentially to the same conclusions, but for the UFD at half the price

Optimising animal experiments by factorial designs - Example (Shaw et al. 2002)

- Fractional Factorial with n =16 animals
- *FrF2* library:

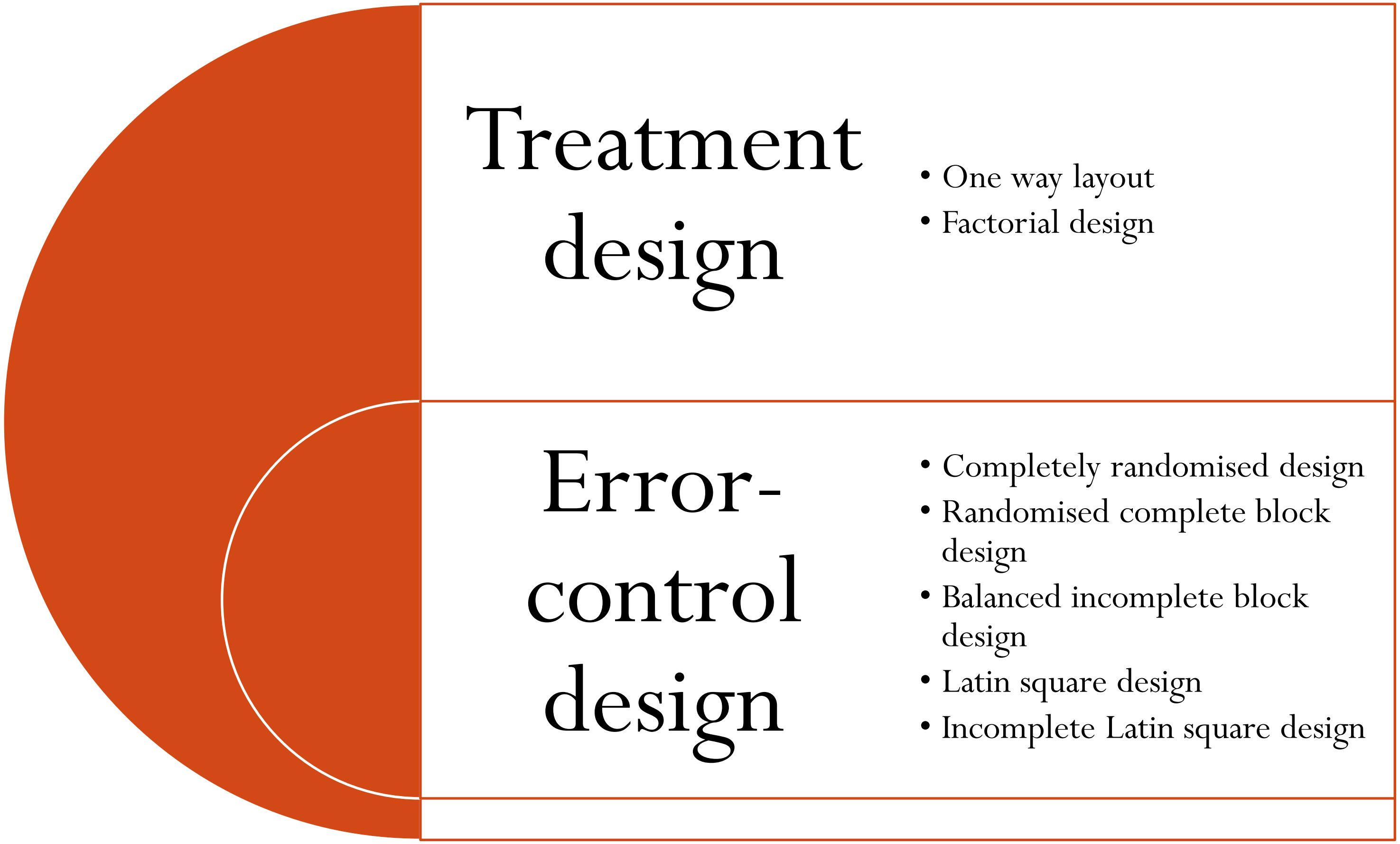
  > *library(FrF2)*
  > *des<-FrF2(16,nfactors=5)*

  See Table 5.5



FrFD correctly identifies major determinants of variability in experiment and can detect some interactions with only ¼ animals of a full factorial design with 64 animals

Combining
Treatment with error-control designs



Treatment design

- One way layout
- Factorial design

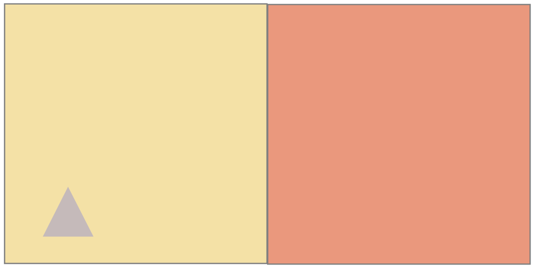
Error-control design

- Completely randomised design
- Randomised complete block design
- Balanced incomplete block design
- Latin square design
- Incomplete Latin square design

More complex designs
Split Plot Design

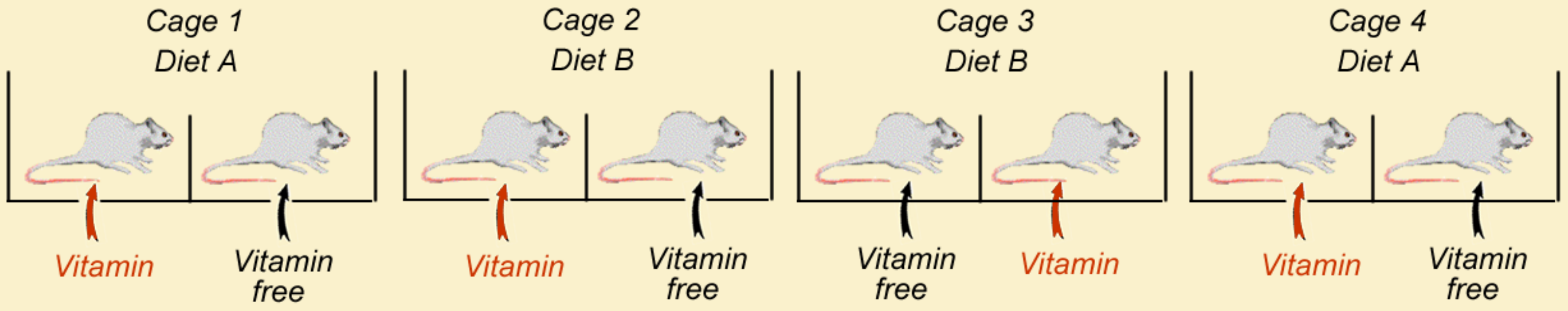
- A split-plot design recognizes two types of experimental units:
  - Plots
  - Subplots
- Two-way crossed (factorial) **treatment design**
  Split plot **error-control design**

- Plots randomly assigned to primary factor (color)
- Subplots randomly assigned to secondary factor (symbol)
- 2 completely randomized designs superimposed



Split Plot design
Example

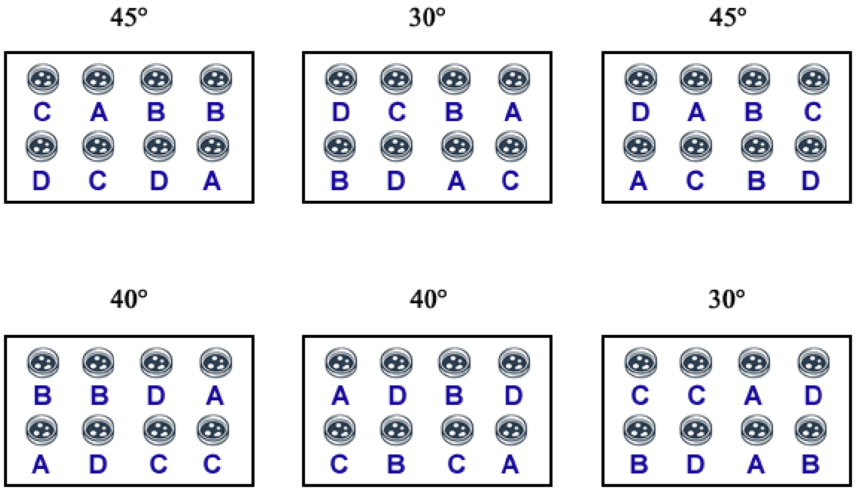
- Effects of 2 diets and vitamin supplement on weight gain in rats
- Rats housed 2 / cage (independence?)
- Cages randomly allocated to diet A or diet B
- Individual rats color marked randomly selected to receive vitamin or not



- Main plot factor = diet
- Subplot factor = vitamin

Split Plot design
Example

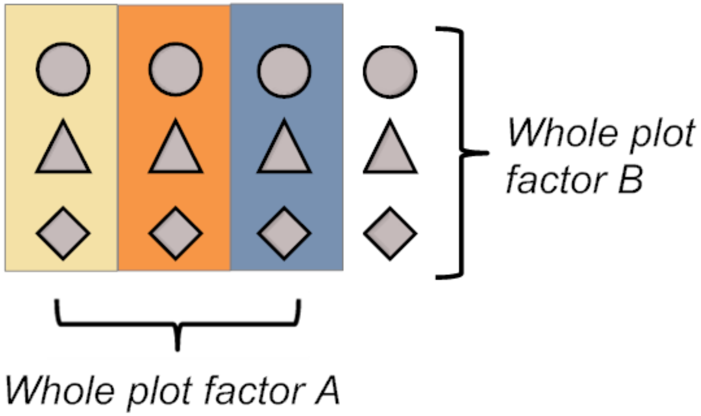
- Effect of temperature 30, 40 and 45 °C and growth medium (A, B, C, D) on yeast growth rate
- 6 incubators randomly assigned to temperature
- Per incubator 8 Petri dishes randomly assigned to growth medium



- Main plot factor = temperature (incubator)
- Subplot factor = growth medium

Strip-Plot or Split-Block Design

- Two whole plots placed orthogonal upon one another
- Two-way crossed (factorial) **treatment design**
  Strip plot **error-control design**



- subunit treatments are the same across entire main plots
- physical operations, e.g. harvesting, use of multipipette system, etc.
- sacrifices precision in main effects
- improves precision in interaction effects
- use R-package ***agricolae***
- found their way to the modern lab in experiments on 96-well plates (Lanski, 2002)

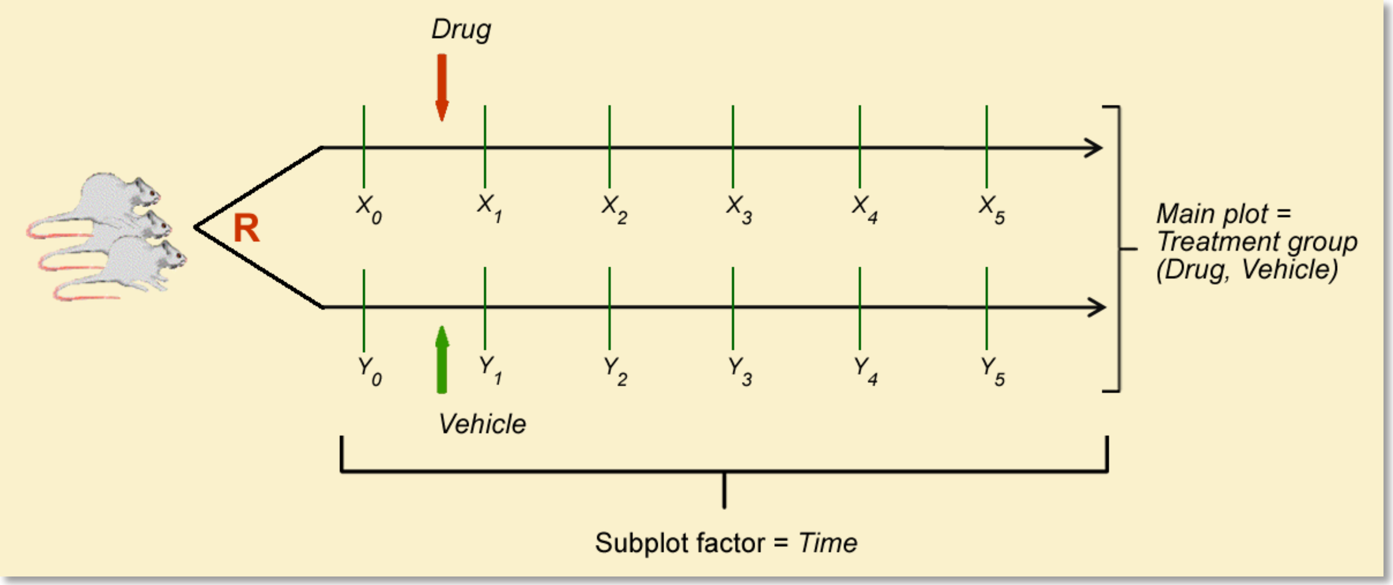
Strip-Plot or Split-Block Design
Example 96-well plate

- 12 columns (1 – 12), 8 rows (A – H)
- Samples A, B, C, D in duplicate A1, A2, etc. in rows
- Dilution levels 1, 2, 3, …, 12 in columns

	1	2	3	4	5	6	7	8	9	10	11	12
A	B1/2	B1/8	B1/10	B1/1	B1/11	B1/3	B1/12	B1/7	B1/4	B1/5	B1/6	B1/9
B	D2/2	D2/8	D2/10	D2/1	D2/11	D2/3	D2/12	D2/7	D2/4	D2/5	D2/6	D2/9
C	B2/2	B2/8	B2/10	B2/1	B2/11	B2/3	B2/12	B2/7	B2/4	B2/5	B2/6	B2/9
D	C1/2	C1/8	C1/10	C1/1	C1/11	C1/3	C1/12	C1/7	C1/4	C1/5	C1/6	C1/9
E	A1/2	A1/8	A1/10	A1/1	A1/11	A1/3	A1/12	A1/7	A1/4	A1/5	A1/6	A1/9
F	A2/2	A2/8	A2/10	A2/1	A2/11	A2/3	A2/12	A2/7	A2/4	A2/5	A2/6	A2/9
G	C2/2	C2/8	C2/10	C2/1	C2/11	C2/3	C2/12	C2/7	C2/4	C2/5	C2/6	C2/9
H	D1/2	D1/8	D1/10	D1/1	D1/11	D1/3	D1/12	D1/7	D1/4	D1/5	D1/6	D1/9

- row 1 sample B1, column 1 = dilution 2
- row 2 sample D2, column 2 = dilution 8

Repeated Measures Design

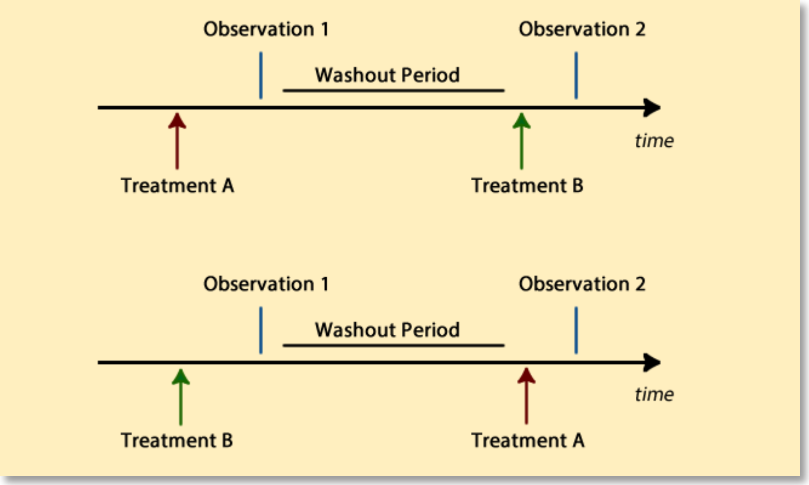


- Two independent variables 'time' and 'treatment'

  Each is associated with different experimental units

- Special case of split plot design

Crossover Design



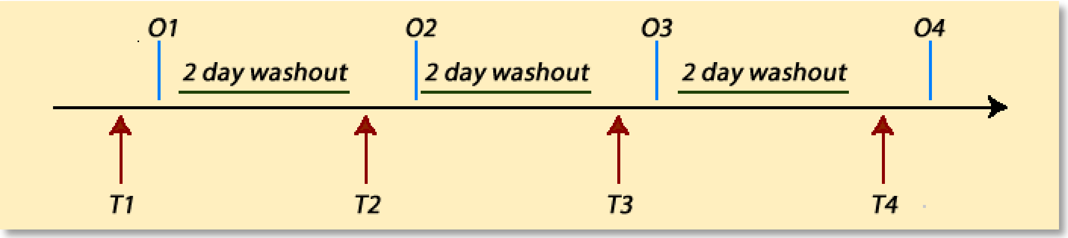
Washout periods – carryover effects

Randomised complete block Latin square error-control design

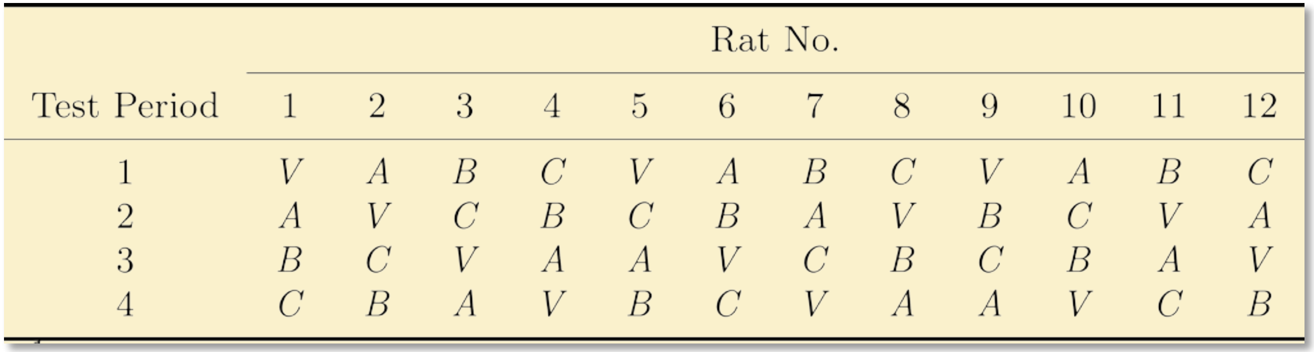
- Closely related to repeated measures design
- Sequence of treatments over different test periods
- 1 period = 1 treatment, all treatments in all subjects
- Sequences switched over between subjects, AB, BA, etc.
- Randomised complete block designs with subjects as block
- Latin square designs (subjects x time)

Crossover Design Example

- 5-HT4 agonists on attentional deficit
- 4 treatment conditions: vehicle (*V*), drug at 2 concentrations (*A*,*B*), positive control (*C*)
- Rats trained for attentional deficit, expensive resource
- Compounds have short-term effect
- Crossover design – each compound to be tested in each animal



Three 4x4 Latin Square Design



	Rat No.											
Test Period	1	2	3	4	5	6	7	8	9	10	11	12
1	*V*	*A*	*B*	*C*	*V*	*A*	*B*	*C*	*V*	*A*	*B*	*C*
2	*A*	*V*	*C*	*B*	*C*	*B*	*A*	*V*	*B*	*C*	*V*	*A*
3	*B*	*C*	*V*	*A*	*A*	*V*	*C*	*B*	*C*	*B*	*A*	*V*
4	*C*	*B*	*A*	*V*	*B*	*C*	*V*	*A*	*A*	*V*	*C*	*B*

Crossover Design

Advantages

- Experimental units are animals or subjects within test period
- No bias due to differences in subjects
- More precise comparisons, less subjects or animals are required.

Drawbacks

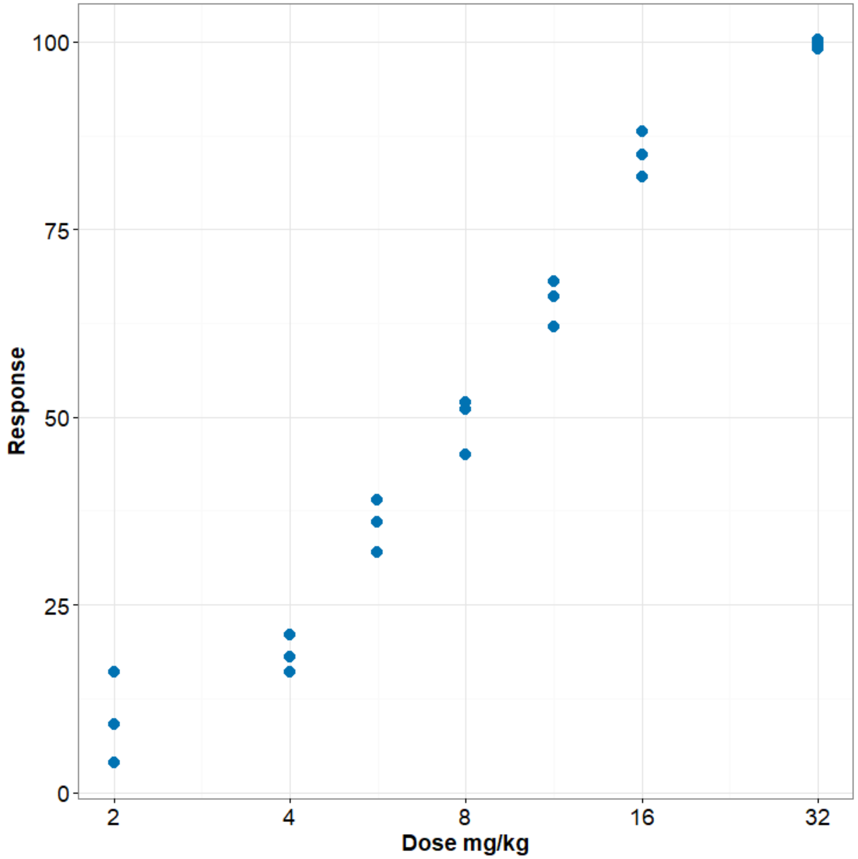
- Carry-over effects
  special designs (Jones & Kenward)
- Take longer to complete
- Ethical concerns

Crossover designs are the standard design in studies for bioequivalence

Dose Response Experiments

- Estimate shape of DRC
- Model fit rather than comparisons
- Estimate parameters of functional model
- Determine threshold dose
- Predict response for intermediate doses

- Treatment design:
  - One-way layout
  - Treatment (dose, concentration) continuous
- Error-control design
  - CRD, RCBD, …..
  - To be taken into account at analysis time (Pinheir & Bates, 2001)



Choice of doses/concentrations free
6 doses in most cases optimal
Dose placement based on optimality criteria

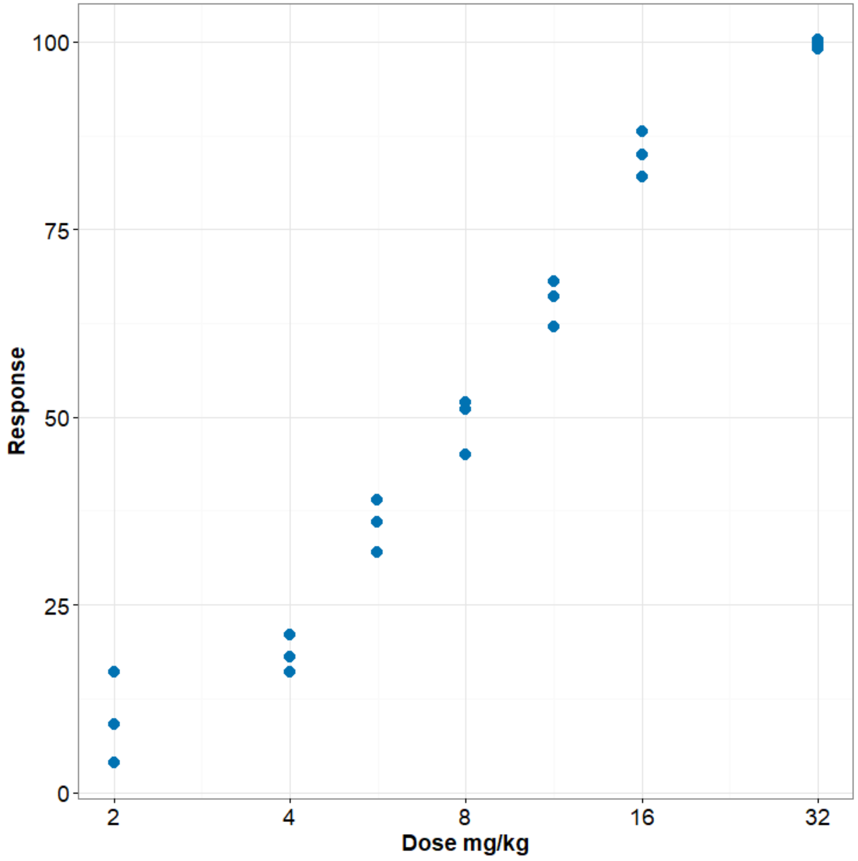
Dose Response Experiments - Example Analgesic action of morphine sulfate

Response = tail flick latency
7 doses of drug (morphine sulphate)
3 rats/dose (randomly assigned) (CRD)

Statistical & Functional model:

$$y_{ij} = f(x_i) + \epsilon_{ij}$$

$$f(x_i; c, b, d, e) = c + \frac{d - c}{1 + exp(b(log(x) - log(e))}$$



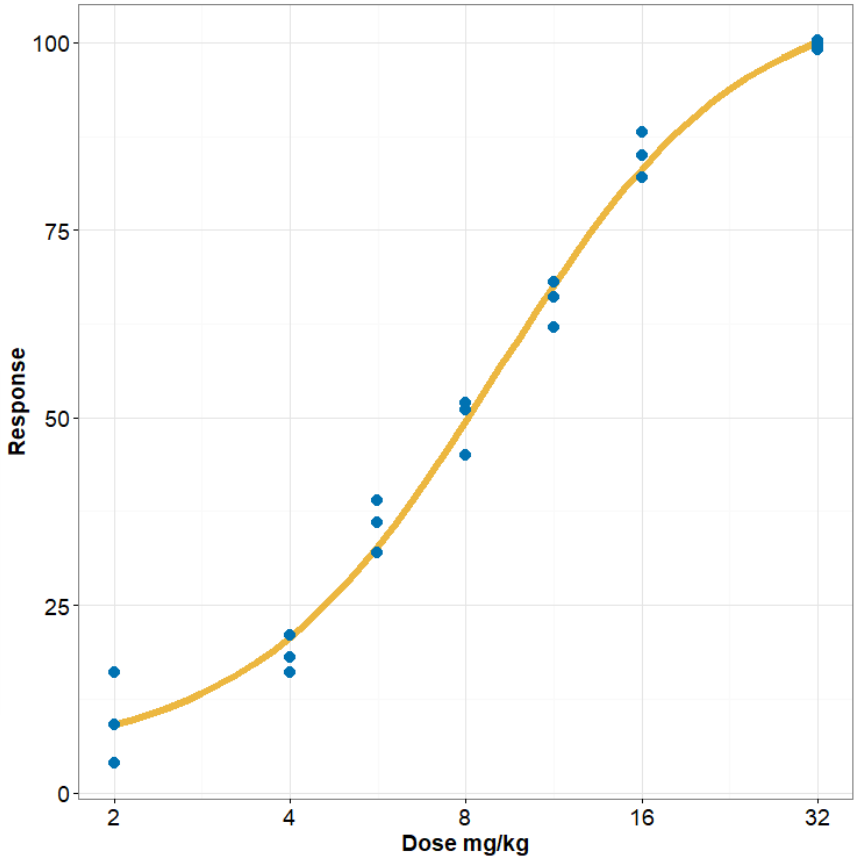
Dose Response Experiments - Example
Analgesic action of morphine sulfate

Response = tail flick latency
7 doses of drug (morphine sulphate)
3 rats/dose (randomly assigned) (CRD)

Statistical & Functional model:

$$y_{ij} = f(x_i) + \epsilon_{ij}$$

$$f(x_i; c, b, d, e) = c + \frac{d - c}{1 + exp(b(log(x) - log(e))}$$



Fitting the 4PL functional model, does not require computation of percentages !

The Smart Design of Animal Experiments

VI. Sample Size and Power

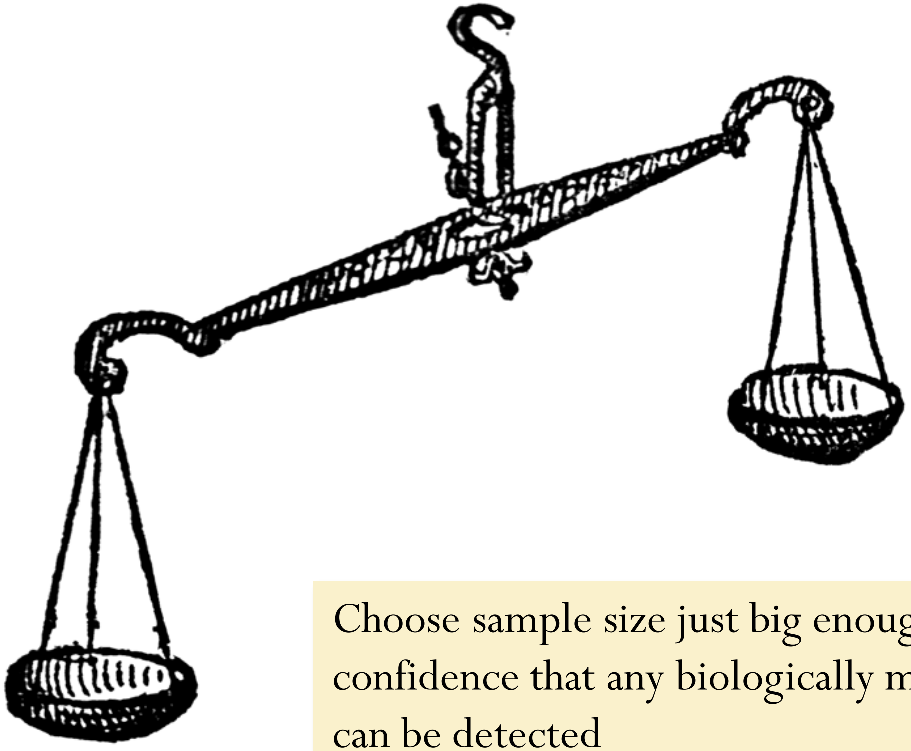
The need for sample size determination

- Key factor for effective and efficient experimentation
- Justification of number of animals requested by animal care committee

Prevent waste of animals:

- ➢ Too few animals:
  - experiment lacks statistical power to detect real treatment effect
  - more studies will be carried out
  - results not trustworthy (effect size inflation)
- ➢ Too many animals:
  - biologically irrelevant effects declared statistically significant
  - animals suffer unnecessary harm

Determining sample size is a risk-cost assessment

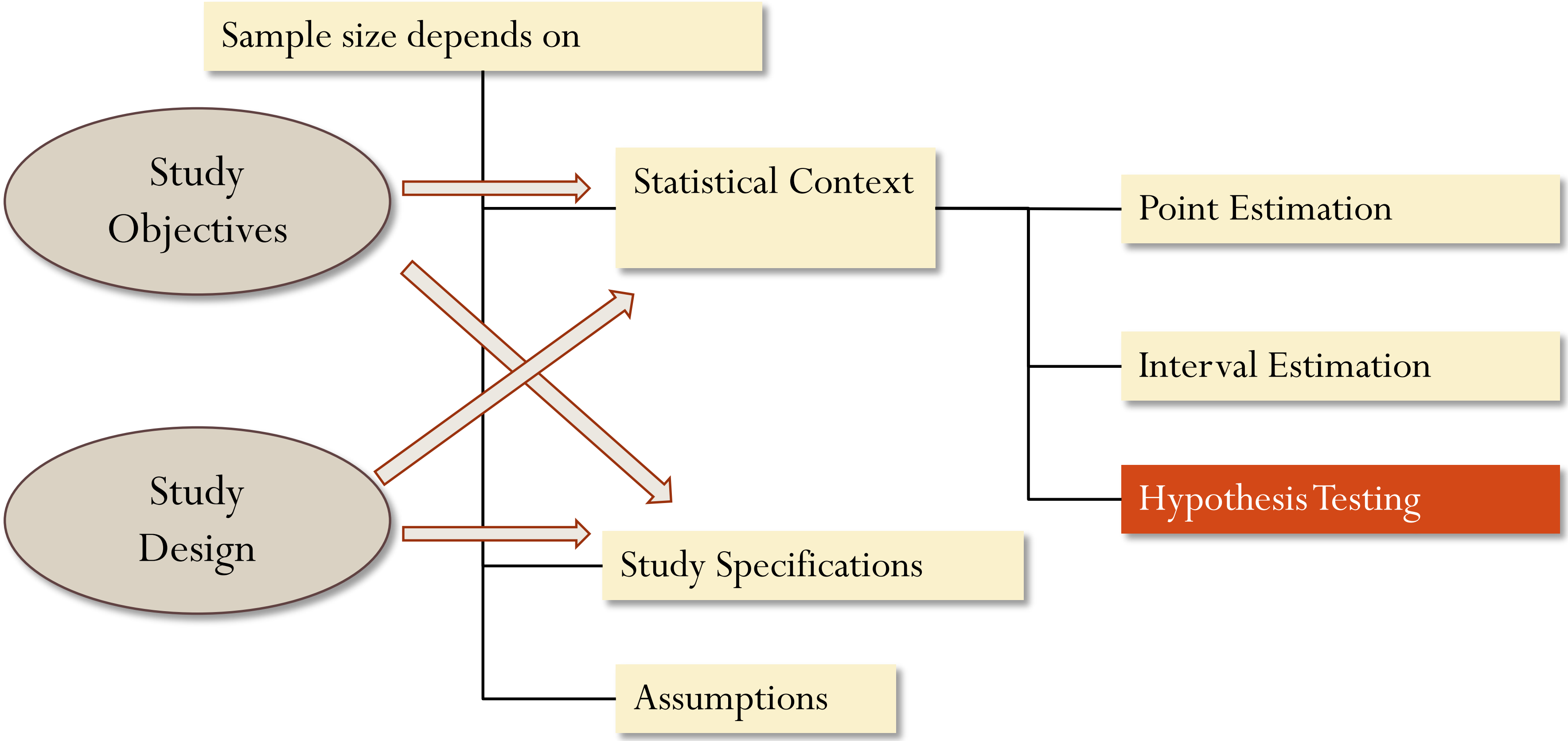


Replicates
Cost

Uncertainty
Confidence

Choose sample size just big enough to give confidence that any biologically meaningful effect can be detected

The context of biomedical experiments



Sample size depends on

Study Objectives

Study Design

Statistical Context

Study Specifications

Assumptions

Point Estimation

Interval Estimation

Hypothesis Testing

The hypothesis testing context
Neyman-Pearson framework

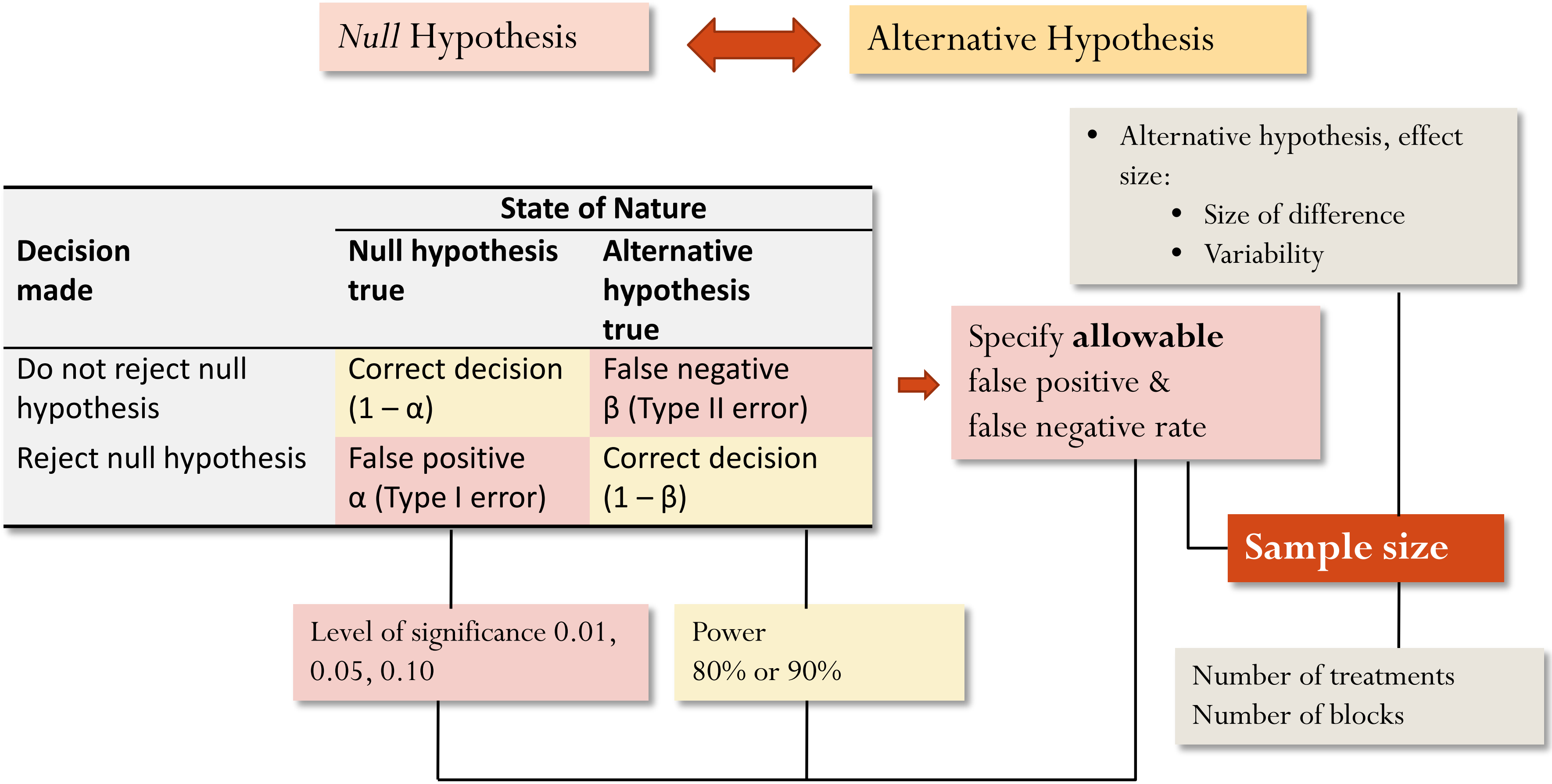
Jerzy Neyman & Egon Pearson (1933)

Null Hypothesis ↔ Alternative Hypothesis

Decision made	**State of Nature**	
	Null hypothesis true	**Alternative hypothesis true**
Do not reject null hypothesis	Correct decision (1 – α)	False negative β (Type II error)
Reject null hypothesis	False positive α (Type I error)	Correct decision (1 – β)

The hypothesis testing context
Sample size

Jerzy Neyman & Egon Pearson (1933)



Null Hypothesis

Alternative Hypothesis

- Alternative hypothesis, effect size:
  - Size of difference
  - Variability

Decision made	State of Nature	
	Null hypothesis true	**Alternative hypothesis true**
Do not reject null hypothesis	Correct decision (1 – α)	False negative β (Type II error)
Reject null hypothesis	False positive α (Type I error)	Correct decision (1 – β)

Specify **allowable** false positive & false negative rate

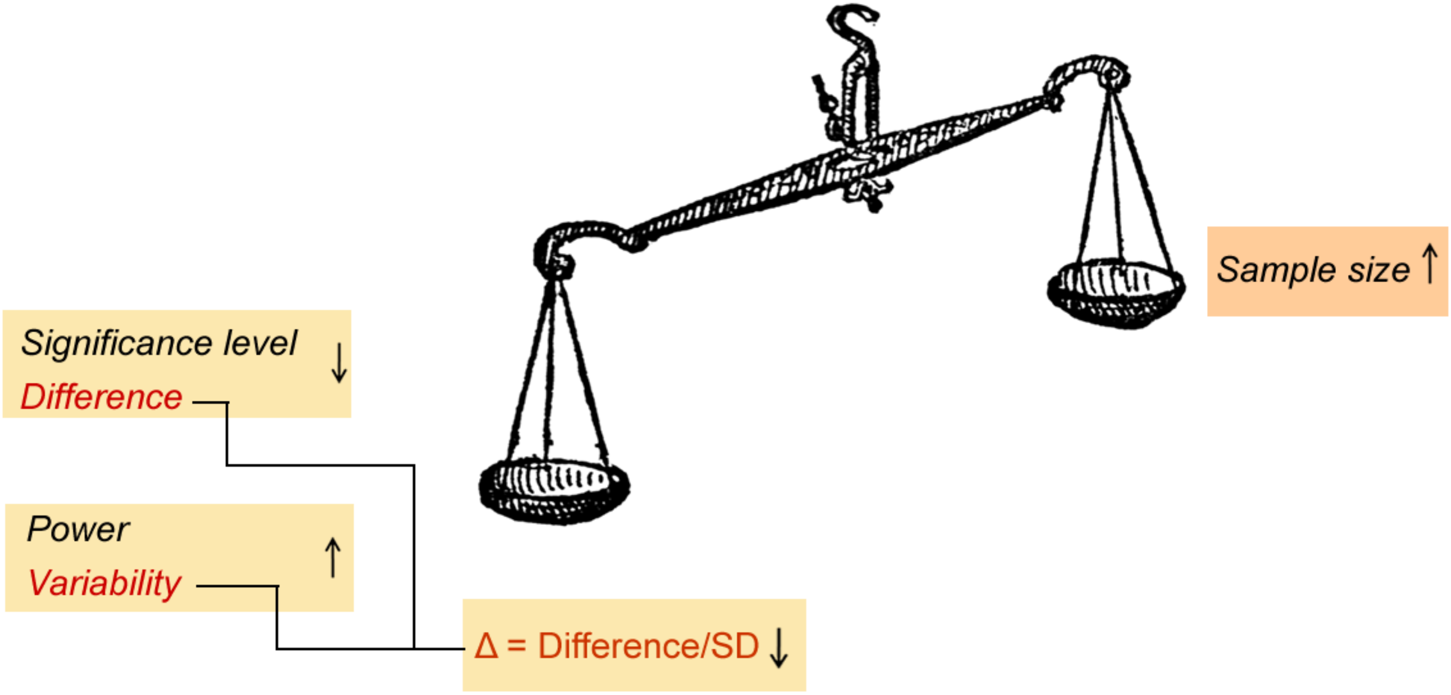
Sample size

Level of significance 0.01, 0.05, 0.10

Power 80% or 90%

Number of treatments
Number of blocks

Major determinants of sample size



Significance level ↓

Difference

Power ↑

Variability

Δ = Difference/SD ↓

Sample size ↑

Sample size calculations

Requires input of

- significance level
- power
- alternative hypothesis (i.e. effect size Δ)

{ smallest difference
  variability (SD)

$$\Delta = \frac{|\mu_1 - \mu_0|}{\sigma}$$

- ➢ R-package *pwr* (Champely, 2009), *StatMate*
- ➢ Dell (2002): $n = 1 + \frac{2C}{\Delta^2}$ *C* depends on power and significance level (Table 6.2, 7.85), *n* per group
- ➢ Lehr's equation: $n \approx {}^{16}/_{\Delta^2}$
  - *n* per group
  - single comparison (2 groups)
  - significance level 0.05, two-sided
  - power 80%, other powers possible

Sample size calculations
Examples

Effect size Δ = difference in means / standard deviation

Δ = 0.2 small; Δ = 0.5 medium; Δ = 0.8 large (Cohen)

Δ = 1.0, Δ = 1.2 (Shaw 2002)

Cardiomyocytes completely randomized design

Treated - Control = 15; SD = 12.4

Δ = 15 / 12.4 = 1.2 (α = 0.05, 100 x (1-β) = 80%)

SD: standard deviation of either group or pooled SD

```
> require(pwr)
> pwr.t.test(d=1.2,power=0.8,sig.level=0.05,type="two.sample",alternative="two.sided")
```

```
Two-sample t test power calculation
              n = 11.94226
              d = 1.2
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
 NOTE: n is number in *each* group
```

Dell (2002):

$$n = 1 + \frac{2 \times 7.85}{1.2^2} \approx 12$$

Lehr's equation:

$$n \approx {}^{16}/_{\Delta^2} = {}^{16}/_{1.44} \approx 12$$

Sample size calculations
Examples – Power of an experiment

Power of **completely randomized experiment** with 5 animals per treatment group to detect an effect $\Delta = 15/12.4 = 1.2$?

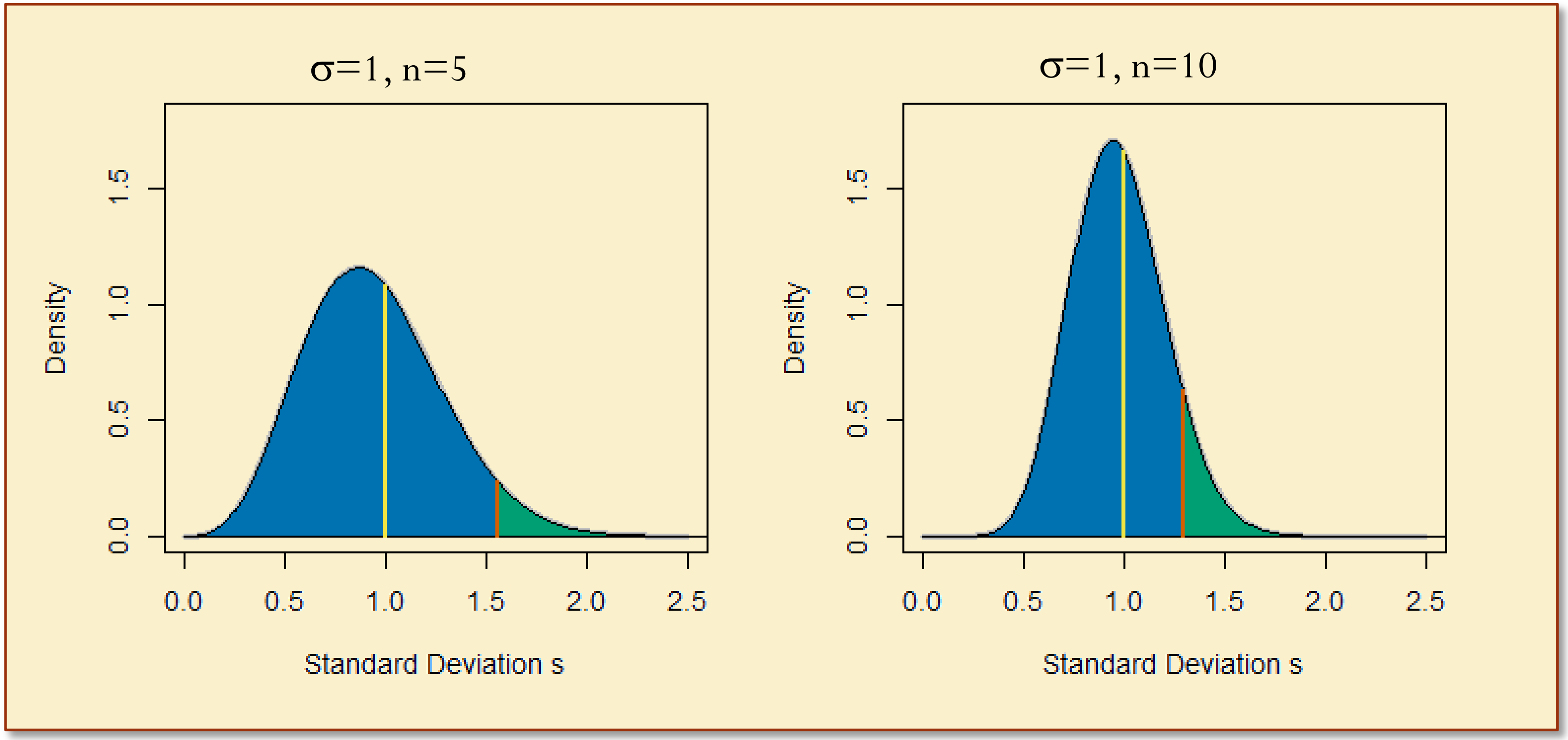
```
> require(pwr)
>  pwr.t.test(d=1.2,n=5,sig.level=0.05,type="two.sample",
+                  alternative="two.sided")

Two-sample t test power calculation

              n = 5
              d = 1.2
      sig.level = 0.05
          power = 0.3864373
    alternative = two.sided

NOTE: n is number in *each* group
```

Power = 39%

Uncertainty in estimating the standard deviation σ

$$\Delta = \frac{|\mu_1 - \mu_0|}{\sigma}$$

- How do we know σ?
- Previous experiments, literature, …
- What if we underestimate σ?



Uncertainty in estimating the standard deviation σ - Example

$$\Delta = \frac{|\mu_1 - \mu_0|}{\sigma}$$

- Conservative value for Δ by using 80% or 90% upper confidence interval for σ

- SD = 12.4 based on two-group comparison of *2 x (n - 1) = 8 d.f.*
- **Table 6.3** upper 80 % C.L.:

  σ multiplication factor = 1.32, Inflation factor for n = 1.742
- 80% upper CL for σ = 12.4 x 1.32 = 16.37
- Δ = 1.2 corresponds to: 1.2 x 16.37 = **20 viable cardiomyocytes**
- Or keep minimum detectable difference of **15 myocytes** and increase sample size (12) of new study by 12 x 1.742 = **21 animals/treatment group**
- 80% confidence that estimated sample size is adequate

Sample size based on coefficient of variation

- Biologists tend to think in terms of percentages
- Example:
  - Two-group experiment
  - Differences in means of 20%
  - Variability is about 30%
  - Level of significance α = 0.05
  - Power = 80%

$$n = 8\frac{c_v^2}{\Delta_p^2}[1 + (1 - \Delta_p^2)]$$

$$c_v = \sigma/\mu$$

$$\Delta_p = (|\mu_1 - \mu_0|)/\mu_0$$

$$n = 8 \times \frac{0.30^2}{0.20^2} \times [1 + (1 - 0.20)^2] \approx 30 \text{ animals per treatment group}$$

Sample size determination
Paired experiments

Example:

- Paired experiment on cardiomyocytes
- SD of pairwise differences = 5.61, 4 d.f.
- 80% confidence to correctly estimate sample size
- Table 6.3: SD x 1.558 = 8.74
- Smallest detectable difference = 20 cardiomyocytes, Δ = 20/8.74 = 2.29
- 80% Power, α = 0.05

$$n = 2 + \frac{C}{\Delta^2}$$

See Table 6.2 for values of C

$$n = 2 + 7.85/(2.29^2) \approx 4$$

```
> require(pwr)
> pwr.t.test(d=2.29,power=0.8,sig.level=0.05, type="paired",alternative="two.sided")
Paired t test power calculation
n = 3.770236
d = 2.29
sig.level = 0.05
power = 0.8
alternative = two.sided
NOTE: n is number of *pairs*
```

Sample size determination
Binary (dichotomous) data

Binary data:

- Alive/death
- Present/absent

Dell (2002):

occurences r_1, r_2 in groups of size n_1, n_2

$$p_1 = \frac{r_1}{n_1},\ p_2 = \frac{r_2}{n_2}$$

$$H_0 : \pi_1 = \pi_2$$

$$n = C\frac{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{\Delta_\pi^2} + \frac{2}{\Delta_\pi} + 2$$

$$\Delta_\pi = |\pi_1 - \pi_2|$$

More accurate:

pwr-package - *pwr.2p.test*

Sample size determination
Binary data - Example

Life span of cardiomyopathic hamsters (Ver Donck, et al. 1991).

- Congestive heart failure
- Death within a year
- Response variable: % surviving hamsters after 300 days
- Expected responses:
- 15% survival control group, 50% survival drug treated group
- Sample size?

$$\pi_1 = 0.15, \pi_2 = 0.5$$

$$\Delta_\pi = |0.15 - 0.50| = 0.35$$

$$n = C\frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{\Delta_\pi^2} + \frac{2}{\Delta_\pi} + 2$$

$$n = 7.85 \times \frac{0.15 \times 0.85 + 0.50 \times 0.50}{0.35^2} + \frac{2}{0.35} + 2 \approx 32$$

Sample size determination
Binary data – Example (cont.)

```
> require(pwr)
> pwr.2p.test(h = ES.h(0.15,0.5),  sig.level = 0.05, power = .80,
+ alternative = "two.sided")

     Difference of proportion power calculation for binomial distribution (arcsine
transformation)

              h = 0.7753975
              n = 26.10885
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

Required sample size:
27 animals/treatment group

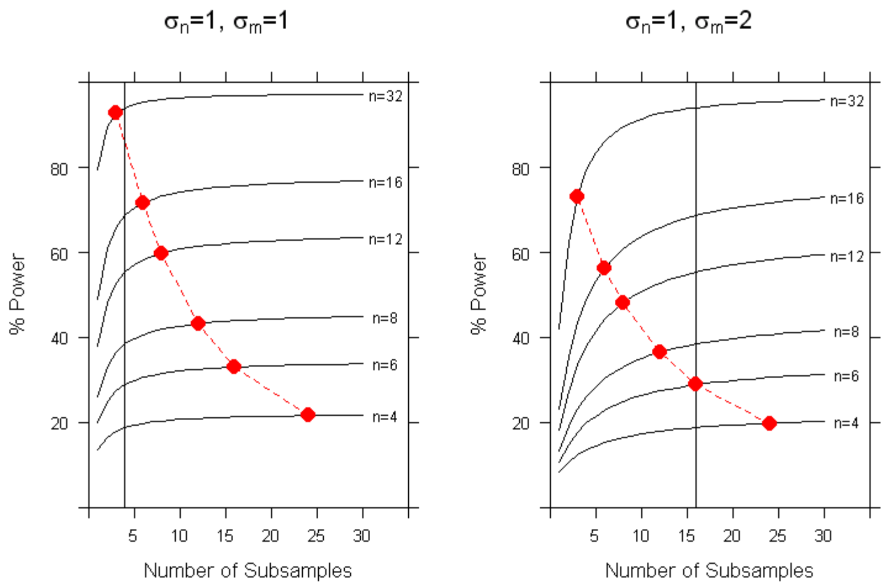
How many subsamples ?

m subsamples, standard deviation of subsamples σ_m

n experimental units, standard deviation σ_n

Standard error of difference:

$$\sqrt{\frac{2}{n}\left(\sigma_n^2 + \frac{\sigma_m^2}{m}\right)}$$



Vertical line corresponds to upper bound: $m = 4\,\frac{\sigma_m^2}{\sigma_n^2}$

Taking differential costs into consideration: $m = \sqrt{\frac{c_n}{c_m} \times \frac{\sigma_m^2}{\sigma_n^2}}$

How many subsamples ? Example

Morphologic study diameter of cardiomyocytes in sheep

7 sheep with intervention

6 sheep control

Diameter of 100 epicardial cells/sheep

Variance components: $\sigma_n^2 = 4.58, \sigma_m^2 = 13.7$

Upper bound (neglecting differential costs): $m = 4 \times 13.7/4.58 \approx 12$

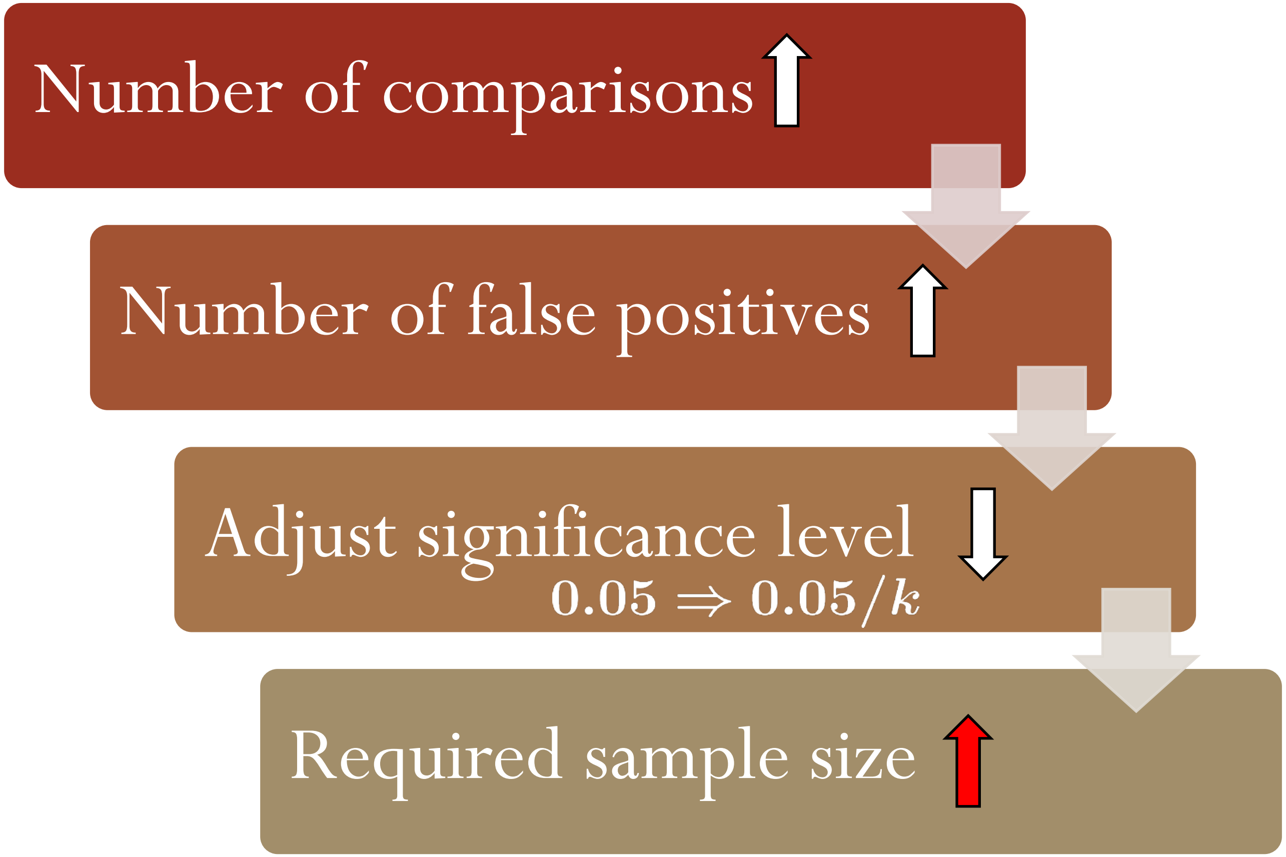
Assume cost of 1 animal is equivalent to 100 subsamples then:

$$m = \sqrt{100 \times 13.7/4.58} \approx 17$$

1 animal equivalent to 1000 subsamples then 55 cells sufficient

Taking a priori 100 subsamples was definitely a waste of time in this case

Multiplicity and sample size



Number of comparisons ↑

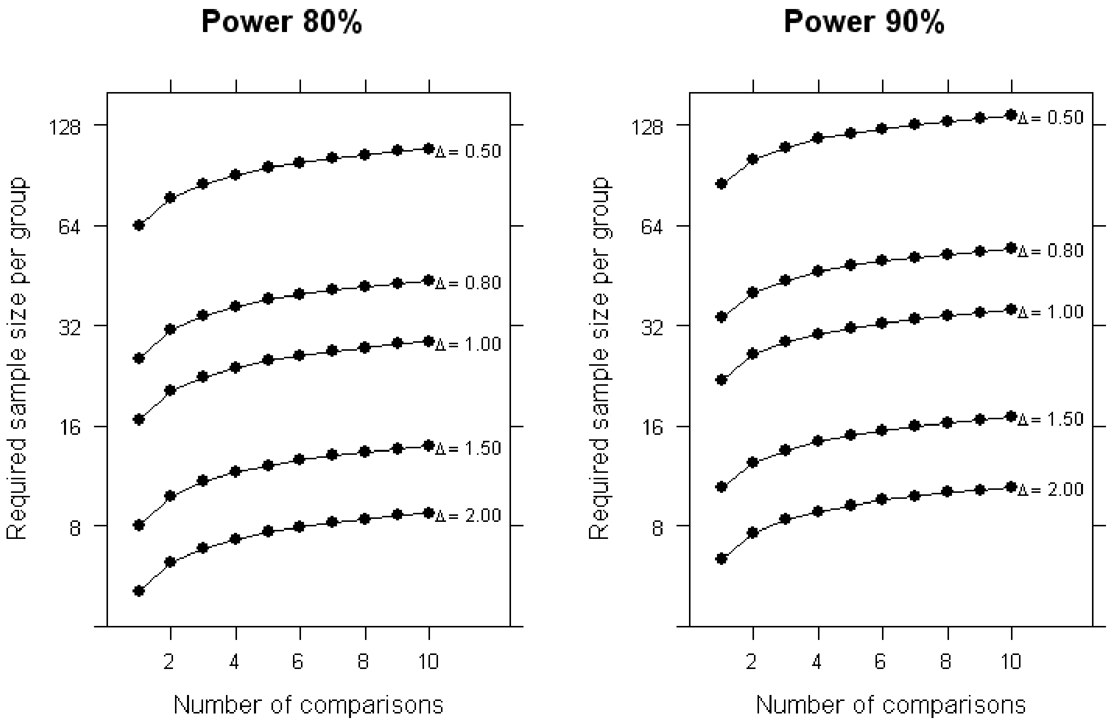
Number of false positives ↑

Adjust significance level ↓

$0.05 \Rightarrow 0.05/k$

Required sample size ↑

Multiplicity and sample size



Power 80%

Power 90%

- 2 tests: 20% increase RSS
- 3 tests: 30% increase RSS
- 4 tests: 40% increase RSS
- 10 tests: 70% increase RSS

- For single comparison, large sample sizes required for medium (Δ = 0.5) to large (Δ = 0.8) effects
- Effects in early research usually extremely large

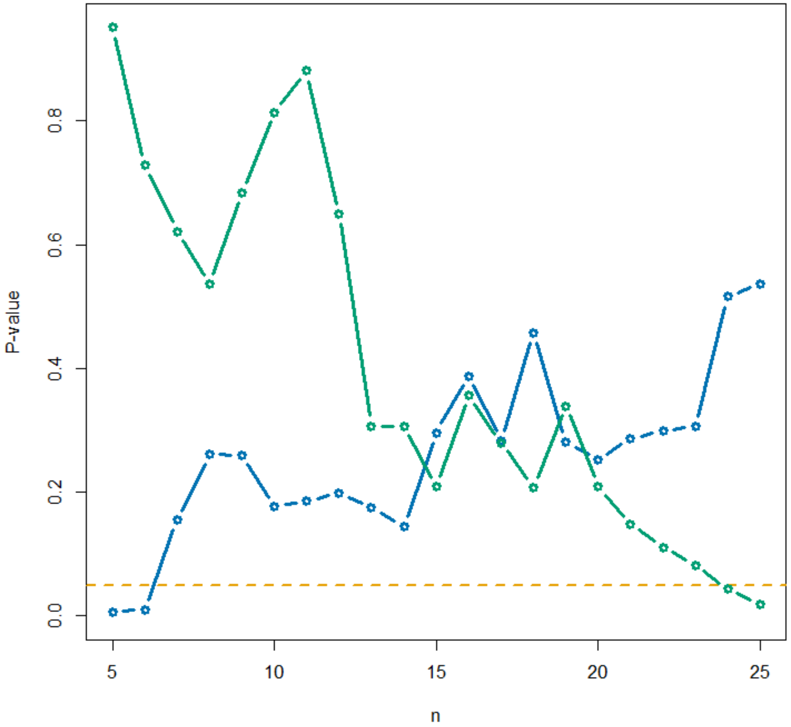
The problem of underpowered experiments

- Power calculations in only fewer than 3% of publications in *Science* and *Nature* (Tressoldi et al. 2013)
- In experimental neurosciences, power is between 8 and 32% (Button et al. 2013)

Consequences of low power:

- Wrongly conclude there is no effect
- In animal research, ethical concern:
  - Underpowered studies will fail to show a true effect
  - More studies will be carried out
  - An adequately powered study at the beginning will require the least animals
- Results are less replicable in small studies
- ***Truth Inflation*** (Reinhart, 2015): effect sizes are overestimated in small significant studies

Dangers of underpowered studies



2 situations:

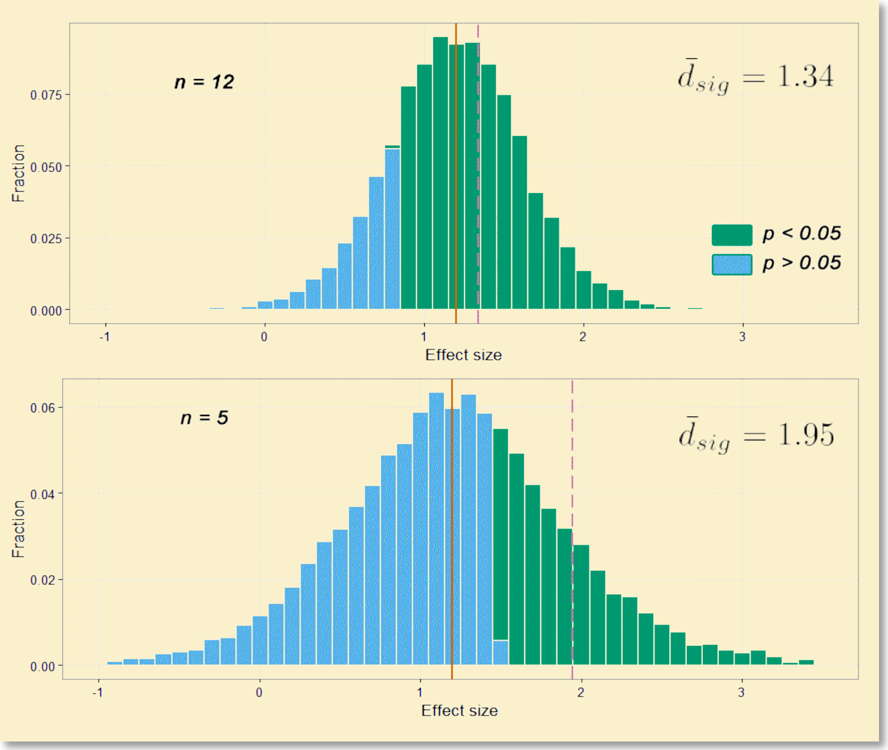
1. (blue) no true effect
2. (green) $\Delta = 0.8$

Required sample size to detect $\Delta = 0.8$ for each group, n = 25

Underpowered studies yield unreliable results

Underpowered experiments
Truth inflation – Type M error

- Cardiomyocyte experiment as completely randomized design
- 12 animals required for a power of 80% to detect a value of Δ=1.2 at two-sided significance level of 0.05
- Simulate 10,000 runs of experiment and estimate effect size



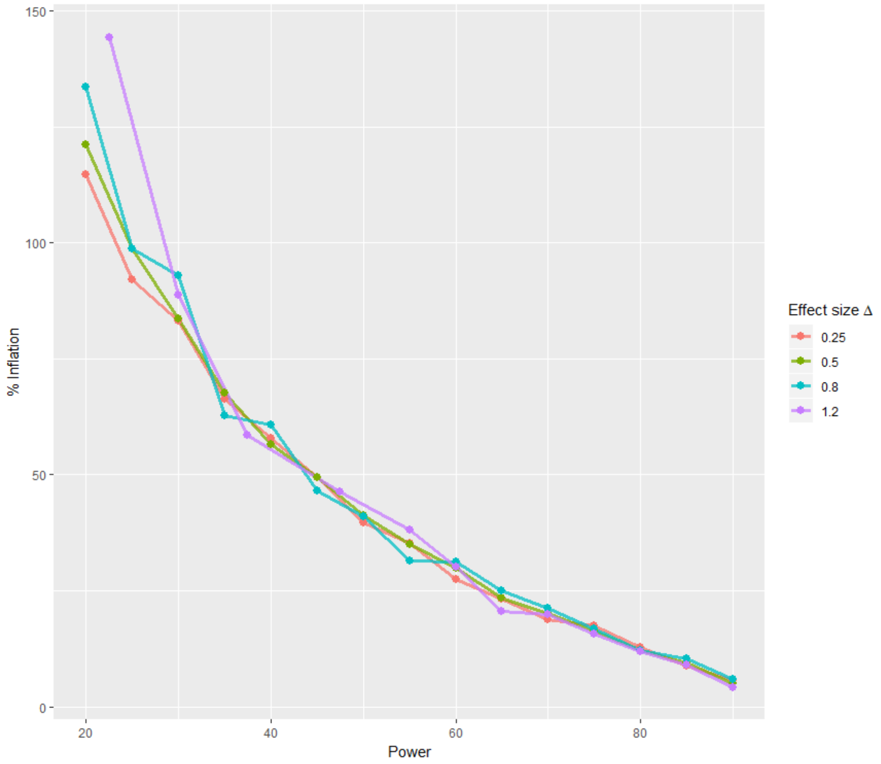
n =12:

- Power =80%
- Effect size inflation = 12%

n =5:

- Power = 39%
- Effect size inflation = **62%**

Underpowered experiments
Truth inflation – The winner’s curse



- Effect size inflation depends on power of study
- Effect size inflation is worst for small low-powered experiments which can only detect very large treatment effects

Consequences:

- Sample size calculations of follow-up studies will yield values that are too small
- Follow-up studies will fail to find a true effect

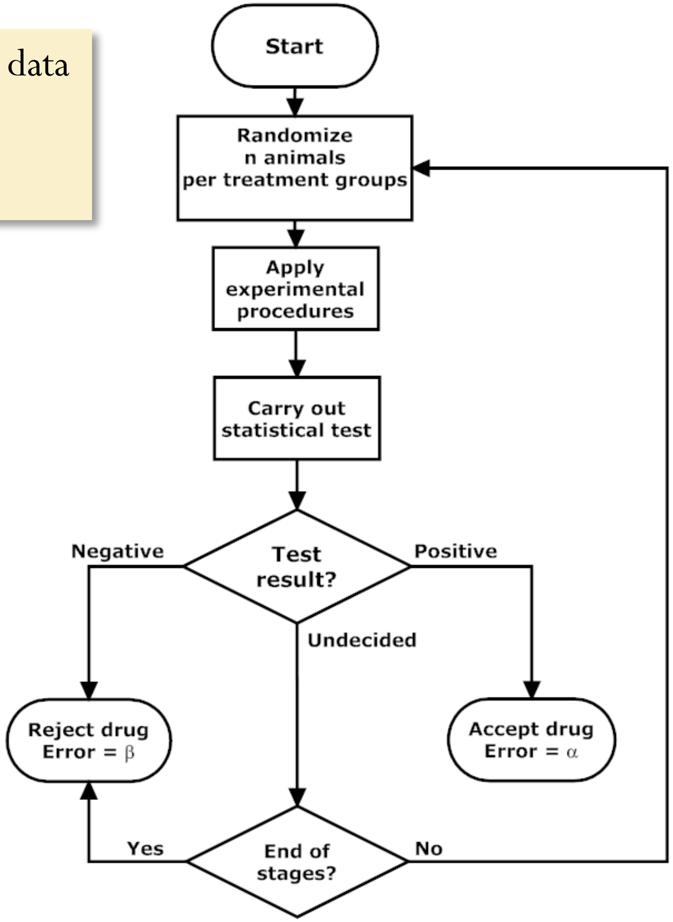
Effect size inflation due to small underpowered experiments is one of the major reasons for the lack of replicability in scientific research

Sequential Plans

- save on experimental material by testing at different stages as data accumulate
- recently advocated in animal experiments

Example screening for drugs that protect against traumatic brain injury:

- Screening, i.e. most compounds inactive
- Large variability in outcome, fixed sample size procedure unethical & inefficient
- One-sided sequential procedure
- After testing 50 compounds, a candidate compound was selected
- Advantage: type I and type II errors known and under control, economize on animals
- Disadvantage: dedicated computer program needed

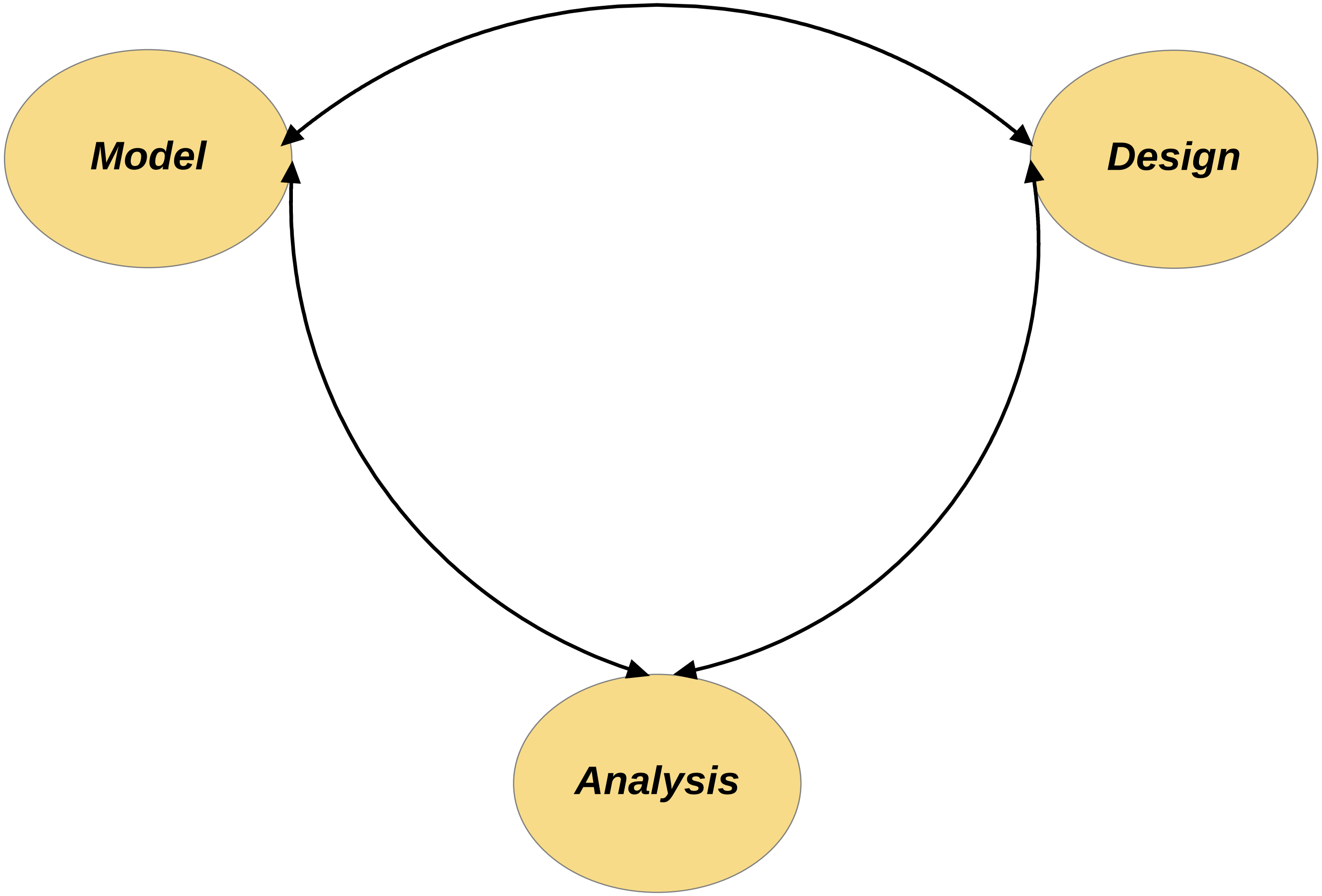


In case of early termination for significance, treatment effects are exaggerated (effect inflation)

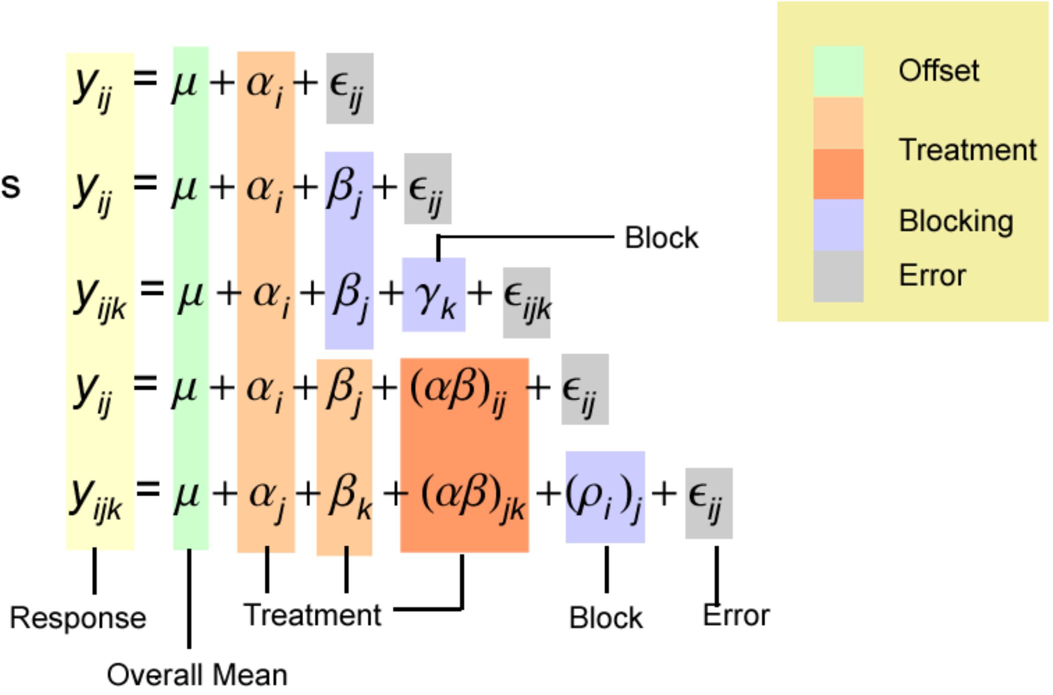
The Smart Design of Animal Experiments

VII. The Statistical Analysis

The Statistical Triangle
Design, *Model*, and *Analysis* are Interdependent



Every experimental design is underpinned by a statistical model



Completely Randomized $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$

Randomized Complete Blocks $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$

Latin Squares Design $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$

Factorial Design $y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$

Split Plot Design $y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + (\rho_i)_j + \epsilon_{ij}$

Statistical analysis:

- Fit model to data
- Compare treatment effect(s) with error

Significance tests

Use of probability for support of a scientific hypothesis

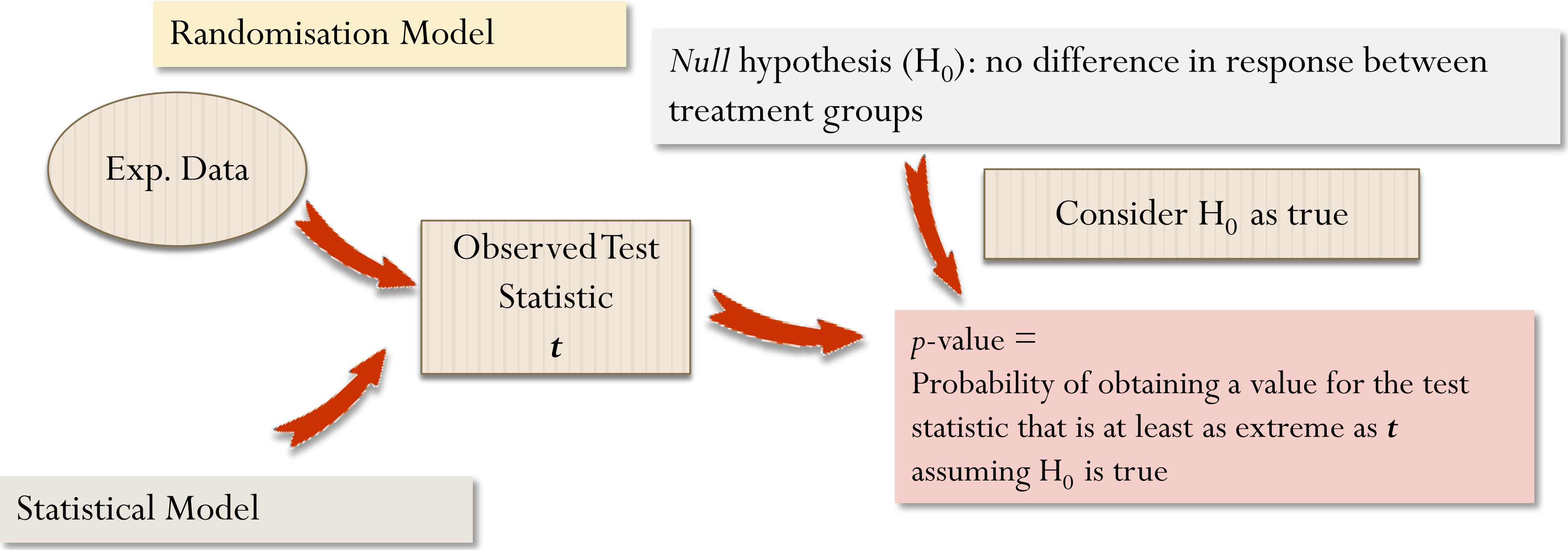
Pierre Simon Laplace
Tidal effect of the moon on the earth's atmosphere (1827)
$P(T>t|H_0)$

Ronald Fisher
Statistical Methods for Research Workers (1925)
$P(T>t|H_0)$ – strength of evidence

Jerzy Neyman – Egon Pearson
Statistical decision theory (1928)
H_0 versus H_1

Significance tests – The *p*-value

The concept of the *p*-value is often misunderstood



Randomisation Model

Null hypothesis (H_0): no difference in response between treatment groups

Exp. Data

Observed Test Statistic ***t***

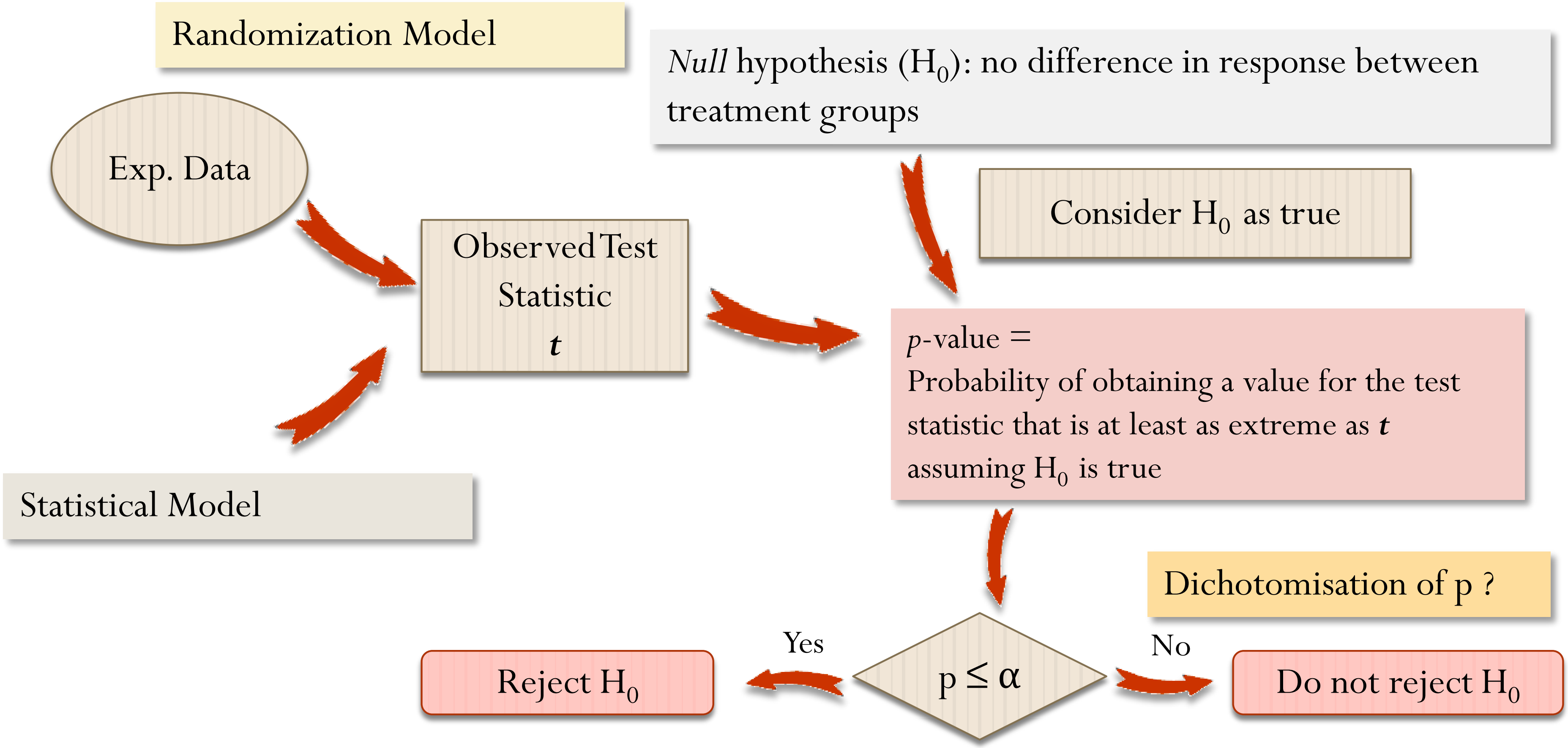
Consider H_0 as true

p-value = Probability of obtaining a value for the test statistic that is at least as extreme as ***t*** assuming H_0 is true

Statistical Model

Significance tests – Hypothesis test

The concept of the *p*-value is often misunderstood



Randomization Model

Exp. Data

Observed Test Statistic ***t***

Statistical Model

Null hypothesis (H_0): no difference in response between treatment groups

Consider H_0 as true

p-value =
Probability of obtaining a value for the test statistic that is at least as extreme as ***t*** assuming H_0 is true

Dichotomisation of p ?

$p \leq \alpha$

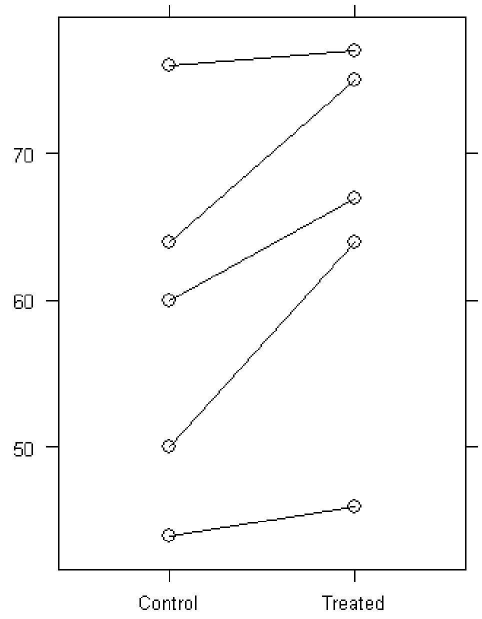
Yes

Reject H_0

No

Do not reject H_0

Significance tests
Example Cardiomyocytes (paired exp.)

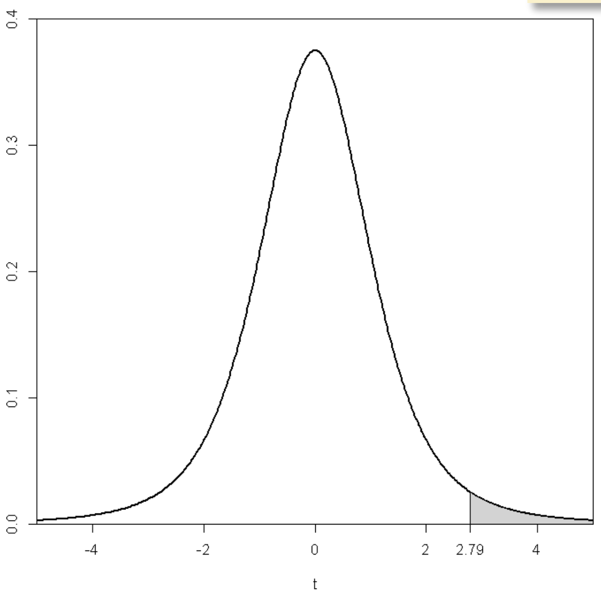


Mean diff. = 7% viable MC
standard error = 2.51

Test statistic
$t = 7/2.51 = 2.79$

Assume H_0 is true

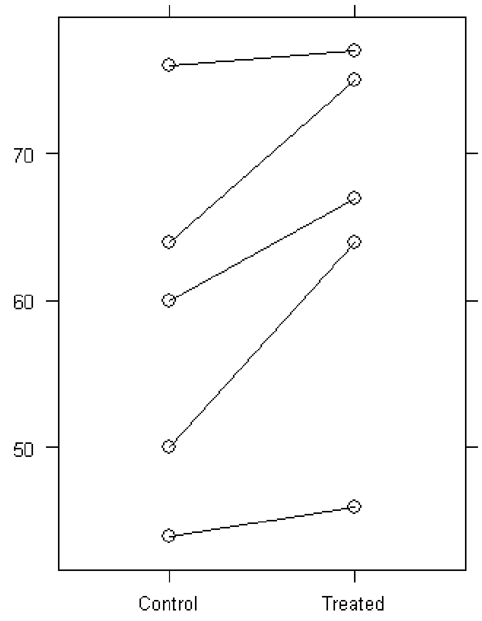
Distribution of possible values of t with mean 0



$P(t \geq 2.79) = 0.024$

Probability of obtaining an increase of 2.79 or more is 0.024, provided H_0 is true

Significance tests
Two-sided tests

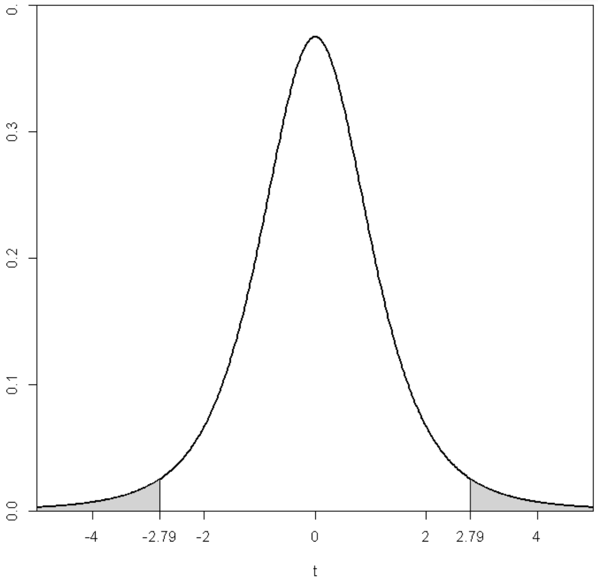


Mean diff. = 7% viable MC
standard error = 2.51

Test statistic
t = 7/2.51 = 2.79

Assume H_0 is true

Distribution of possible values of t with mean 0



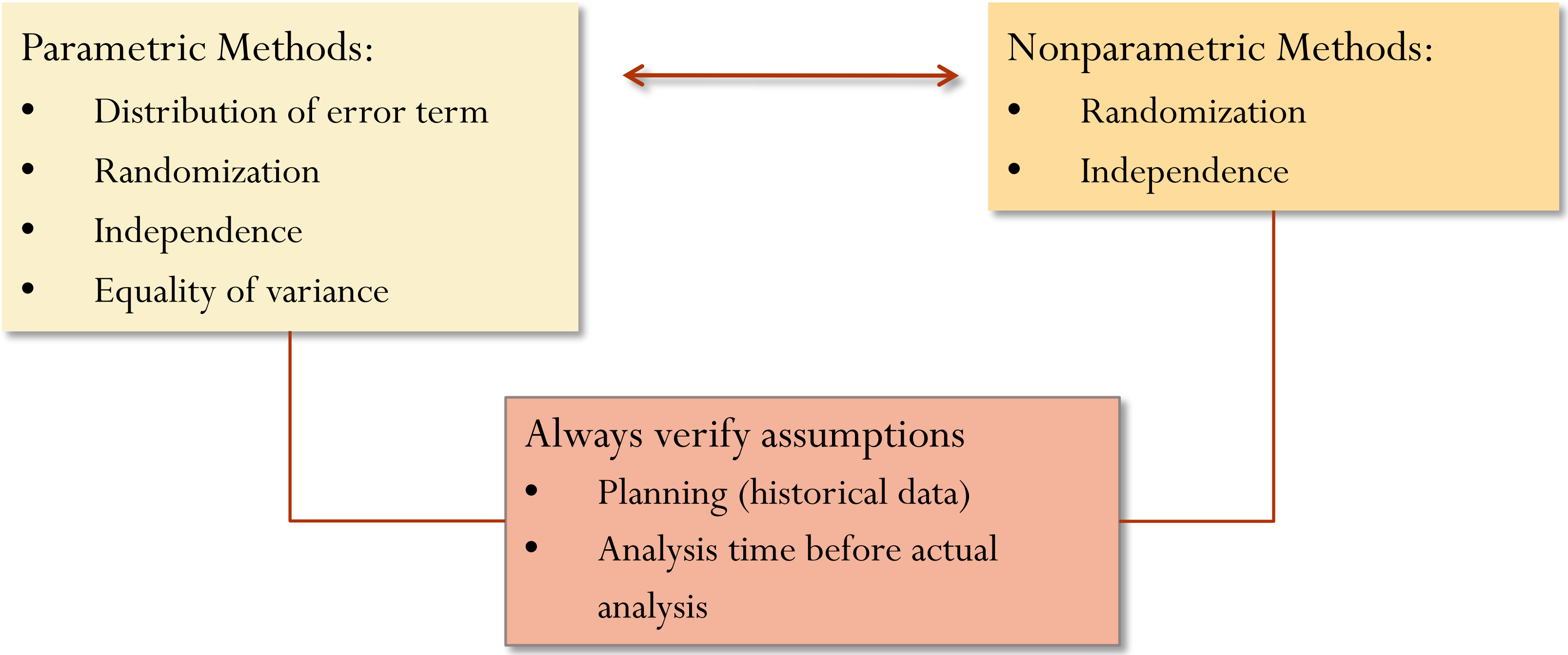
Results in opposite direction are also of interest

$P(|t| \geq 2.79) = 0.048$

Probability of obtaining a difference of ±2.79 or more is 0.048, provided H_0 is true

Two-sided tests are the rule!

Statistics always makes assumptions



Parametric Methods:

- Distribution of error term
- Randomization
- Independence
- Equality of variance

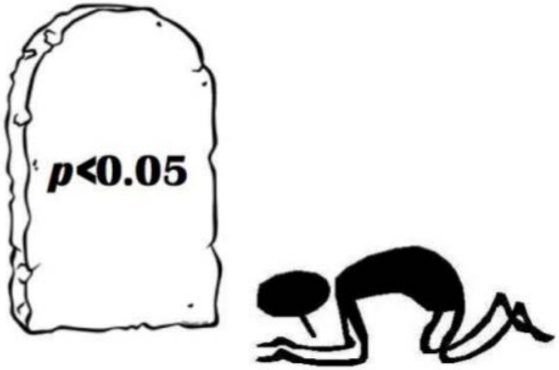
Nonparametric Methods:

- Randomization
- Independence

Always verify assumptions

- Planning (historical data)
- Analysis time before actual analysis

The p-value and statistical significance



Interpretation of p-value:

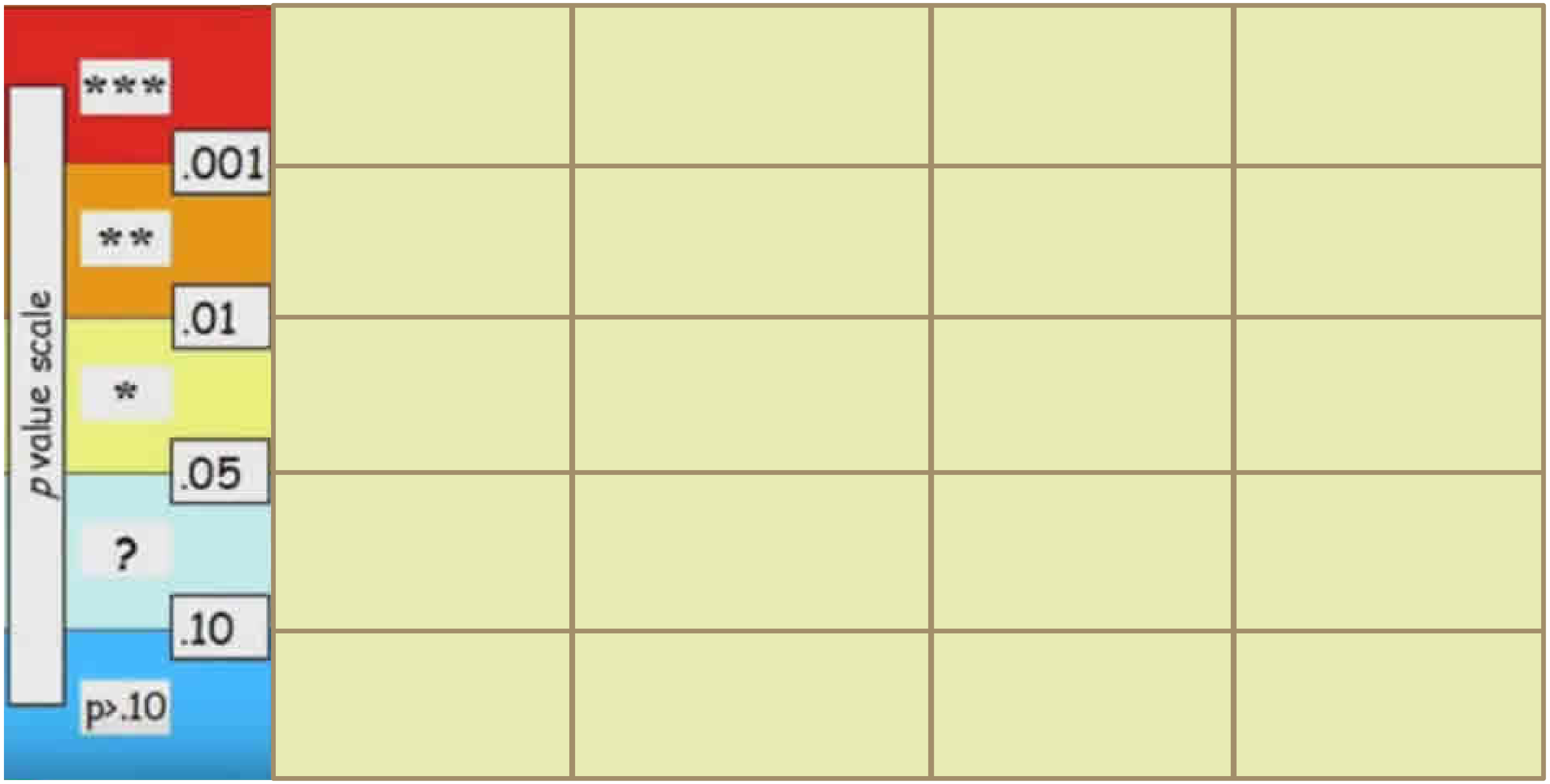
- A measure for the strength of evidence (Fisher)
- Given the data, the probability of obtaining a result as extreme or more extreme than the one observed, assuming that H_0 (no difference) is true.

> *"…it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation"*

- p> 0.1 does not mean no difference
- p-value not related to scientific relevance
- large enough samples make most uninteresting minimal difference significant

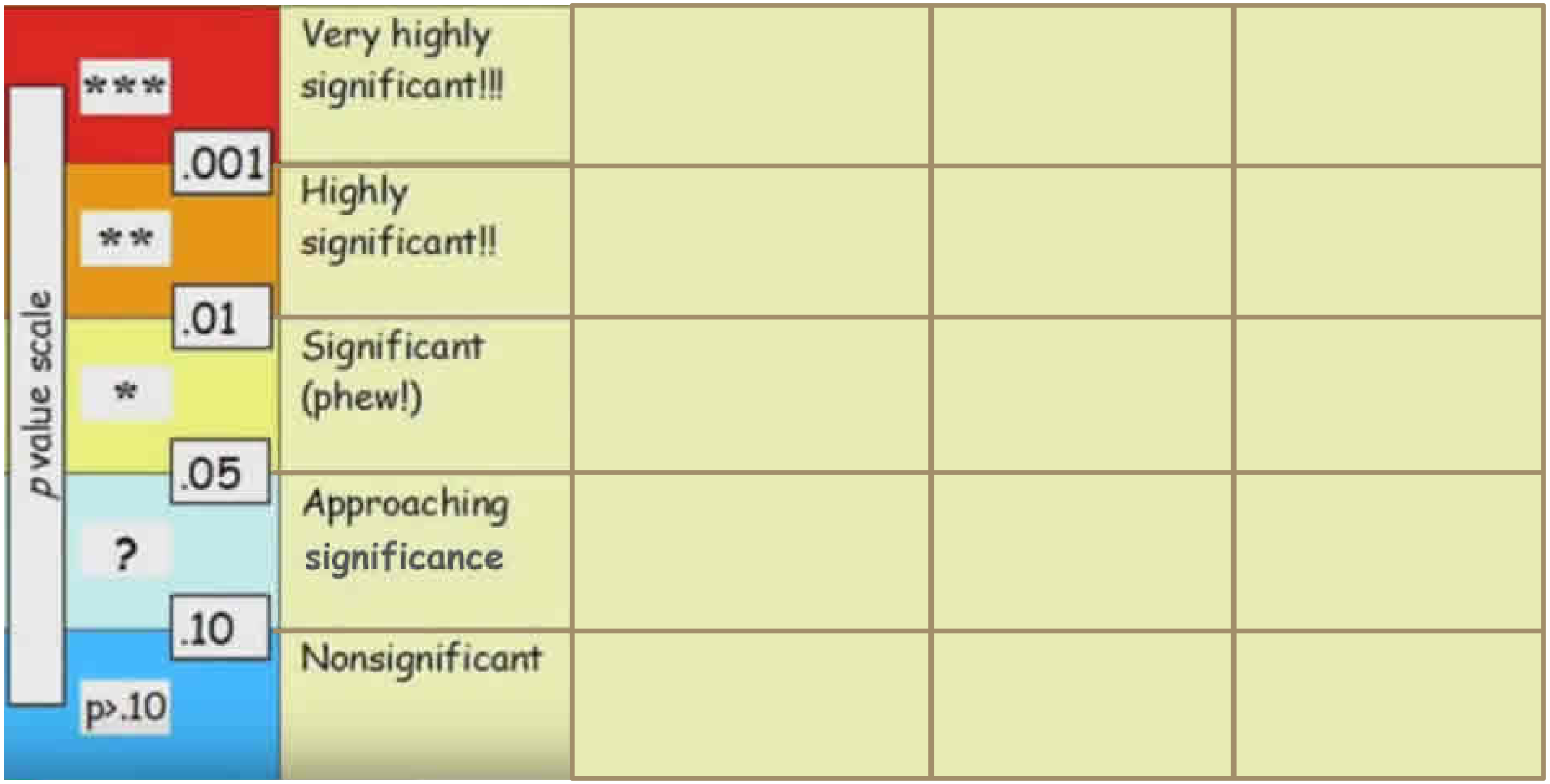
The meaning of statistical significance
The dance of the p-values (G. Cumming)

p as a measure of strength of evidence (Fisher)



The meaning of statistical significance
The dance of the p-values (G. Cumming)

… that elicits a significance language

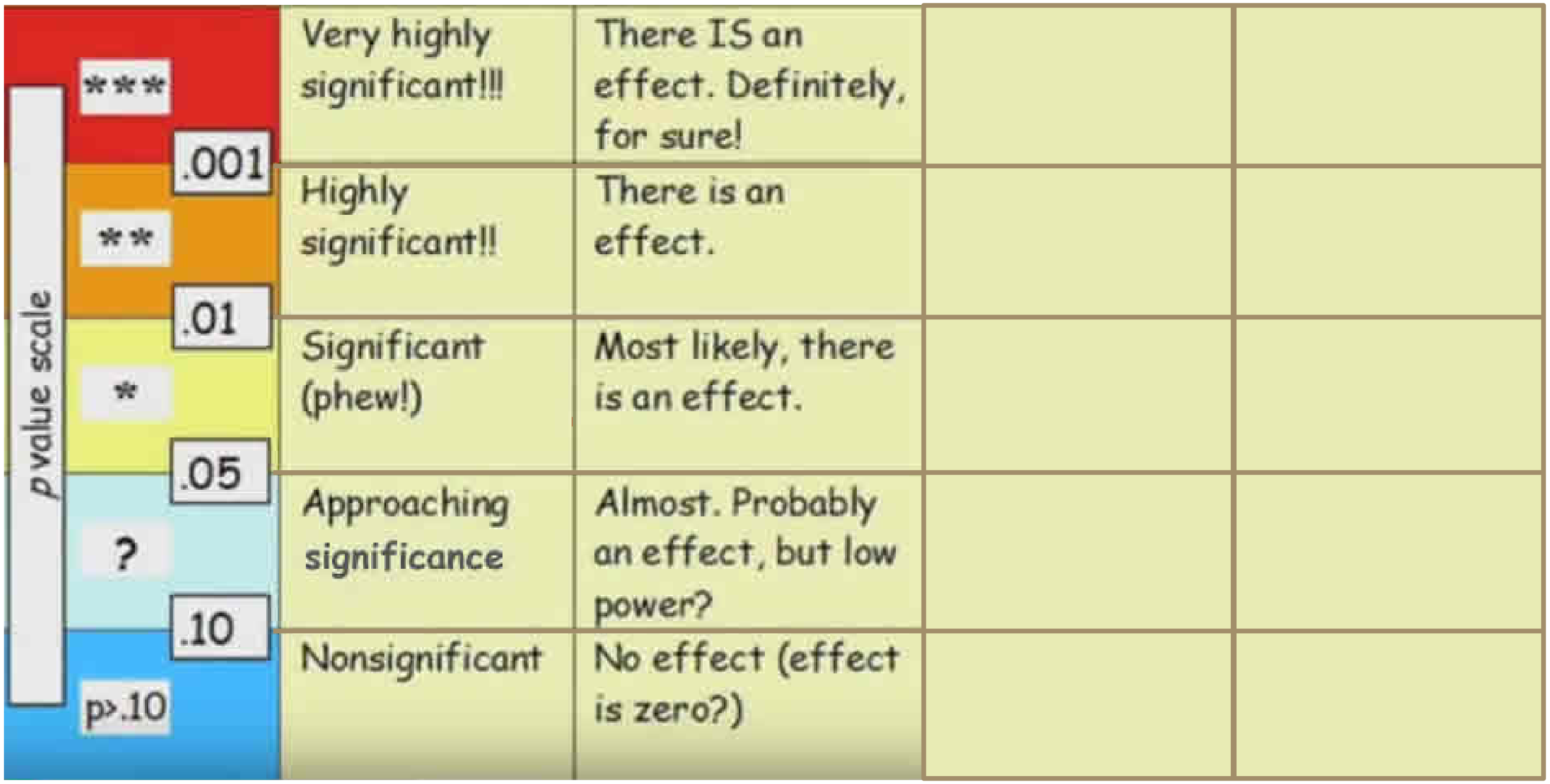


p value scale

***	Very highly significant!!!			
.001				
**	Highly significant!!			
.01				
*	Significant (phew!)			
.05				
?	Approaching significance			
.10				
p>.10	Nonsignificant			

The meaning of statistical significance
The dance of the p-values (G. Cumming)

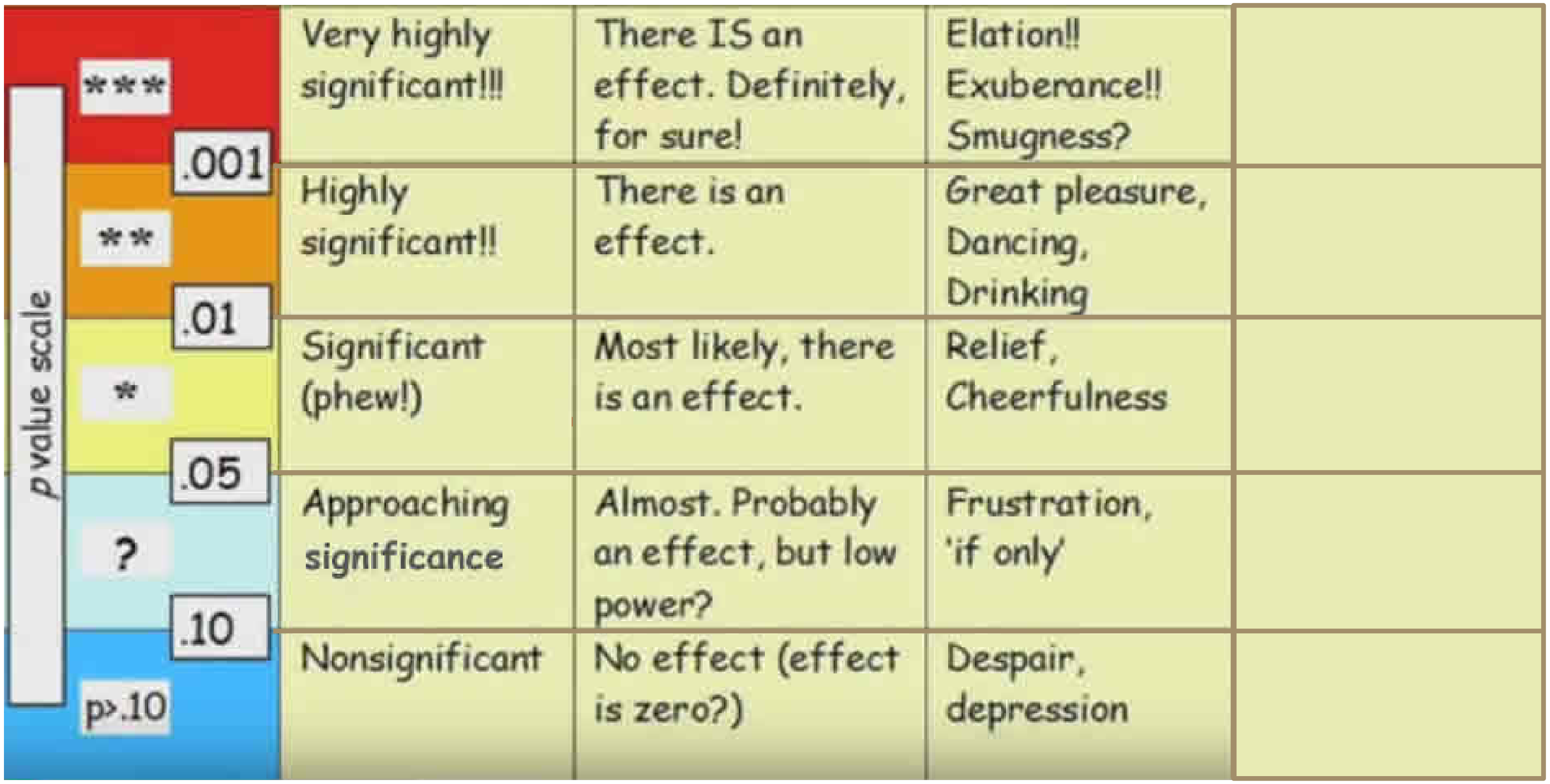
… which suggests *truth*



p value scale					
***	.001	Very highly significant!!!	There IS an effect. Definitely, for sure!		
**	.01	Highly significant!!	There is an effect.		
*	.05	Significant (phew!)	Most likely, there is an effect.		
?	.10	Approaching significance	Almost. Probably an effect, but low power?		
p>.10		Nonsignificant	No effect (effect is zero?)		

The meaning of statistical significance
The dance of the p-values (G. Cumming)

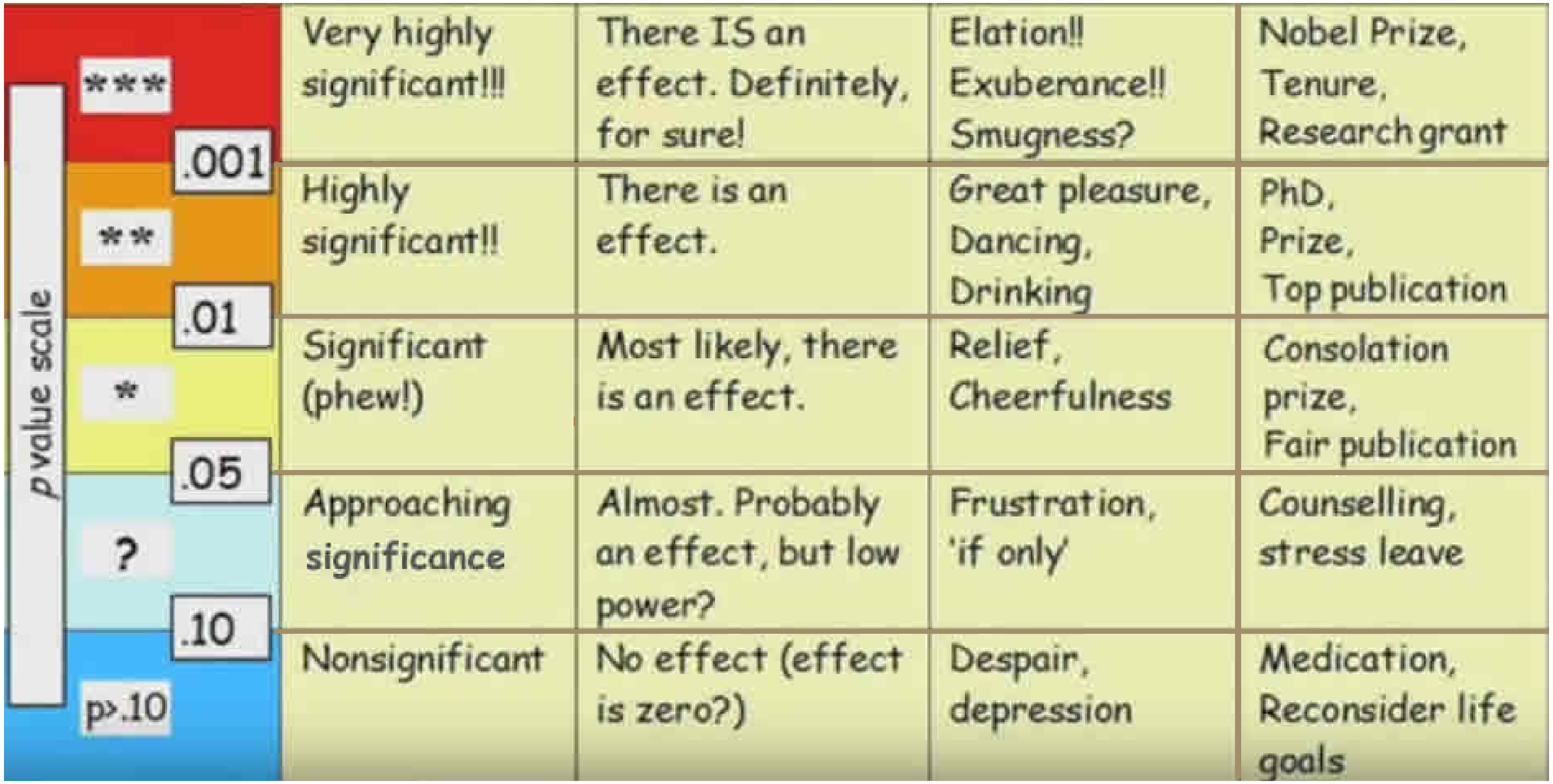
… evokes emotion



p value scale					
***	.001	Very highly significant!!!	There IS an effect. Definitely, for sure!	Elation!! Exuberance!! Smugness?	
**	.01	Highly significant!!	There is an effect.	Great pleasure, Dancing, Drinking	
*	.05	Significant (phew!)	Most likely, there is an effect.	Relief, Cheerfulness	
?	.10	Approaching significance	Almost. Probably an effect, but low power?	Frustration, 'if only'	
p>.10		Nonsignificant	No effect (effect is zero?)	Despair, depression	

The meaning of statistical significance
The dance of the p-values (G. Cumming)

… and has real-life consequences

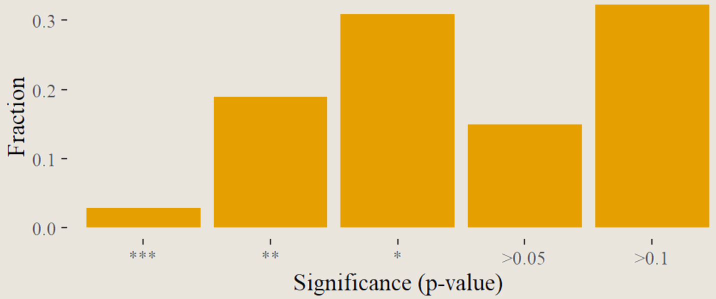


p value scale					
***		Very highly significant!!!	There IS an effect. Definitely, for sure!	Elation!! Exuberance!! Smugness?	Nobel Prize, Tenure, Research grant
**	.001	Highly significant!!	There is an effect.	Great pleasure, Dancing, Drinking	PhD, Prize, Top publication
*	.01	Significant (phew!)	Most likely, there is an effect.	Relief, Cheerfulness	Consolation prize, Fair publication
?	.05	Approaching significance	Almost. Probably an effect, but low power?	Frustration, 'if only'	Counselling, stress leave
p>.10	.10	Nonsignificant	No effect (effect is zero?)	Despair, depression	Medication, Reconsider life goals

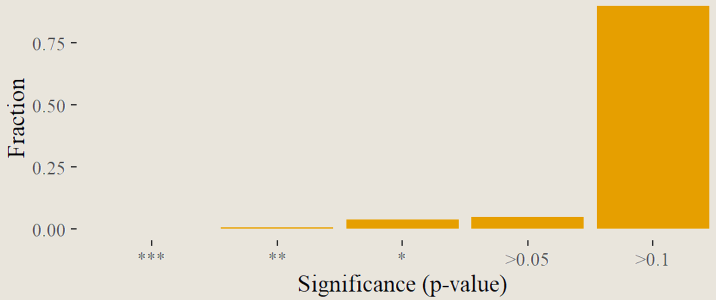
The meaning of statistical significance
The dance of the p-values (G. Cumming)

Can we trust p?

- Two group experiment
- n =7, Δ = 1.2, power = 54%
- 10,000 simulations
- $p \begin{cases} <0.001 \\ 0.986 \end{cases}$



- n =12, Δ = 0
- 10,000 simulations
- $p \begin{cases} <0.001 \\ 1.000 \end{cases}$



The meaning of statistical significance

Can we trust p?

- Experiment yields two-sided $p_{obt} = 0.03$
- Replication experiment
- $P(p_{rep} < 0.05 \mid p_{obt} = 0.03)$?

The meaning of statistical significance

Can we trust p?

- Experiment yields two-sided $p_{obt} = 0.03$
- Replication experiment
- $P(p_{rep} < 0.05 \,|\, p_{obt} = 0.03)$ $= 0.586$

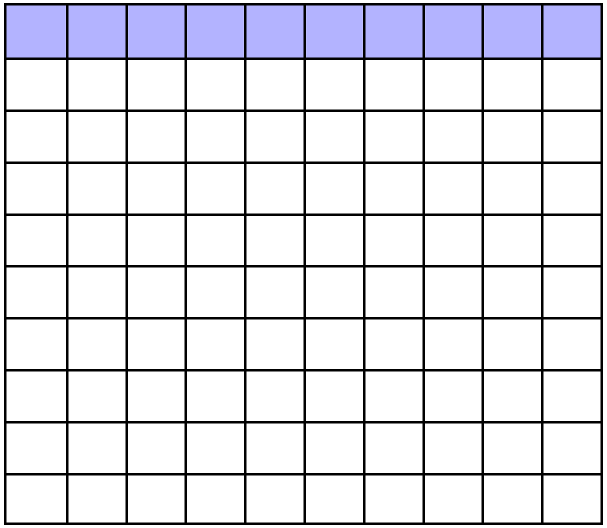
- Prediction intervals for one-tailed p-value p_{rep} in replication experiment, when p_{obt} was the two-tailed p-value in the initial experiment.
- One-sided prediction interval goes from 0 to 80% percentile
- Two-sided prediction interval goes from 10% to 90% percentile

Cumming, G. (2008). Persp Psychol Sci,3: 286-300

P_{obt}	One-sided p_{rep}	Two-sided p_{rep}
0.001	(0, 0.018)	(<0.001, 0.070)
0.01	(0,0.083)	(<0.001,0.22)
0.02	(0, 0.13)	(<0.001, 0.30)
0.05	(0,0.22)	(<0.001, 0.44)
0.1	(0,0.32)	(<0.001, 0.57)
0.2	(0,0.46)	(<0.001,0.70)
0.4	(0,0.64)	(0.004, 0.83)
0.6	(0,0.75)	(0.0098,0.90)

The meaning of statistical significance

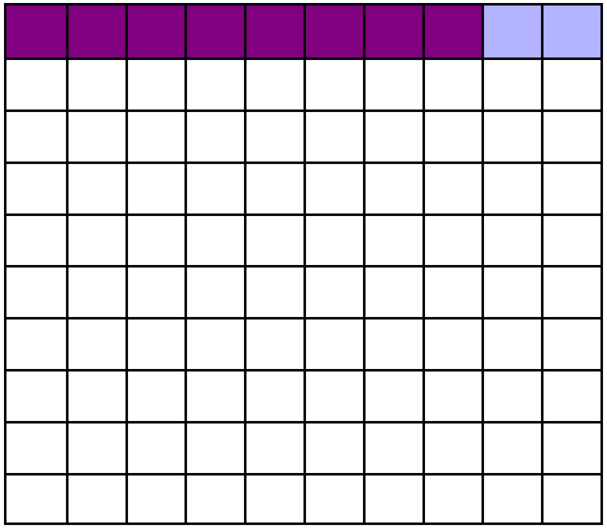
$p \leq 0.05$, what does it mean?



- 100 drugs tested against biological target
- 10% are known to be active
- Prevalence, $\pi = 0.1$

The meaning of statistical significance

$p \leq 0.05$, what does it mean?

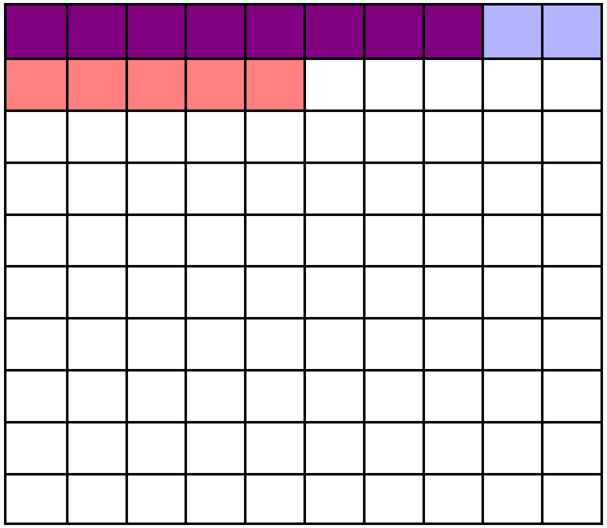


- 100 drugs tested against biological target
- 10% are known to be active
- Prevalence, $\pi = 0.1$

- Power = 80%, 8 active drugs determined

The meaning of statistical significance

$p \leq 0.05$, what does it mean?



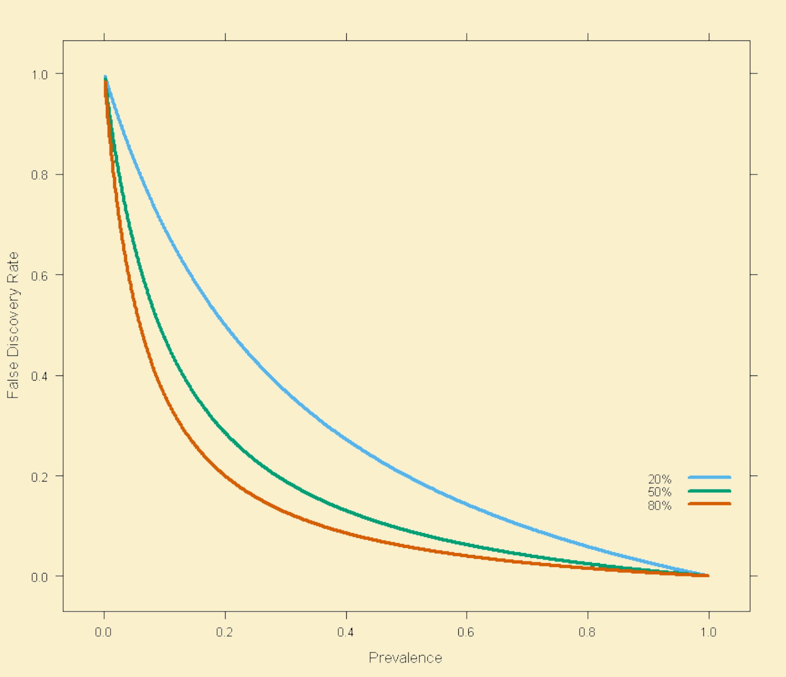
- 100 drugs tested against biological target
- 10% are known to be active
- Prevalence, $\pi = 0.1$
- Power = 80%, 8 active drugs correctly determined

- $\alpha = 0.05$, 5% of inactive drugs are falsely declared active
- 13 drugs declared active
- 8 drugs truly active
- PPV = 62% , FDR = 38%

The meaning of statistical significance

$$FDR = \frac{1}{1 + \frac{\pi}{1-\pi}\,\frac{1-\beta}{\alpha}}$$

- FDR depends highly on prevalence π
- Example, π = 1/10
  Power 80%, FDR = 36%
  Power 20%, FDR = 69%
- π = 1/100, power 80%, α = 0.05,
  FDR = 0.69, 69% findings are false
- Some journals pounce upon such unexpected findings
- FDR key factor for lack of replicability



What is the value of p ≈ 0.05 (0.045 – 0.05) ?

MFDR = minimum value FDR

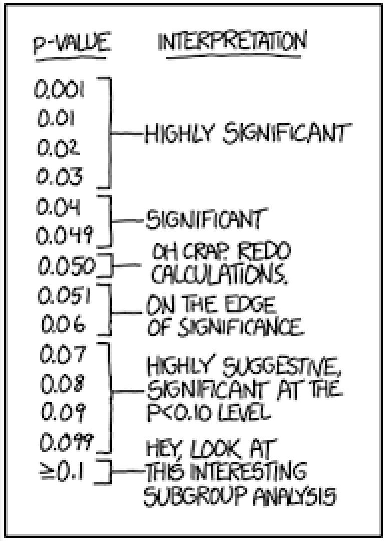
p = 0.05, MFDR = 0.289

p = 0.01, MFDR = 0.111

For p ≈ 0.05, in at least 29% of the cases there is no true effect

Dictatorship of significance
ASA statement March 6, 2016

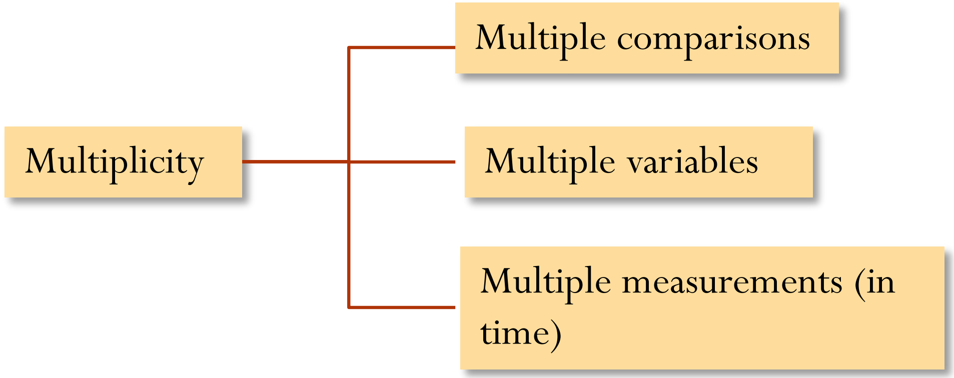
- *P*-values can indicate how incompatible the data are with a specified statistical model
- *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
- Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold
- Proper inference requires full reporting and transparency



P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

- A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result
- By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis

Multiplicity



	State of Nature	
Decision made	**Null hypothesis true**	**Alternative hypothesis true**
Do not reject null hypothesis	Correct decision (1 – α)	False negative β
Reject null hypothesis	False positive α	Correct decision (1 – β)

- 20 doses of a drug tested against control, each at a significance level α of 0.05
- Assume all null hypotheses are true, no dose differs from control
- P(correct decision of no difference) = 1 - 0.05 = 0.95
- P(ALL 20 decisions correct) = $0.95^{20} = 0.36$
- P(at least 1 mistake) =
  P(NOT ALL 20 correct) = 1 – 0.36 = 0.64

- The "Curse of Multiplicity" is of particular importance in microarray data (30,000 genes ⟶ 1,500 FP)
- Multiplicity and how to deal with it must be recognized at the planning stage

The Smart Design of Animal Experiments

VIII. The Study Protocol

The writing of the study protocol finalizes the research design phase

Research protocol

- **Rehearses research logic**
- Forms the basis for reporting

Technical protocol

- **Describes practical actions**
- Guidelines for lab technicians

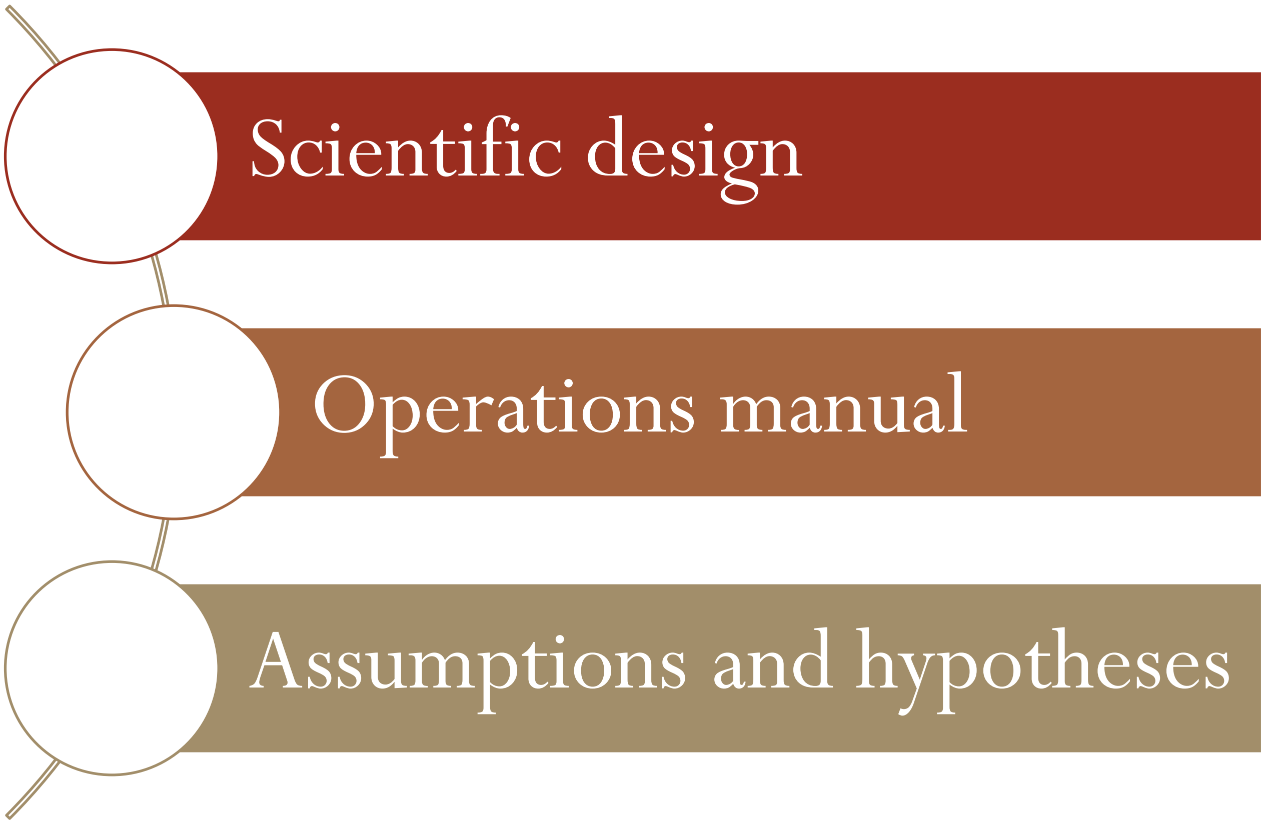
- General hypothesis
- Working hypothesis
- Experimental design rationale (treatment, error control, sampling design)
- Measures to minimize bias (randomization, etc.)
- Measurement methods
- Statistical analysis

prevent data dredging (data snooping) & p-value hacking

- Manipulation of animals
- Materials
- Logistics
- Data collection & processing
- Personal responsibilities

Writing the Study Protocol is already working on the Study Report

Crucial role of the protocol



Scientific design

Operations manual

Assumptions and hypotheses

The Smart Design of Animal Experiments

IX. The Research Report

The ARRIVE Guidelines

- Kilkenny (2009): issues with quality of reporting
- Kilkenny (2010): ARRIVE guidelines
- Accommodate most serious pitfalls in reporting of studies in animals
- List of 20 items to be included in scientific publications
- See Appendix E

The ARRIVE Guidelines
Introduction Section (items 3 & 4)

Requirements about scientific background, experimental approach and rationale, primary and secondary hypotheses

→ Study Protocol

ARRIVE Guidelines
Methods Section – 6. Study Design

- Number & size of experimental groups and control group
- Weaknesses and strengths of the study design
- use of randomization, blinding, etc.
- Was blocking used?
  Reasons for blocking and blocking factors
- Statistical analysis of blocking
- Ambiguity about experimental unit, unit used in the statistical analysis (single animal, litter, cage, etc.)
- Justification of choice of EU

Animals housed as a group !

ARRIVE Guidelines
Methods Section - 10. Sample Size

- Specify total number of animals used in each experiment and each experimental group
- How was the total number of animals decided, details of sample size calculations, consider a factorial design to reduce sample size
- Multiple statistical tests increase risk of finding false positives
- How many independent experiments were carried out, did they confirm each other or not

Factorial experiments allow to investigate treatment effects under different circumstances

ARRIVE Guidelines
Methods Section – 11. Allocation of animals

- Full details on how animals were allocated to experimental groups
- Randomisation? Matching?
- Randomisation **must** be used after choosing suitable experimental design

❗ Order of treatment or assessment must be random, else systematic bias can be introduced

ARRIVE Guidelines
Methods Section – 13. Statistical Methods

- Should be described in enough detail to enable reader to verify results
- Report & Justify the methods
- “*tests of significance*”is too vague
- Level of significance and direction (one-sided, two-sided)
  e.g. two-sided p-values smaller than or equal to 0.05 were considered to indicate statistical significance.
- Multiplicity? How was it dealt with?
- Specify unit of analysis in each dataset (single animal, group, cage, single cell)
- How was the aptness of the statistical model diagnosed, i.e. assessment whether data met assumptions of statistical test.
- Software and version used in statistical analysis

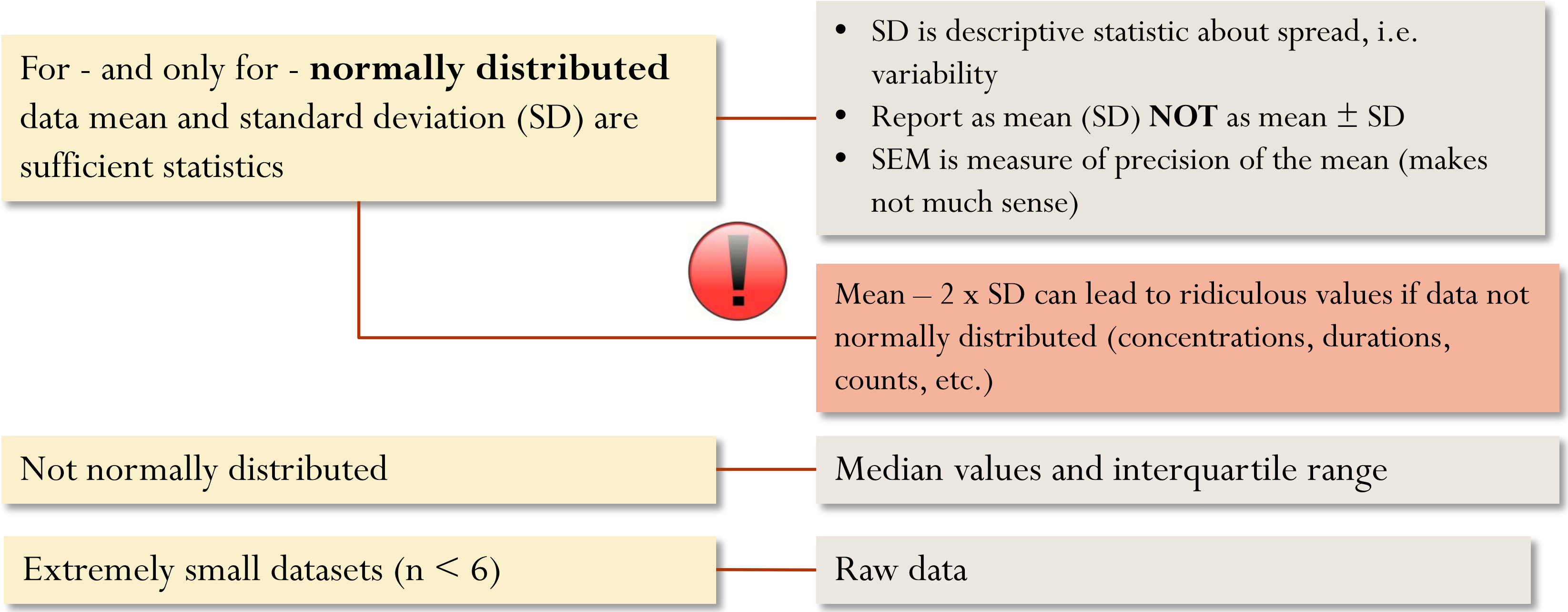
ARRIVE Guidelines
Results Section – 15. Numbers Analyzed

Always report number of experimental units analyzed in each group

- Explain why animals were excluded from analysis and exclusion criteria
- Explain discrepancies with number randomized

ARRIVE Guidelines
Results Section – 16. Outcomes and Estimation

Present findings with appropriate indicators of uncertainty



For - and only for - **normally distributed** data mean and standard deviation (SD) are sufficient statistics

- SD is descriptive statistic about spread, i.e. variability
- Report as mean (SD) **NOT** as mean ± SD
- SEM is measure of precision of the mean (makes not much sense)

Mean – 2 x SD can lead to ridiculous values if data not normally distributed (concentrations, durations, counts, etc.)

Not normally distributed

Median values and interquartile range

Extremely small datasets (n < 6)

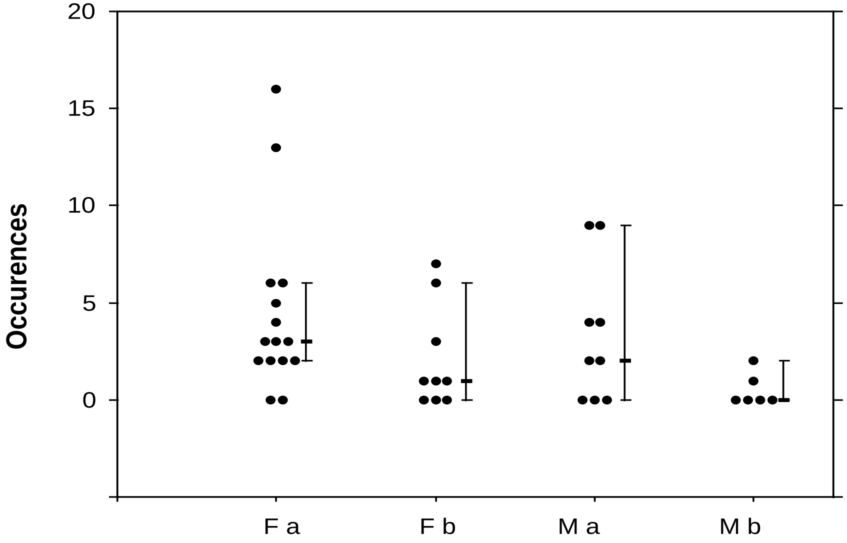
Raw data

Spurious precision detracts a paper's readability and credibility

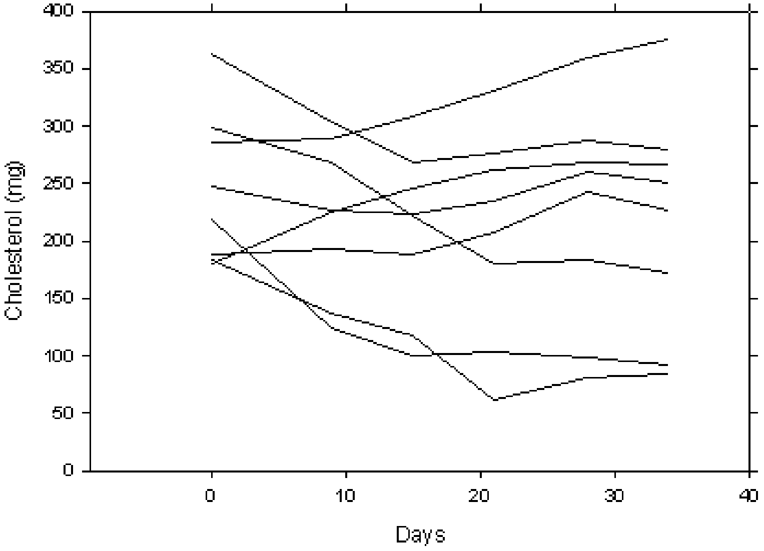
Points to consider in the Results Section Graphical Displays

- Complement tabular presentations
- Better suited for identifying patterns:

  "A picture says more than a thousand words"
- Whenever possible display individual data (e.g. use of *beeswarm* package in R)



Individual data, median values and 95% confidence intervals



Longitudinal data (measurements over time)
Individual subject profiles

Points to consider in the Results Section
Percentage of Control – A common misconception

- Scientist often prefer to re-express data as percent of control
- Example Experiment:
  Effect of 3 treatments Control, A, B **independent** groups of 6 mice each

	Control	A	B
	80	99	52
	106	73	18
	75	72	61
	79	94	61
	54	85	29
	84	62	49
Mean	79.7	80.8	45.0
SD	16.69	14.25	17.67

- Divide all data by mean value of control (79.7) and express as %
- Calculate mean values and SD of % data

	A	B
	124	65
	92	23
	90	77
	118	77
	107	36
	78	61
Mean	101.4	56.5
SD	17.88	22.19

Points to consider in the Results Section
Percentage of Control – A common misconception

- Scientist often prefer to re-express data as percent of control

- Example Experiment:
  Effect of 3 treatments Control, A, B **independent** groups of 6 mice each

	Control	A	B
	80	99	52
	106	73	18
	75	72	61
	79	94	61
	54	85	29
	84	62	49
Mean	79.7	80.8	45.0
SD	16.69	14.25	17.67

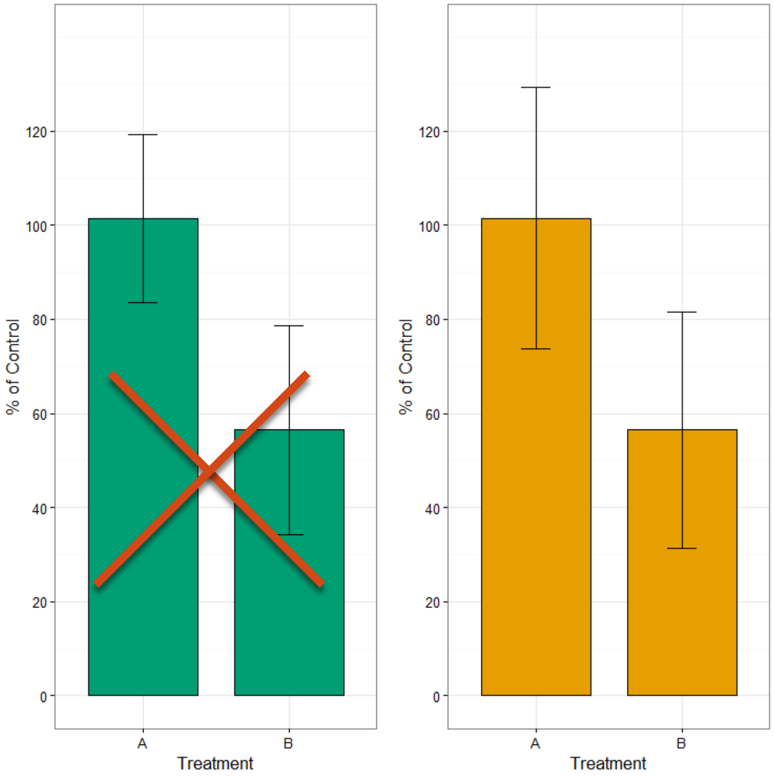
- Divide all data by mean value of control (79.7) and express as %

- Calculate mean values and SD of % data

	A	B
	124	65
	92	23
	90	77
	118	77
	107	36
	78	61
Mean	101.4	56.5
SD	17.88	22.19

Ignore variability in control group

Points to consider in the Results Section
Percentage of Control – A common misconception



$$\sigma_{X/Y} = \sqrt{\frac{1}{\mu_Y^2}\sigma_X^2 + \frac{\mu_X^2}{\mu_Y^4}\sigma_Y^2}$$

Assumes normality of X/Y

Points to consider in the Results Section Significance Tests

- Specify method of analysis do not leave ambiguity

e.g. "*statistical methods included ANOVA, regression analysis as well as tests of significance*" in the methods section is not specific enough

- Tests of significance should be two-sided

- Two group tests allow **direction** of alternative hypothesis, e.g. treated > control, this is a one-sided alternative
- Use of one-sided tests must be justified and the direction of the test must be stated beforehand (in protocol)
- For tests that allow a one-sided alternative it must be stated whether two-sided or one-sided was used

- Report exact p-values rather than $p<0.05$ or NS

- 0.049 is significant, 0.051 not ?
- Allows readers to choose their own critical value
- Avoid reporting $p = 0.000$; but report $p<0.001$ instead
- Report p-values to the third decimal
- For one-sided tests if result is in "wrong" direction then $p' = 1 - p$

Points to consider in the Results Section
Significance Tests – Confidence Intervals

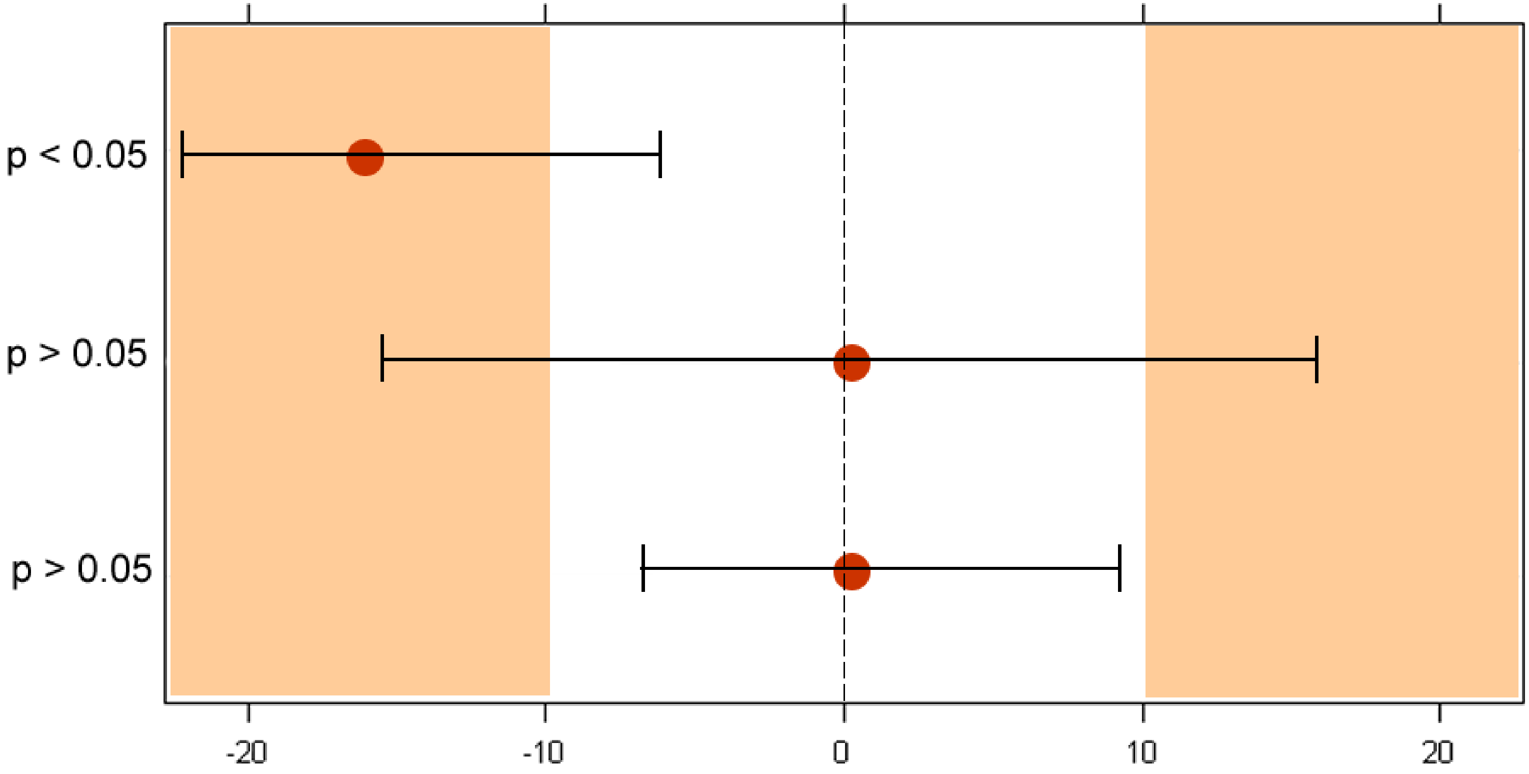
Statistical significance ≠ Biomedical Importance

Provide **confidence intervals** to interpret size of effect

Lack of significance ≠ Null hypothesis true (i.e. no effect)

- Null hypothesis is never proved
- Lack of evidence is no evidence for lack of effect
- **Confidence intervals** provide a region of plausible values for the treatment effect

Points to consider in the Results Section
Significance Tests – Confidence Intervals

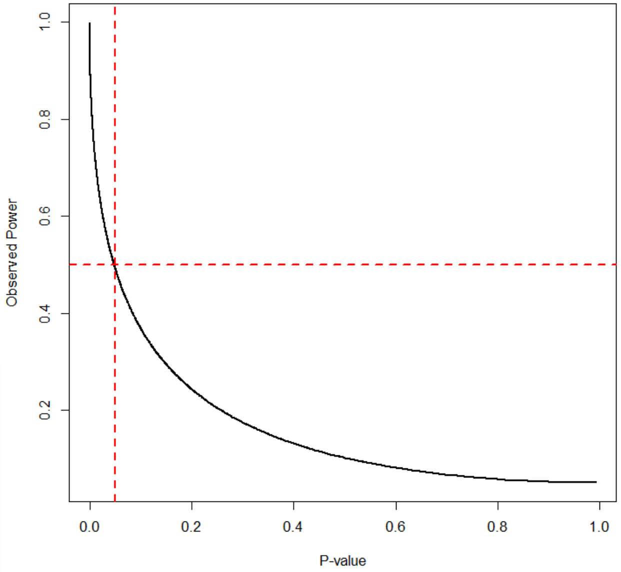


Post-hoc Power Calculations

Power calculated retrospectively using observed treatment effect

- Advocated by many authors
- Requested by some journal editors
- Present in software (e.g. SPSS)

- There is a perfect relationship between obtained *p*-value and observed power
- Large reported *p*-values will always have low observed power
- Only a priori power or sample size calculations make sense



Never perform an observed or post-hoc power analysis, even if an editor requests it

X is significant, while Y is not
False claims of interaction

“The percentage of neurons showing cue-related activity increased with training in the mutant mice (P<0.05), but not in the control mice (P>0.05)”

Invalid reasoning from nonsignificant result (P>0.05)

The difference between “significant” and “not significant” is not itself significant
Test for equalitiy of **effect** of training in 2 groups

Only a **factorial experiment** with proper method of analysis (ANOVA) and test for interaction can make such a claim

The Smart Design of Animal Experiments

X. Case Studies

The Smart Design of Animal Experiments

XI. Concluding Remarks

Concluding Remarks

- Biomedical research struggles with many problems of replicability, reproducibility, acceptance and efficiency
- Statistical thinking provides a conceptual framework and generic tools to deal with these problems
- Statistical thinking provides tools to design efficient insightful experiments
- Statistical thinking urged us to critically think about the outcome of the statistical analysis

Concluding Remarks

Statistical thinking is the key to succesful experiments

1. Time spent thinking on the conceptualization and design of an experiment is time wisely spent
2. The design of an experiment reflects the contributions from different sources of variability
3. The design of an experiment balances between its internal validity and external validity
4. Good experimental practice provides the clue to bias minimization
5. Good experimental design is the clue to the control of variability
6. Experimental design integrates various disciplines
7. A priori consideration of statistical power is an indispensable pillar of an effective experiment

Role of statistician?

- Professional skilled in solving research problems
- Team member, collaborator, possible co-author
- Consult when doubt about design, sample size, or statistical analysis
- Include statistician from start of project

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Fisher about the Role of the Statistician

"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."

(Presidential Address to the First Indian Statistical Congress, 1938; Fisher, 1938).

Teşekkür ederim