Designing Experiments

Oxford Reproducibility School

Stanley E. Lazic

stan.lazic@cantab.net
https://stanlazic.github.io

26 Sept 2018

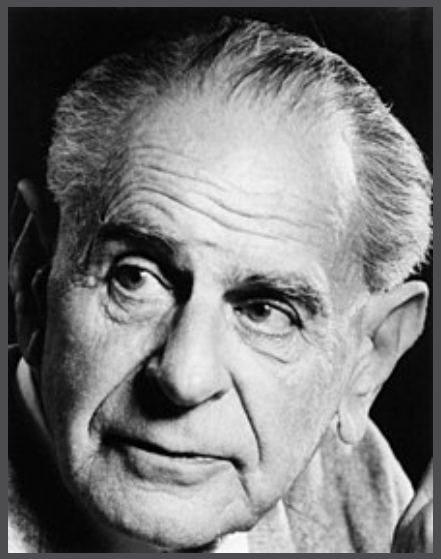
Topics covered

- Experimental aims (it's not all hypothesis testing)
- Properties of a good experiment
- Fundamental Experimental Design Equation
- A mental model for experimental designs

What do scientists do? Philosophers concluded:

"A scientist, whether theorist or experimenter, puts forward statements, or systems of statements, and tests them step by step. In the field of the empirical sciences, more particularly, he constructs hypotheses, or systems of theories, and tests them against experience by observation and experiments."

– Karl Popper (1934/1959)



...and statisticians provided the tools:

Fisher

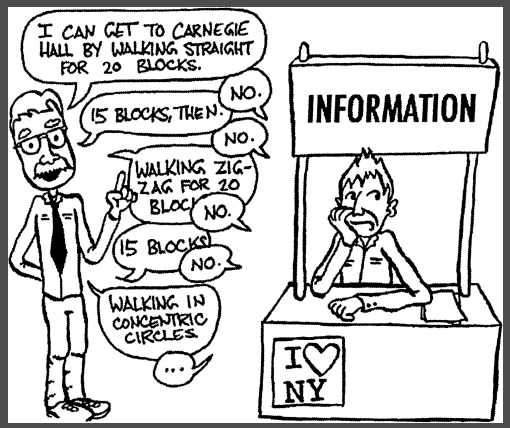
Pearson

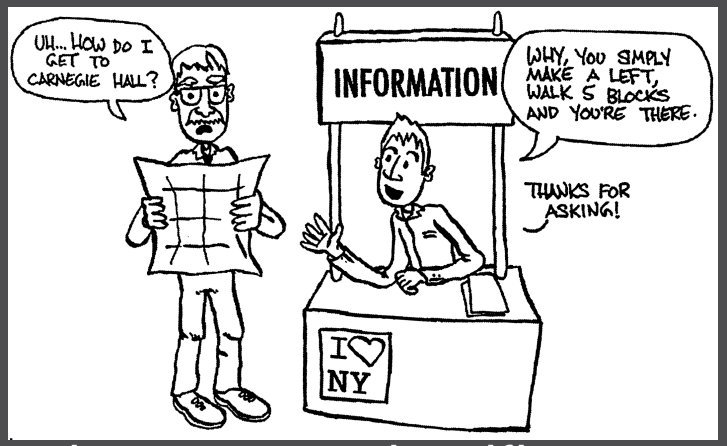
Neyman

Hypothesis tests, p-values, H_0, H_1, power, Type I and Type II errors, etc.

What's wrong with this?

- For Popper, scientist = physicist. But other sciences are different.
- People believed Popper $\rightarrow$ to be scientific we must construct and test hypotheses!
- Science is more than hypothesis testing (which is best used at the end of a long process of discovery).
- We don't learn in everyday life by proposing and rejecting hypotheses $\rightarrow$ inefficient way to learn.





What is the consequence of shoehorning every scientific problem into a hypothesis testing framework?

Glass DJ (2007)

Other scientific activities:

- Describing
- Estimating
- Predicting
- Optimising
- Learning

These activities have different objectives and influence the experimental design, methods for determining an appropriate sample size, and what is reported.

Other scientific activities: Describing

Aim is to describe the data without making inferences to a larger population (maybe you have the whole population).

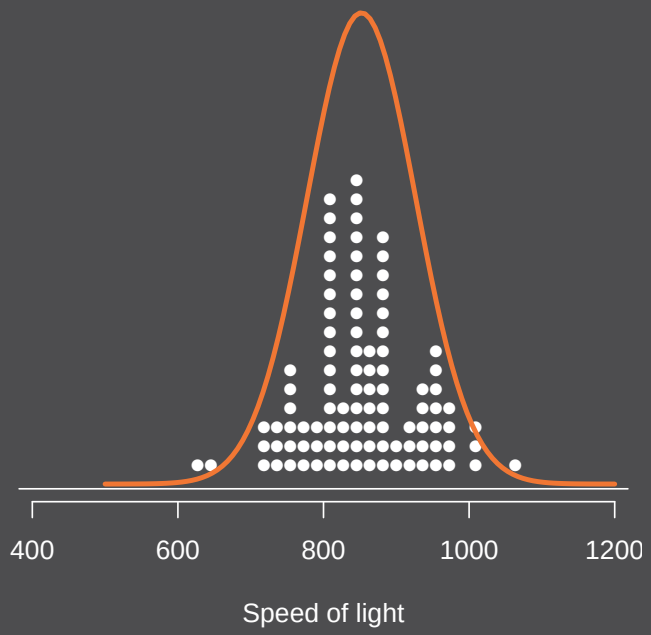
- How many cyclists passed by a given point today?
- How many Catholics are there in Scotland (census)?
- Demographic or baseline characteristics of participants in a randomised experiment.
- The feeding behaviour of a new species of beetle.
- The number of times Shakespeare uses the word “thou”.

www.cambridge-news.co.uk/news/cambridge-news/why-is-cambridge-cycling-city-14475326

Other scientific activities: Estimating

Aim is to determine the value of something when you don't have the whole population, or when there is measurement error.

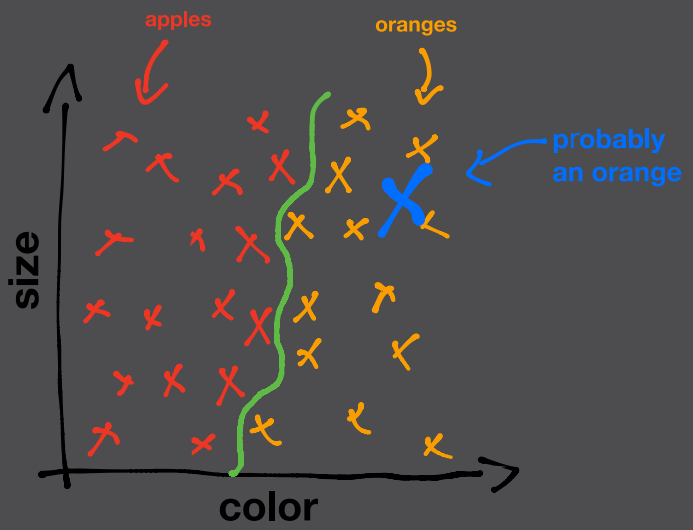
- What is the speed of light?
- How many badgers are there in East Anglia?
- What proportion of fish are infected with parasites?
- What concentration of a drug gives half the maximal response?



Other scientific activities: Predicting

Aim is to predict future observations or classify samples. Also known as “Machine Learning”.

- Predicting response to a treatment.
- Predicting if a drug will have toxic side effects.
- Calculating the probability of a flood in the next 5 years.

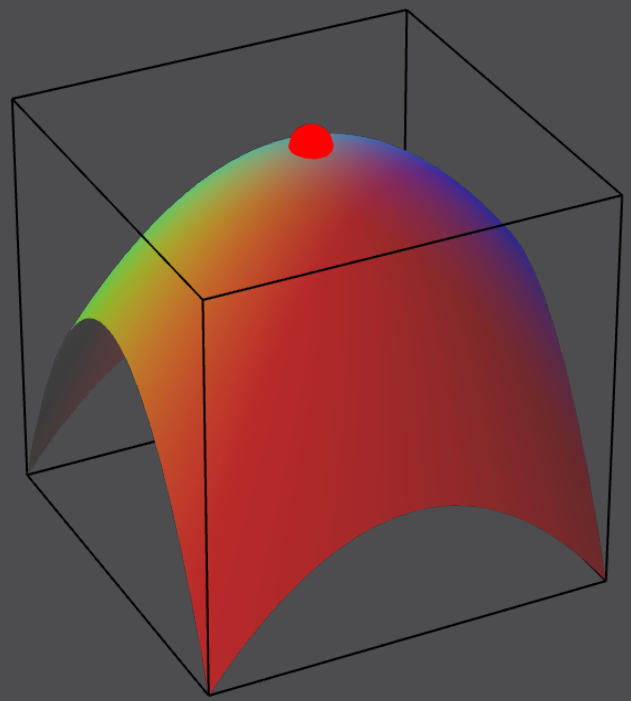


www.kdnuggets.com/2017/05/machine-learning-crash-course-part-1.html

Other scientific activities: Optimising

Aim is to find inputs (experimental interventions) that lead to a desirable output, such as a high, low, or target value of an outcome (e.g. response-surface models or Bayesian optimisation).

- Which reaction conditions maximise the yield of a chemical product?
- Which properties of the advert lead to the most clicks?
- Which stimulus settings give the largest effect?
- Which assay conditions give the largest assay window?

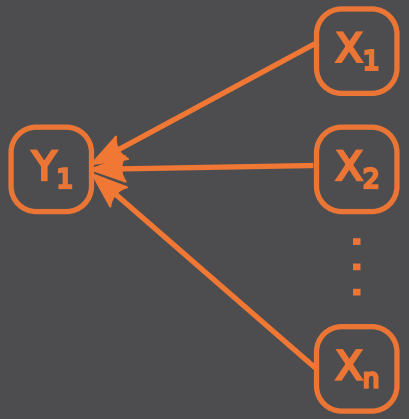


https://mathinsight.org/applet/local_maximum_two_variables

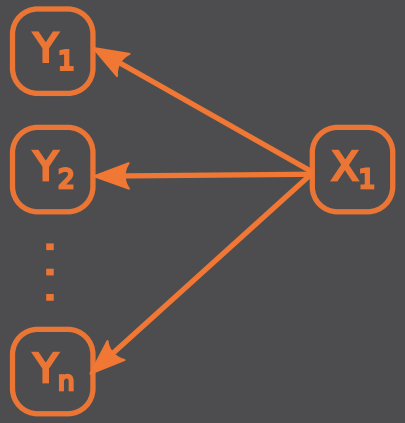
Other scientific activities: Learning

Aim is to learn as much as possible and come in two types:

1) Screening experiments: many inputs and one or few outputs
   - High-throughput drug discovery screen.
   - Finding the vital few vs. the trivial many (80/20 principle).



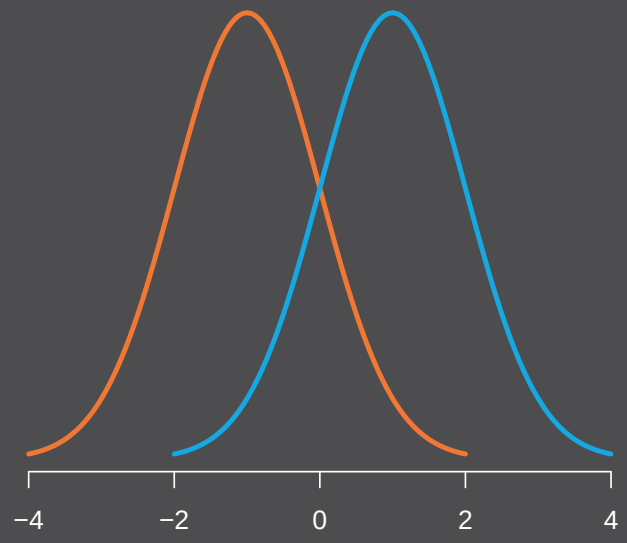
2) "-Omics" (in biology): one or few inputs and many outputs.
   - RNA-seq, proteomics, or metabolomics experiments.
   - Brain imaging and behaviour.



. . . and for completeness: Hypothesis testing (comparing)

Aim is to test a clearly defined hypothesis, either derived from theory or as a follow-up to a learning or optimising experiment (also accounting for background knowledge).

- Does 5 mg/kg of a drug given once a day for 5 days increase blood albumin levels in rats?
- Does 30 min of aerobic activity (treadmill running) at 60% VO_2 max, 3 days a week for 6 weeks, in males between 55–70 years of age, improve performance on a mental rotation task?



Aims influence the design and reporting

	Describe	Estimate	Predict	Optimise	Learn	Compare
Outcomes (y)	Few/Many	Few/Many	Few	Few	Few/Many	Few
Predictors (x)	NA	Few	Many	Few	Few/Many	Few
Design space	NA	Large	Large (diverse samples)	Large/small	Large	Small
Sample size/ Power	NA (all)	AIPE	Sample/ param ratio	AIPE, NA	Resource equation, AIPE, NA	Classic power calcs
Reporting	Descriptive stats	Estimate $\pm$ CI	Accuracy, AUC, MSE	Opt value, $\pm$ CI, x-values	Ranked lists (effect size)	P-vals, estimate $\pm$ CI

Digression: How big is a Gluebik?

Table 8.5 Items to Illustrate When Making Powers of Ten Posters

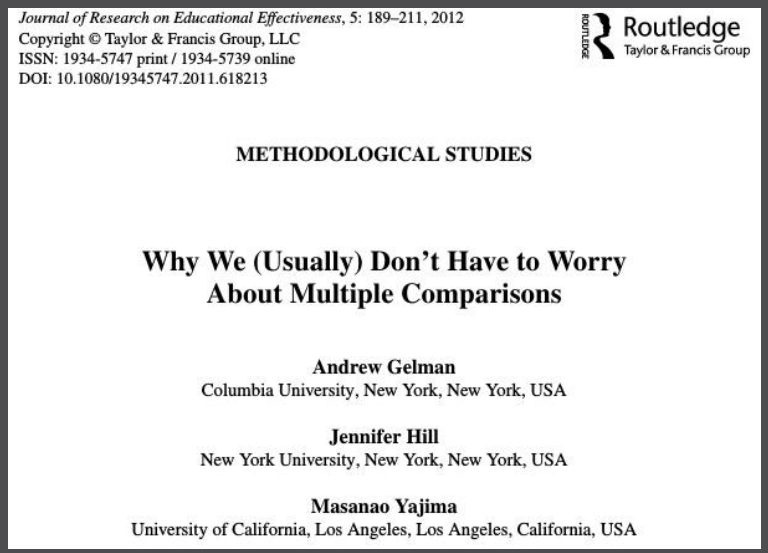
Distance	Comparison (Approximate)	Distance	Comparison (Approximate)
10^{0} m	Distance from floor to door knob	10^{26} m	Radius of observable universe
10^{-1} m	Width of hand	10^{25} m	Distance to the 3C273, brightest quasar
10^{-2} m	Width of fingernail on smallest finger	10^{24} m	Distance to the nearest large supercluster
10^{-3} m	Thickness of a U.S. dime	10^{23} m	Distance to galaxies beyond our local group
10^{-4} m	Length of a dust mite	10^{22} m	Distance to Andromeda galaxy
10^{-5} m	Diameter of human red blood cells	10^{21} m	Diameter of the disc of the Milky Way
10^{-6} m	Diameter of small bacteria	10^{20} m	Diameter of the Small Magellanic Cloud
10^{-7} m	Length of a virus	10^{19} m	Approximate thickness of the Milky Way
10^{-8} m	Thickness of bacteria flagellum	10^{18} m	Diameter of a typical globular cluster
10^{-9} m	Width of DNA helix	10^{17} m	Distance from Earth to Vega
10^{-10} m	Width of ice or quartz cell	10^{16} m	Inner radius of Oort cloud
10^{-11} m	Radius of a hydrogen atom	10^{15} m	100 × diameter of the solar system
10^{-12} m	Wavelength of X-rays	10^{14} m	10 × diameter of the solar system
10^{-13} m	Wavelength of an electron	10^{13} m	Diameter of solar system
10^{-14} m	Diameter of a nucleus	10^{12} m	Distance from Sun to Saturn
10^{-15} m	Diameter of a proton	10^{11} m	Distance from Sun to Venus
10^{-16} m	One-tenth the diameter of a proton	10^{10} m	One half the distance light travels in a minute
10^{-17} m	One-hundredth the diameter of a proton	10^{9} m	Diameter of the Sun
10^{-18} m	Radius of an electron	10^{8} m	Diameter of Saturn
		10^{7} m	North Pole to equator
		10^{6} m	Length of California (north to south)
		10^{5} m	Length of Connecticut (north to south)
		10^{4} m	Depth of Mariana Trench, deepest point
		10^{3} m	One kilometer; 2.5 times around a track
		10^{2} m	One side of a running track
		10^{1} m	Distance for a first down in football
		10^{0} m	Distance from floor to door knob

Increasing Size →

When you know little, a single data point tells you a lot!

Another digression: Don't need hypothesis testing for omics experiments

- *Hypothesising* that some genes out of 20K are differentially expressed is as scientific as predicting that "some famous people will die next year".
- It's better treated as an estimation problem.
- The problem of multiple hypothesis testing then becomes a problem of overfitting, for which there are good solutions (e.g. regularisation/shrinkage).


Journal of Research on Educational Effectiveness, 5: 189–211, 2012

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2011.618213

Routledge
Taylor & Francis Group

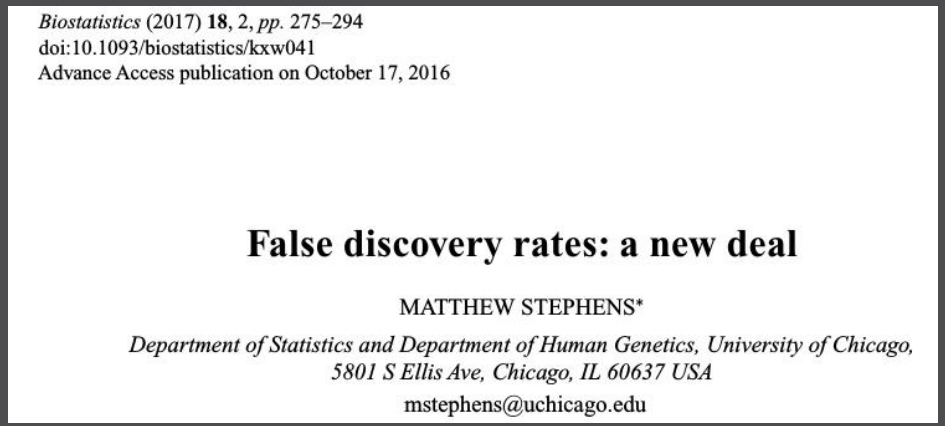
METHODOLOGICAL STUDIES

Why We (Usually) Don't Have to Worry About Multiple Comparisons

Andrew Gelman
Columbia University, New York, New York, USA

Jennifer Hill
New York University, New York, New York, USA

Masanao Yajima
University of California, Los Angeles, Los Angeles, California, USA


Biostatistics (2017) **18**, 2, *pp.* 275–294
doi:10.1093/biostatistics/kxw041
Advance Access publication on October 17, 2016

False discovery rates: a new deal

MATTHEW STEPHENS*

Department of Statistics and Department of Human Genetics, University of Chicago, 5801 S Ellis Ave, Chicago, IL 60637 USA

mstephens@uchicago.edu

Summary

- It's ok to do something other than hypothesis testing.
- A research programme will often have multiple aims, and should use diverse designs.
- Better to do several experiments with different goals than one big experiment to achieve multiple goals.

Properties of a good experiment

1) Effects can be estimated unambiguously and without bias.
2) Estimates are precise.
3) Estimates are protected from possible one-off events that might compromise the results.
4) The experiment is easy to conduct.
5) The data are easy to analyse and interpret.
6) Maximum information is obtained for fixed time, resources, and samples.
7) The findings are applicable to a wide variety of subjects, conditions, and situations.

$$1 + 3 + 4 + 7 \rightarrow \text{Reproducibility}$$

Fundamental Experimental Design Equation

$$\begin{aligned} \text{Outcome} = {} & \text{Treatment effects} + \\ & \text{Biological effects} + \\ & \text{Technical effects} + \\ & \text{Error} \end{aligned}$$

- Treatment effects caused by the manipulations and interventions of the experimenter.
- Biological effects arise from intrinsic properties of the samples or sample material.
- Technical effects arise from properties of the experimental system.
- Error is the remaining unexplained variation in the outcome.

What could affect my outcome? (Don't know, do a screening experiment!)

- What are my main treatments or interventions?
- What biological effects need to be accounted for?
  - Sex
  - Age
  - Weight
  - Litter
  - Cell line
  - Species/Strain
- What technical effects need to be accounted for?
  - Cage
  - Plate
  - Batch
  - Position/Location
  - Experimenter
  - Day
  - Order
  - Machine

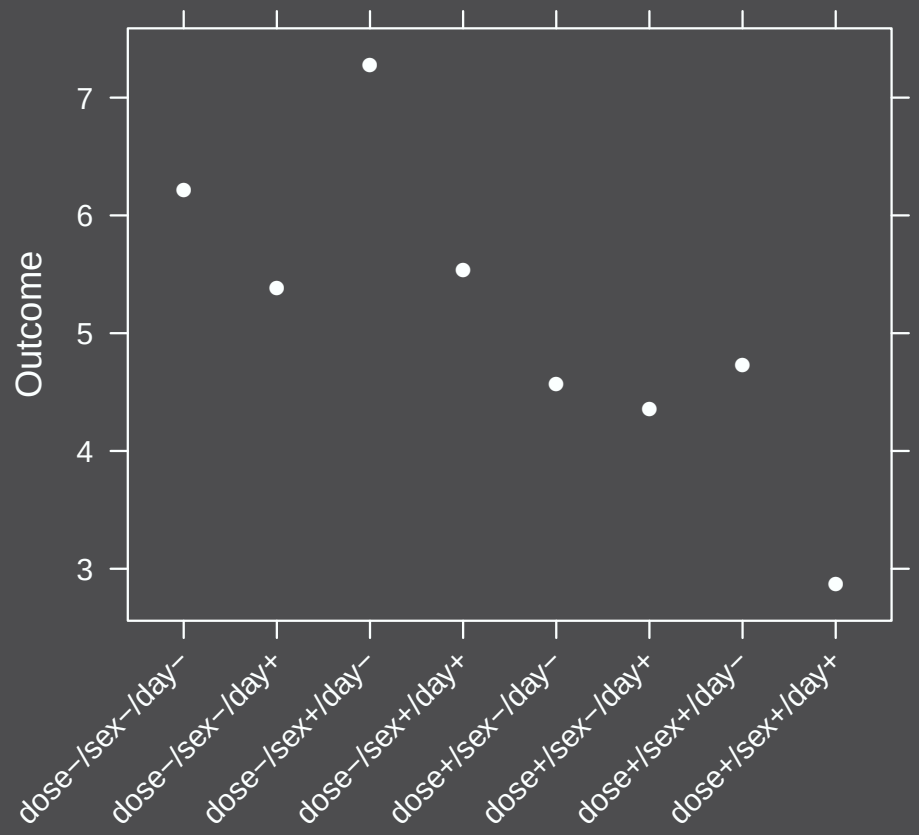
What am I going to do about it?

Options include:

1) Ignore. The effect is small and we need to compromise.
2) Hold constant (e.g. animals are the same age and weight).
3) Balance across treatment conditions (use blocking).
4) Measure and adjust during the analysis. Rely on randomisation to balance across treatment conditions (useful if it can only be measured at the end of the experiment. . . but should not be affected by the treatment/intervention).

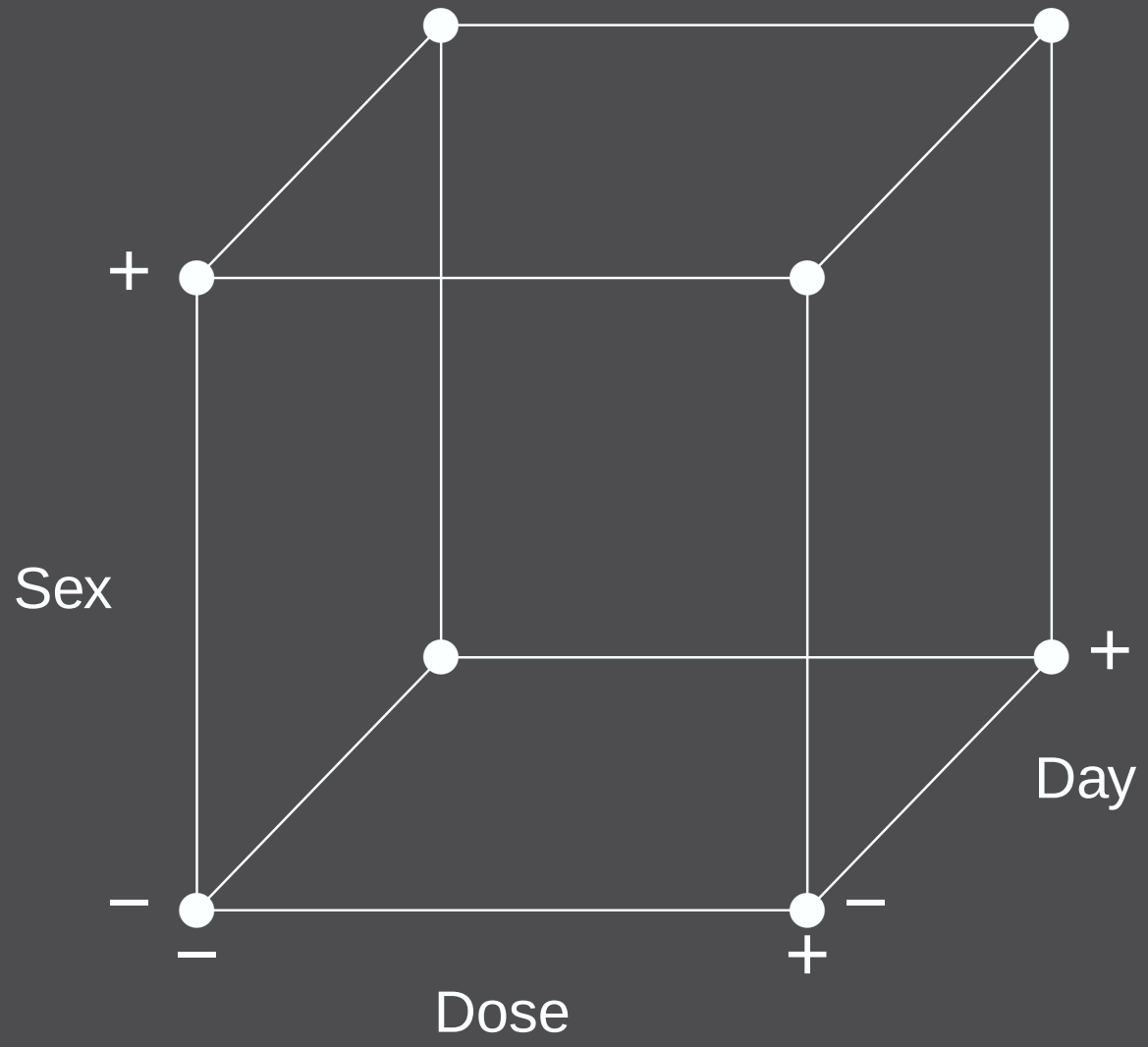
A mental model for experimental designs

Value	Day	Dose	Sex
6.2	-	-	-
5.4	-	-	+
7.3	-	+	-
5.5	-	+	+
4.6	+	-	-
4.4	+	-	+
4.7	+	+	-
2.9	+	+	+

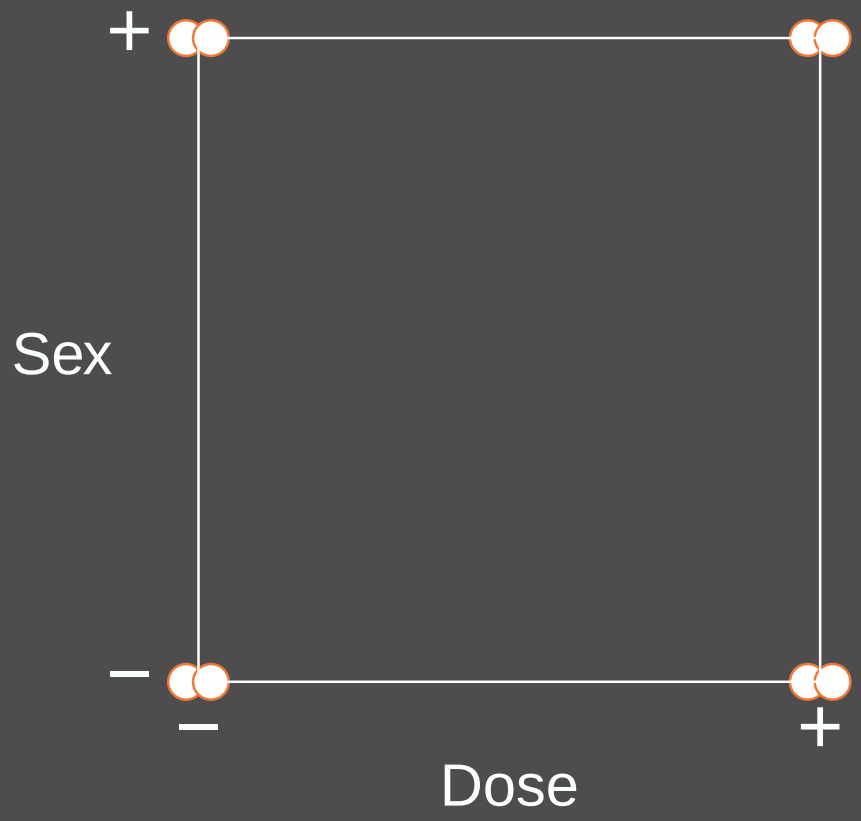


This is not a one-way design with 8 factor levels; it's a $2 \times 2 \times 2$ design. Think "variables and factor levels", not "unique groups".

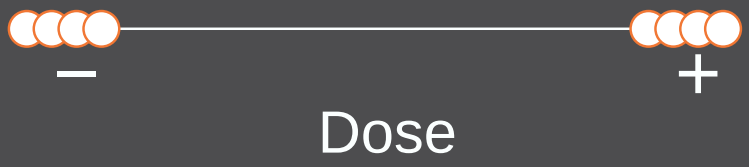
We need to think multidimensionally



To compare doses, collapse across day and sex

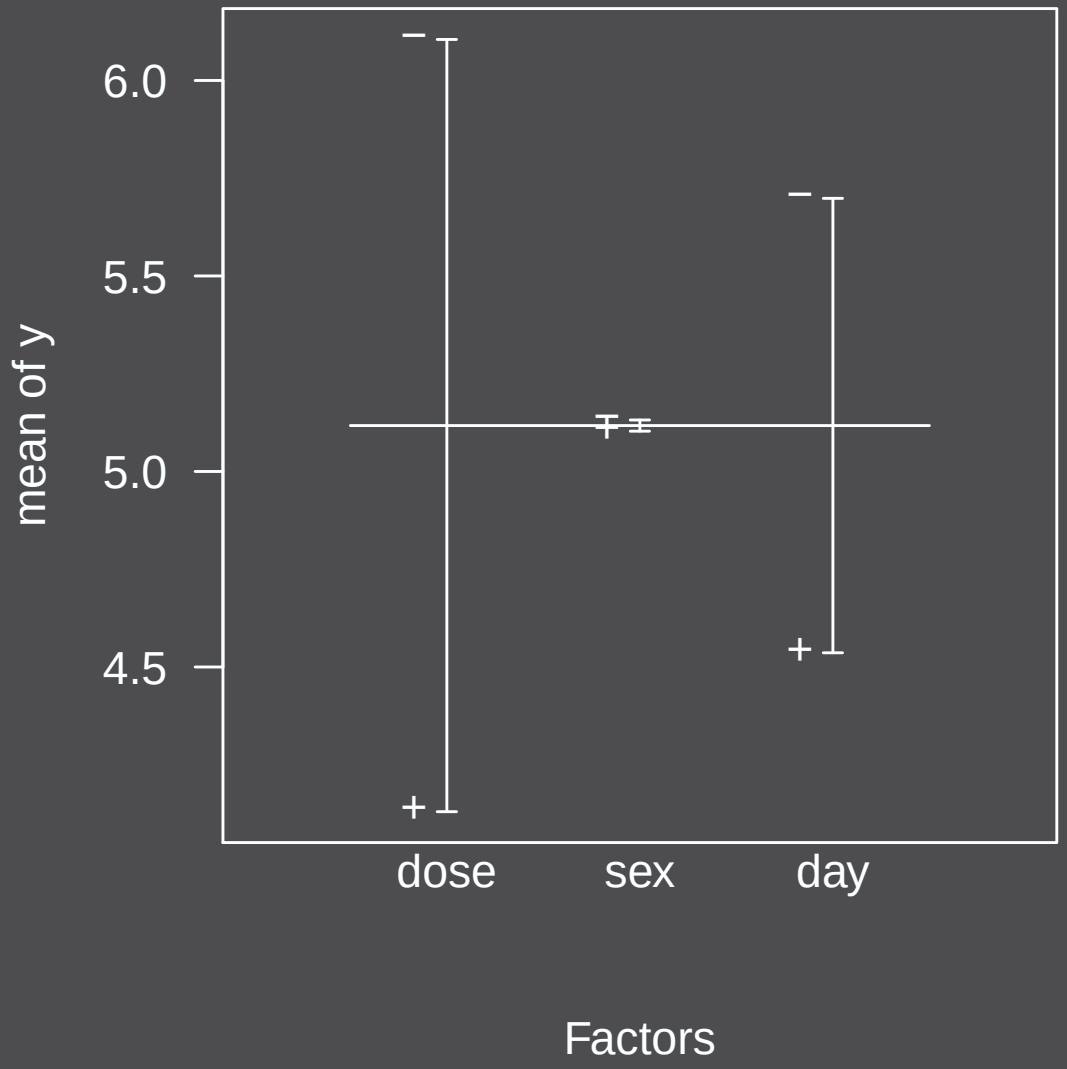


To compare doses, collapse across day and sex



Plot collapsed mean values

```
plot.design(y ~ dose + sex + day, data=d)
```



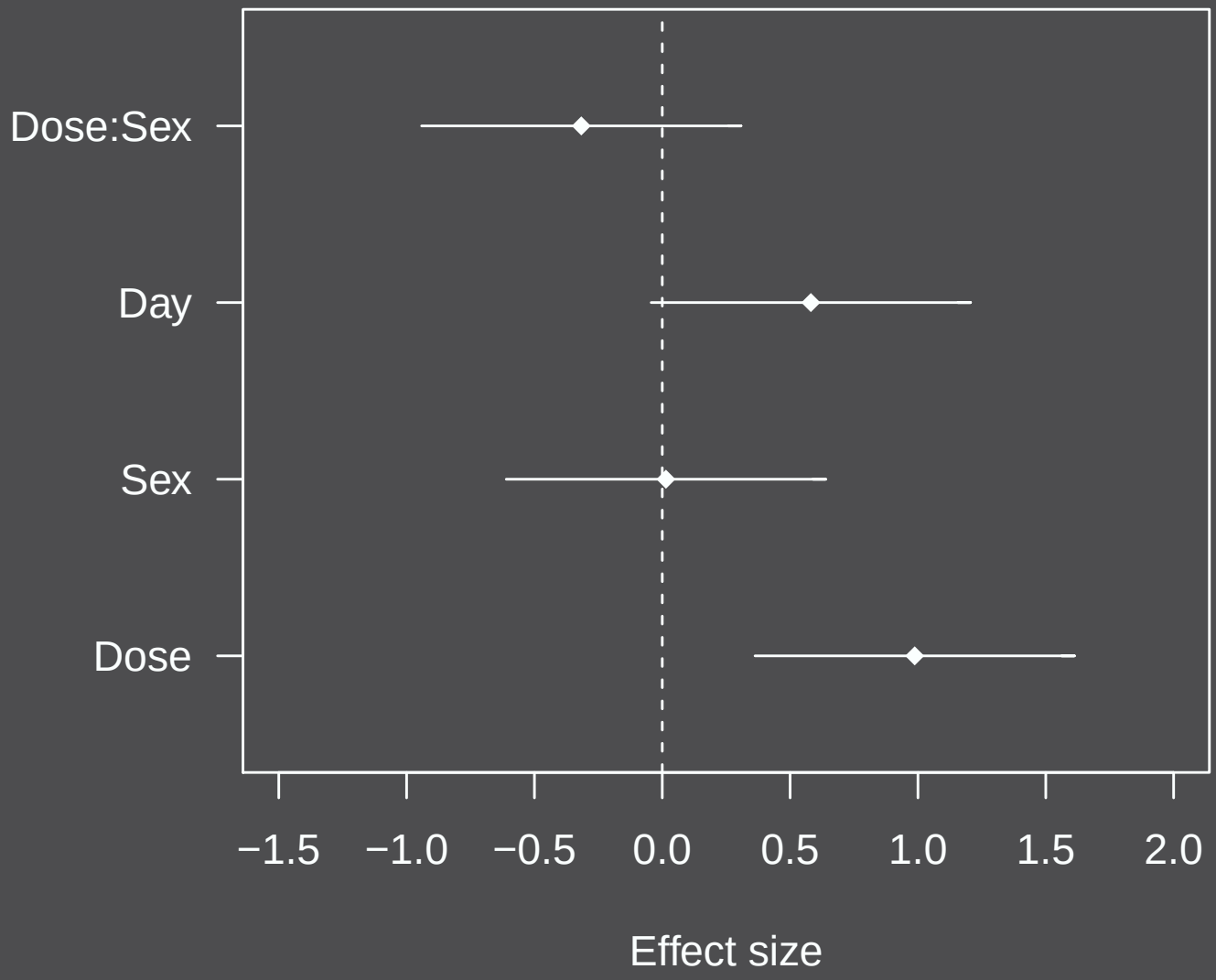
Hypothesis testing

Analysis with dose, sex, day, and the dose-by-sex interaction (i.e. does the dose differ between sexes).

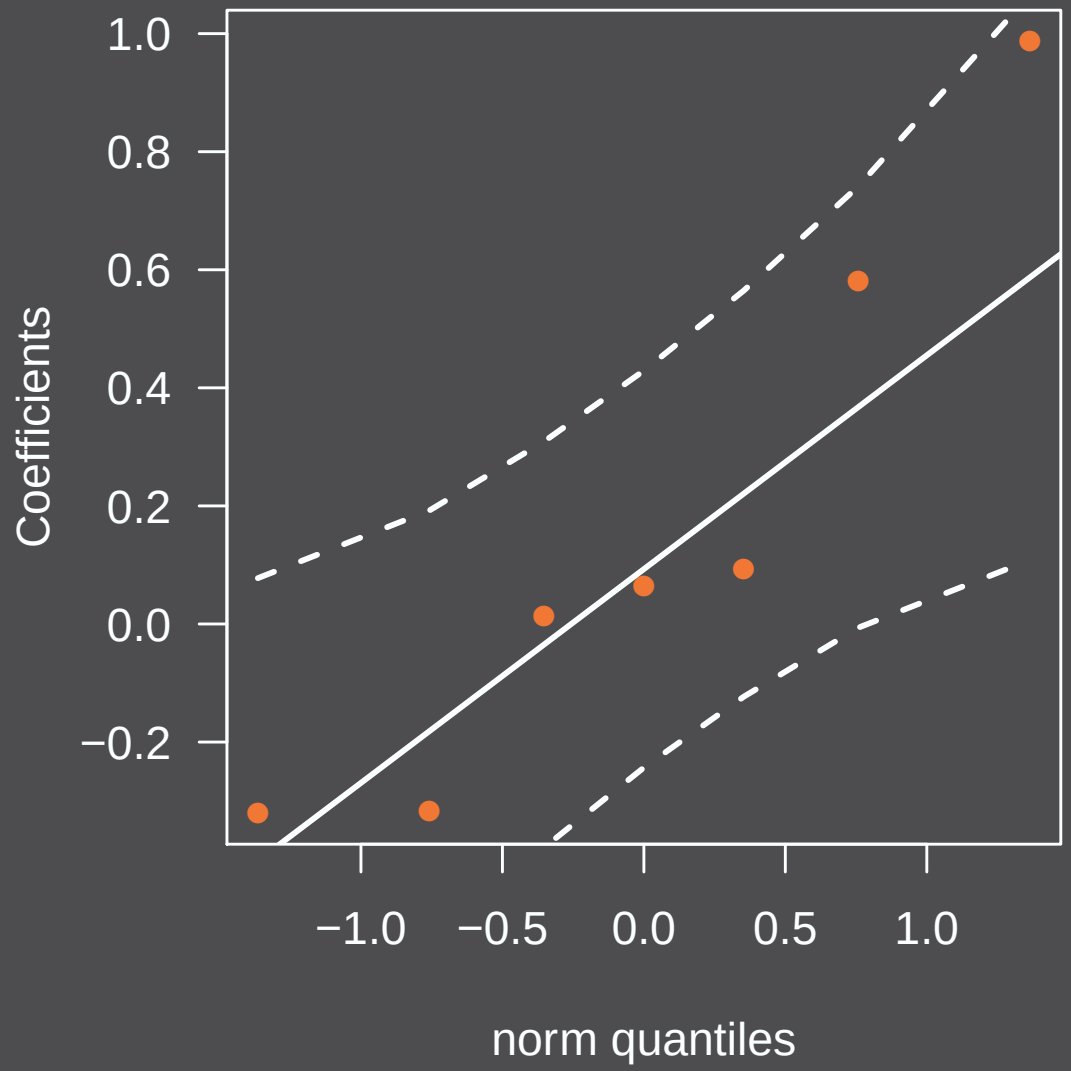
```
> car::Anova(lm(value ~ dose * sex + day, data=d))
Anova Table (Type II tests)

Response: value
          Sum Sq Df F value Pr(>F)
dose       7.800  1  25.309 0.0151 *
sex        0.002  1   0.005 0.9465
day        2.702  1   8.767 0.0595 .
dose:sex   0.803  1   2.605 0.2049
Residuals  0.925  3
```

Parameter estimation



Learning (screening)



References

1) Gelman A, Hill J, Yajima M (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5(2): 189–211.

2) Glass DJ (2014). *Experimental Design for Biologists*. Cold Spring Harbor Press: NY.

3) Lazic SE (2016). *Experimental Design for Laboratory Biologists Maximising Information and Improving Reproducibility*. Cambridge University Press: Cambridge, UK.

4) Stephens M (2017). False discovery rates: a new deal. *Biostatistics* 18(2): 275–294.