How to design a good experiment

Advanced methods for reproducible science

Stanley E. Lazic

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

5 April 2018

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?
- What will you measure (1^{o}, 2^{o}, ancillary outcomes)?

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?
- What will you measure (1^o, 2^o, ancillary outcomes)?
- What will you manipulate?

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?
- What will you measure (1^{o}, 2^{o}, ancillary outcomes)?
- What will you manipulate?
- What biological or technical effects might influence the outcome(s)?

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?
- What will you measure (1^o, 2^o, ancillary outcomes)?
- What will you manipulate?
- What biological or technical effects might influence the outcome(s)?
- Which of these will you ignore, hold constant, or allow to vary and include in the design?

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?
- What will you measure (1^o, 2^o, ancillary outcomes)?
- What will you manipulate?
- What biological or technical effects might influence the outcome(s)?
- Which of these will you ignore, hold constant, or allow to vary and include in the design?
- What are the specific factor levels, doses, time points, etc.?

Questions to answer before conducting an experiment

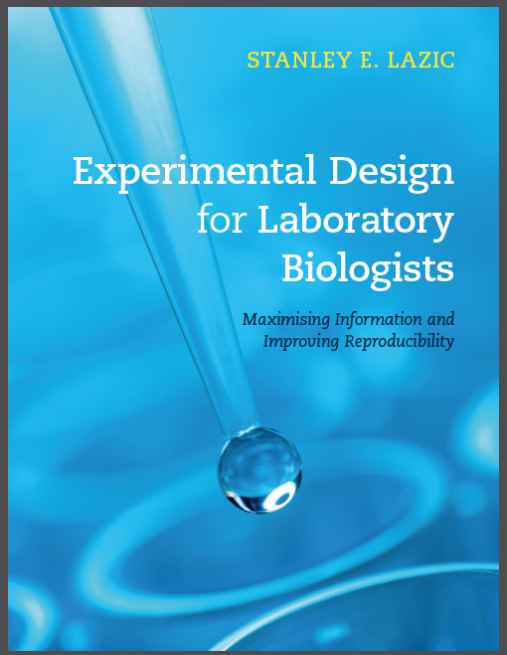
- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?
- What will you measure (1^o, 2^o, ancillary outcomes)?
- What will you manipulate?
- What biological or technical effects might influence the outcome(s)?
- Which of these will you ignore, hold constant, or allow to vary and include in the design?
- What are the specific factor levels, doses, time points, etc.?
- What are the biological, experimental, and observational units?

Questions to answer before conducting an experiment

- What is the question or hypothesis (what does this add)?
- Is the experiment (mainly) exploratory or confirmatory?
- What will you measure (1^{o}, 2^{o}, ancillary outcomes)?
- What will you manipulate?
- What biological or technical effects might influence the outcome(s)?
- Which of these will you ignore, hold constant, or allow to vary and include in the design?
- What are the specific factor levels, doses, time points, etc.?
- What are the biological, experimental, and observational units?
- What sample size do you need (resource equation or sample size calculation)?

Topics covered

- Properties of a good experiment
- Two experimental goals
- Fundamental experimental design equation
- Replication
- Power/sample size calculations



Properties of a good experiment

1) Effects can be estimated unambiguously and without bias.

Properties of a good experiment

1) Effects can be estimated unambiguously and without bias.
2) Estimates are precise.

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.

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.

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.

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.

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.

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.

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

Two experimental goals

1) Learn/explore: Discover as much as possible about phenomenon under investigation, given time and resource constraints.

Two experimental goals

1) Learn/explore: Discover as much as possible about phenomenon under investigation, given time and resource constraints.
2) Confirm: Verify or validate a result or hypothesis (often derived from an earlier learning experiment or from theory).

Two experimental goals

1) Learn/explore: Discover as much as possible about phenomenon under investigation, given time and resource constraints.
2) Confirm: Verify or validate a result or hypothesis (often derived from an earlier learning experiment or from theory).

Problem: Many learning experiments are designed and presented as confirming experiments.

Learning vs. confirming examples

Learning questions	Confirming questions
Does the drug have toxic side effects (at what dose, given for how long, in which tissue)?	Does 5 mg/kg of the drug given once a day for 5 days increase blood creatinine concentration?

Learning vs. confirming examples

Learning questions	Confirming questions
Does the drug have toxic side effects (at what dose, given for how long, in which tissue)?	Does 5 mg/kg of the drug given once a day for 5 days increase blood creatinine concentration?
Does stress affect rodent behaviour (what kind of stress, for how long, on what behavioural tasks)?	Does fox urine odour affect the amount of food Wistar rats consume during the first 24 hours after exposure?

Learning vs. confirming examples

Learning questions	Confirming questions
Does the drug have toxic side effects (at what dose, given for how long, in which tissue)?	Does 5 mg/kg of the drug given once a day for 5 days increase blood creatinine concentration?
Does stress affect rodent behaviour (what kind of stress, for how long, on what behavioural tasks)?	Does fox urine odour affect the amount of food Wistar rats consume during the first 24 hours after exposure?
Does exercise affect cognitive functioning in older people (what type of exercise, how much, which aspect of cognition)?	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?

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised
Treatments	Many	Few

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised
Treatments	Many	Few
Factor levels	Many	Few

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised
Treatments	Many	Few
Factor levels	Many	Few
Design space	Large	Small

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised
Treatments	Many	Few
Factor levels	Many	Few
Design space	Large	Small
Outcomes	Many	Few

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised
Treatments	Many	Few
Factor levels	Many	Few
Design space	Large	Small
Outcomes	Many	Few
Time points	Many	Few

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised
Treatments	Many	Few
Factor levels	Many	Few
Design space	Large	Small
Outcomes	Many	Few
Time points	Many	Few
Controls	Few	Many

Design options for both goals

Design feature	Learning	Confirming
Subjects	Heterogeneous	Homogeneous
Environment	Varied	Standardised
Treatments	Many	Few
Factor levels	Many	Few
Design space	Large	Small
Outcomes	Many	Few
Time points	Many	Few
Controls	Few	Many
Analysis	Bayesian	Hypothesis testing

Sheiner LB (1997).

Fundamental Experimental Design Equation

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

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.

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.

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.

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?

- What are my main treatments or interventions?

What could affect my outcome?

- What are my main treatments or interventions?
- What biological effects need to be accounted for?
  - Sex
  - Age
  - Weight
  - Litter
  - Cell line
  - Species/Strain

What could affect my outcome?

- 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.

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).

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).

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).

On the NIH's recommendation to use both sexes

Duplicating studies to 'compare and contrast experimental findings in male and female animals and cells' is rarely practical, affordable, prudent, scientifically warranted or ethically justifiable.

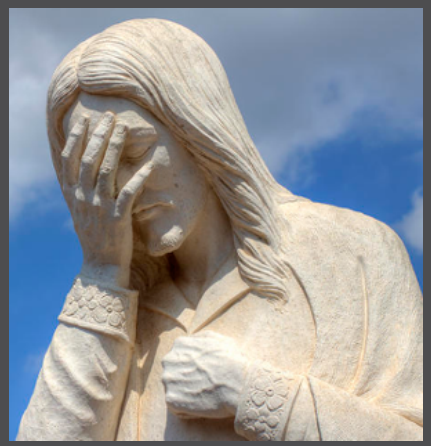
. . . using both sexes halves sample size while increasing variance, making it less likely that an observed difference not due to sex can be detected at a statistically significant level. Thus, an increased number of samples would be needed to reach firm conclusions.

Fields RD (2014).

On the NIH's recommendation to use both sexes

Duplicating studies to 'compare and contrast experimental findings in male and female animals and cells' is rarely practical, affordable, prudent, scientifically warranted or ethically justifiable.

. . . using both sexes halves sample size while increasing variance, making it less likely that an observed difference not due to sex can be detected at a statistically significant level. Thus, an increased number of samples would be needed to reach firm conclusions.



Fields RD (2014).

Does adding another variable reduce *N* by a half?

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

Does adding another variable reduce *N* by a half?

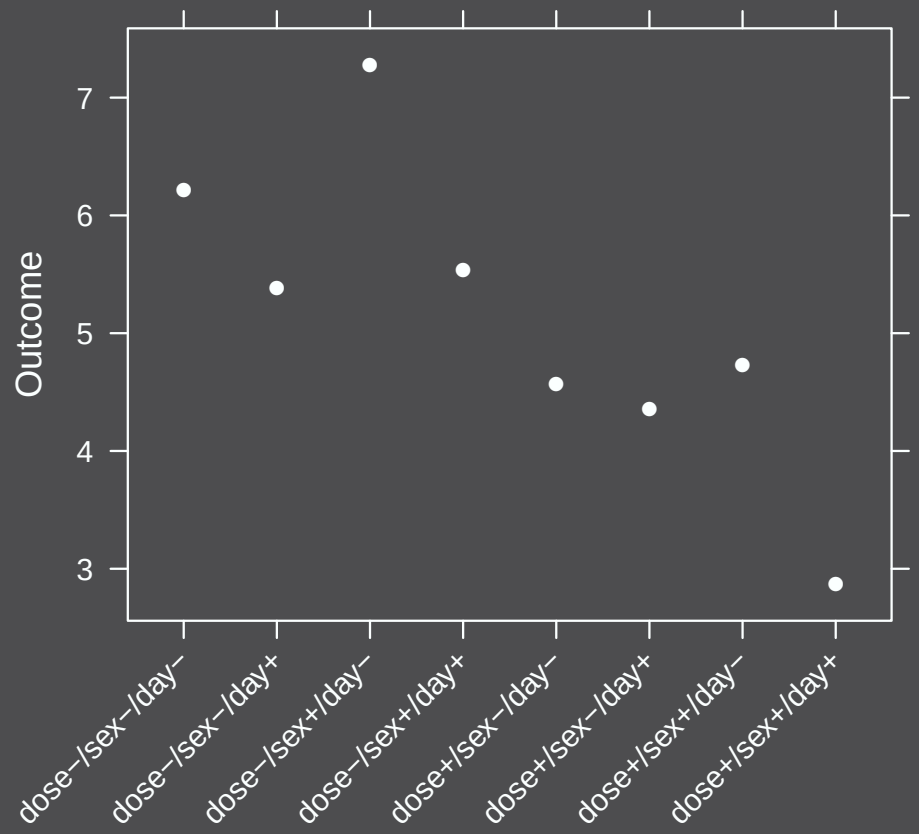
Value	Dose	Sex
6.2	-	-
5.4	-	-
7.3	-	+
5.5	-	+
4.6	+	-
4.4	+	-
4.7	+	+
2.9	+	+

Does adding another variable reduce *N* by a half?

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

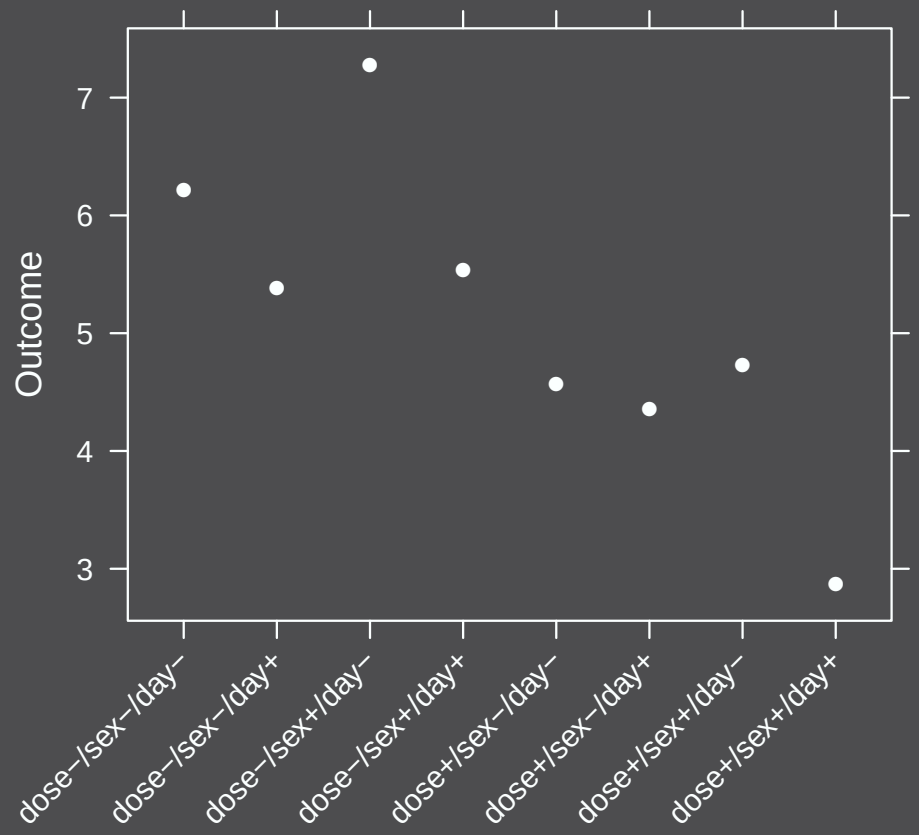
Does adding another variable reduce *N* by a half?

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



Does adding another variable reduce *N* by a half?

Value	Dose	Sex	Day
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.

Does adding another variable reduce *N* by a half?

Analysis with dose only.

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

Response: value
          Sum Sq Df F value Pr(>F)
dose       7.800  1   10.56 0.0175 *
Residuals  4.431  6
```

Note: residual df = 6.

Does adding another variable reduce *N* by a half?

Analysis with dose and sex.

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

Response: value
          Sum Sq Df F value Pr(>F)
dose       7.800  1   8.805 0.0313 *
sex        0.002  1   0.002 0.9674
Residuals  4.429  5
```

Note: residual df = 5. We lost the equivalent of one sample and asked another question!

Does adding another variable reduce *N* by a half?

Analysis with dose, sex, and day.

```
> 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  18.062 0.0132 *
sex        0.002  1   0.004 0.9539
day        2.702  1   6.257 0.0667 .
Residuals  1.727  4
```

Note: residual df = 4. We lost the equivalent of one more sample and asked one further question!

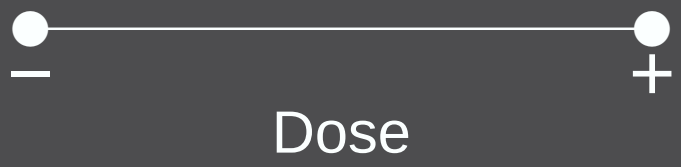
Does adding another variable reduce *N* by a half?

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

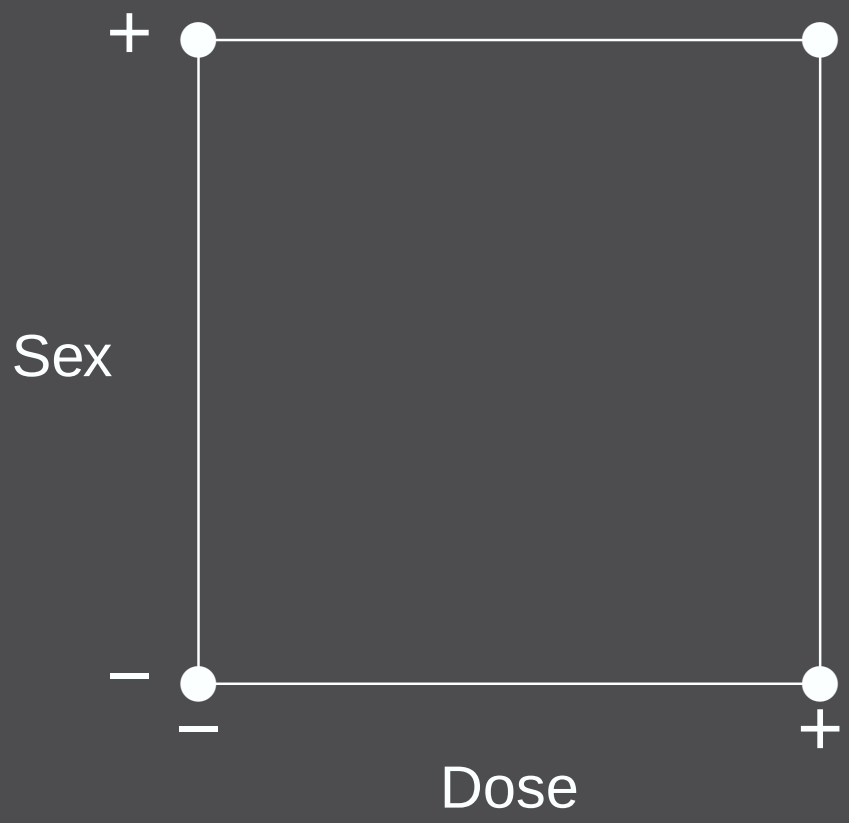
```
> car::Anova(lm(value ~ dose * sex + day, data=d))
Anova Table (Type II tests)

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
```

Note: residual df = 3. We lost the equivalent of one more sample but asked yet another question!

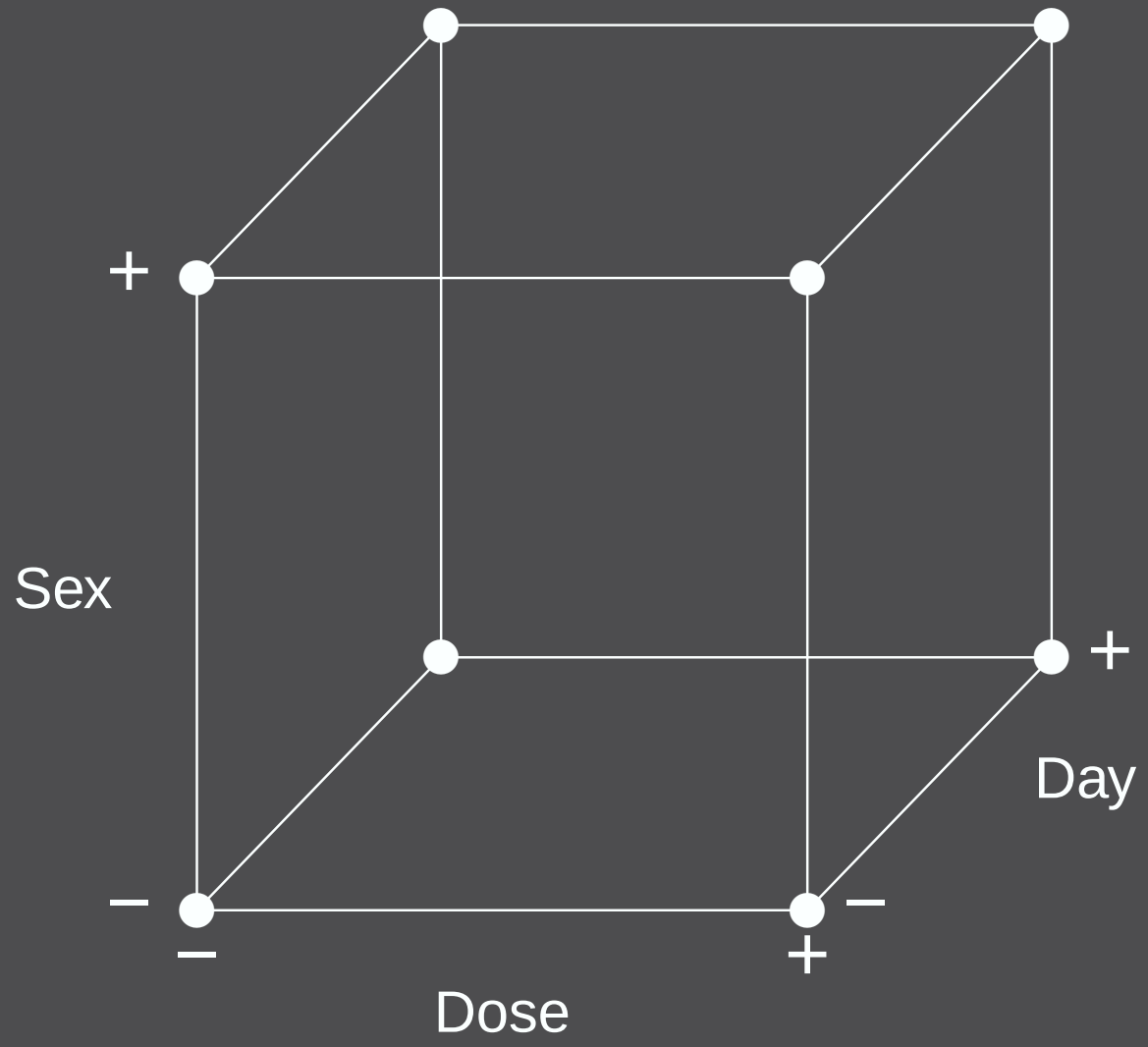
We need to think multidimensionally



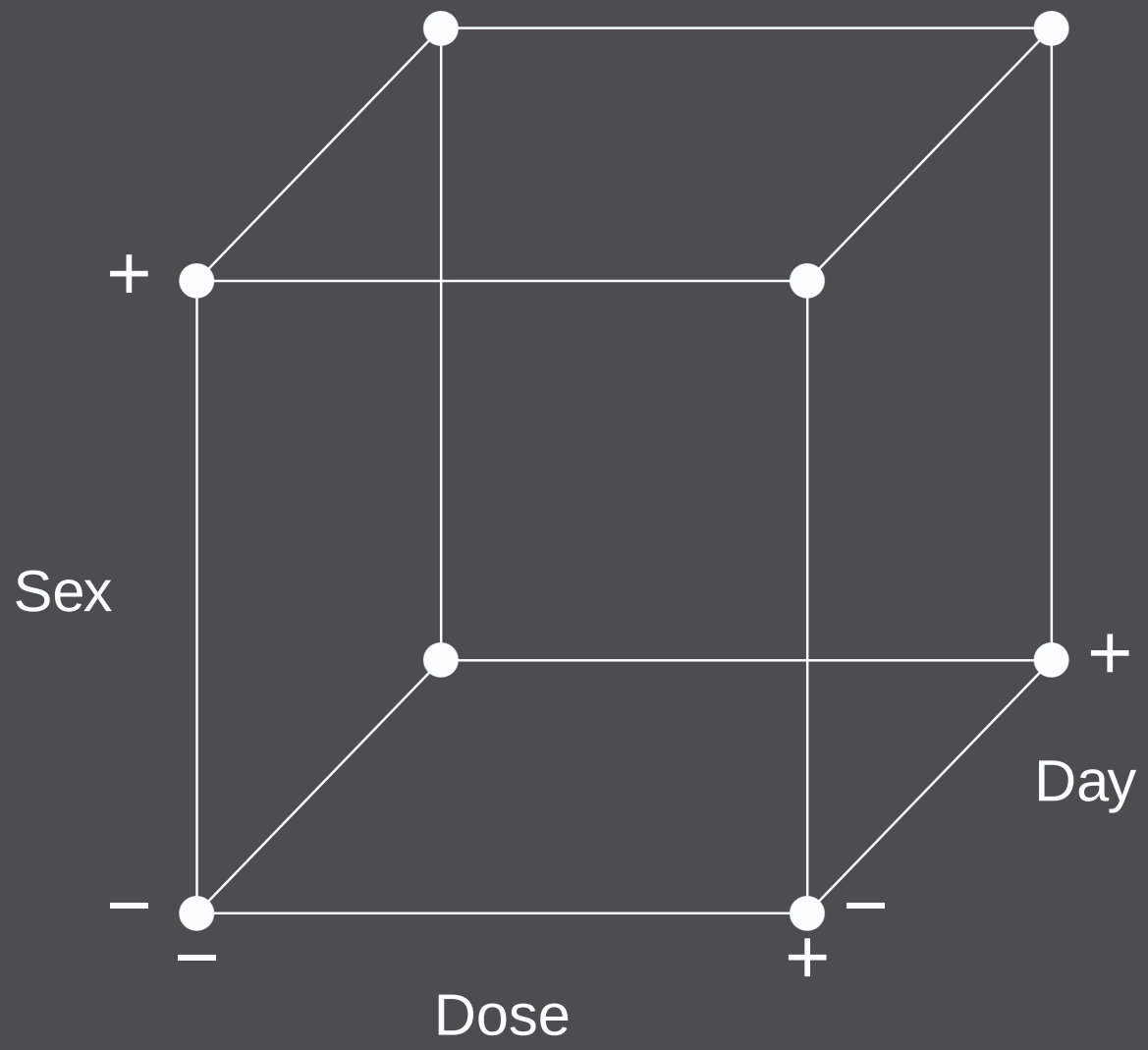
We need to think multidimensionally



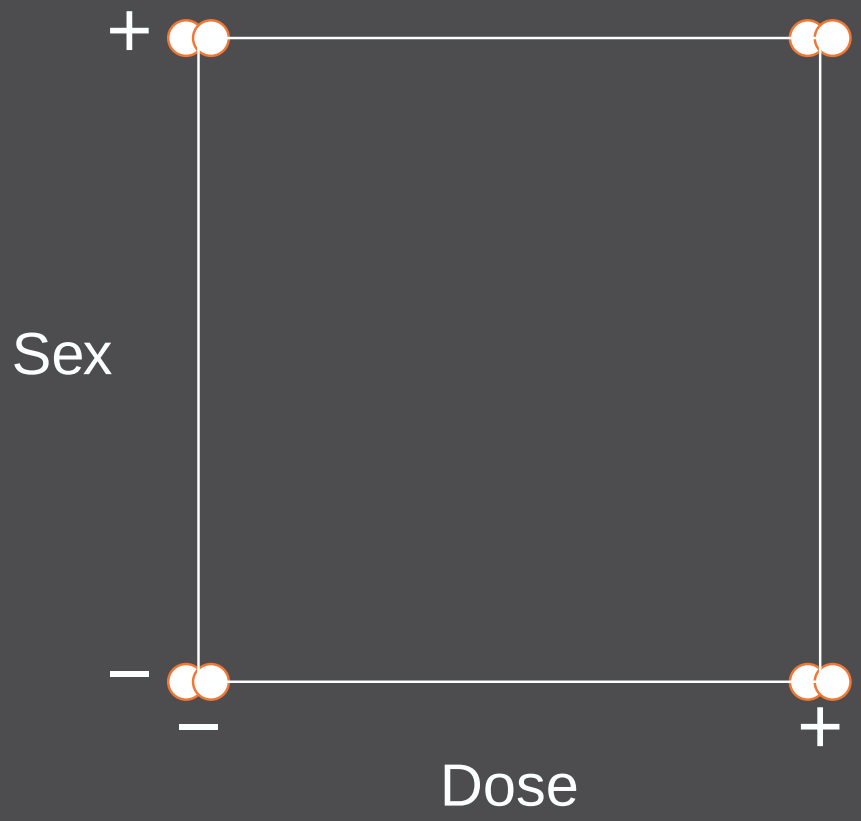
We need to think multidimensionally



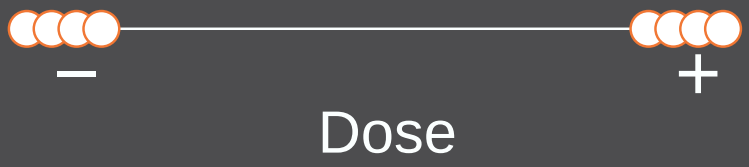
To compare doses, collapse across day and sex



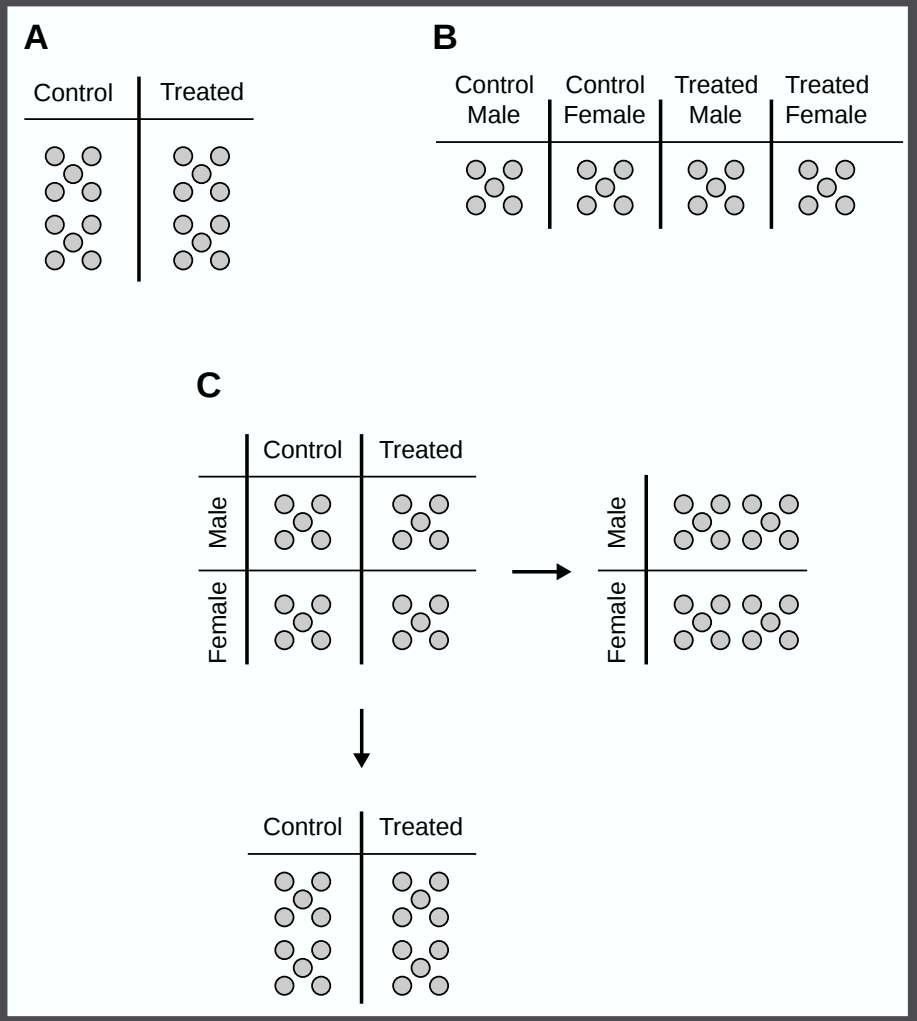
To compare doses, collapse across day and sex



To compare doses, collapse across day and sex

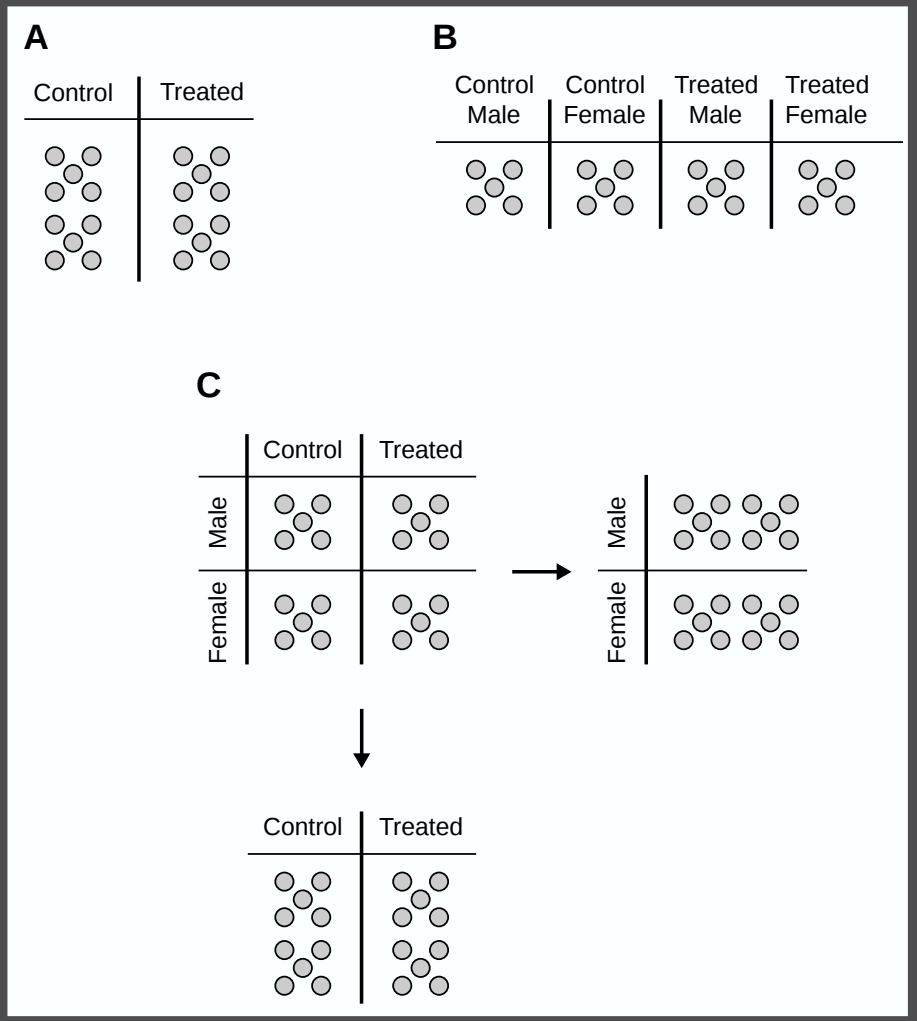


One more example



Conclusions:

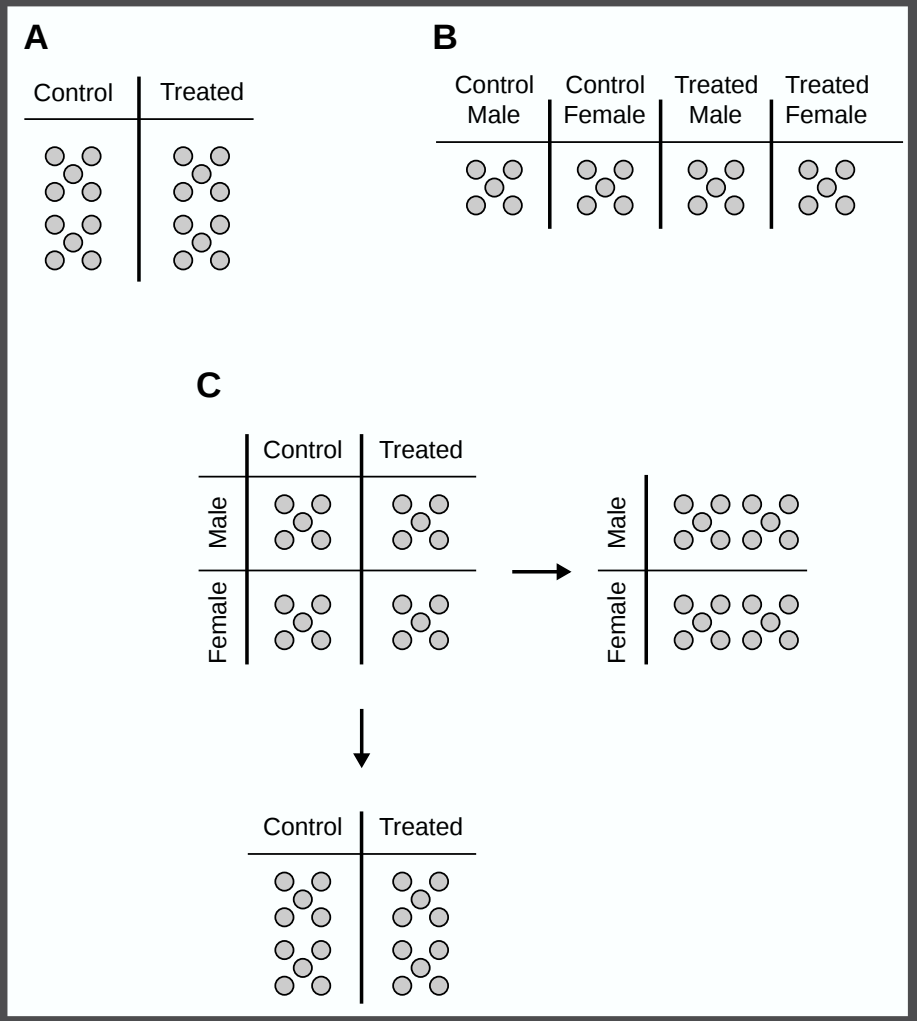
One more example



Conclusions:

1) We do not decrease the sample size by 50% when including new variables (power hardly changed).

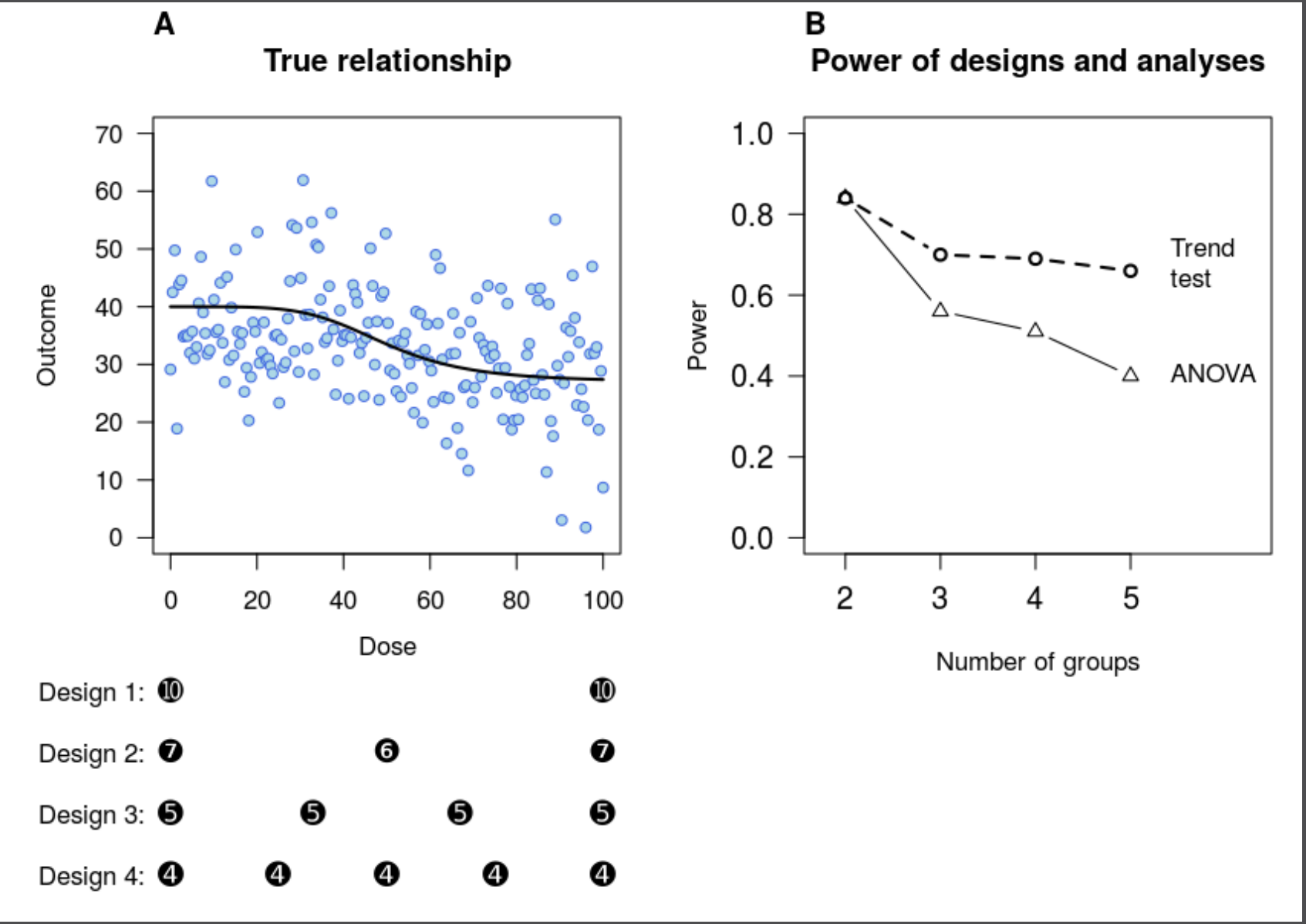
One more example



Conclusions:

1) We do not decrease the sample size by 50% when including new variables (power hardly changed).
2) Increased variance due to the new variables is irrelevant when they are included in the model. Only residual variation matters.

Choose the design that best addresses the question



Lazic SE (2018).

Replication

Biological unit of interest (BU): is the entity about which inferences are made.

Replication

Biological unit of interest (BU): is the entity about which inferences are made.

Experimental unit (EU): the entity that is randomly and independently assigned to one treatment or another. The sample size (*N*) is equal to the number of EUs. They may correspond to:

(1) a biological unit of interest
(2) groups of biological units
(3) parts of a biological unit
(4) a sequence of observations on a biological unit

Replication

Biological unit of interest (BU): is the entity about which inferences are made.

Experimental unit (EU): the entity that is randomly and independently assigned to one treatment or another. The sample size (*N*) is equal to the number of EUs. They may correspond to:

(1) a biological unit of interest
(2) groups of biological units
(3) parts of a biological unit
(4) a sequence of observations on a biological unit

Ideally, the treatment should be applied independently to each EU, and the EUs should not influence each other.

Replication

Biological unit of interest (BU): is the entity about which inferences are made.

Experimental unit (EU): the entity that is randomly and independently assigned to one treatment or another. The sample size (*N*) is equal to the number of EUs. They may correspond to:

(1) a biological unit of interest
(2) groups of biological units
(3) parts of a biological unit
(4) a sequence of observations on a biological unit

Ideally, the treatment should be applied independently to each EU, and the EUs should not influence each other.

Observational unit (OU): the entity on which measurements are taken, which may be different from the EU and BU.

Lazic SE, Clarke-Williams CJ, Munafo MR (2018).

Power/sample size calculations

Power analysis: A prediction about the success of a planned experiment,

Power/sample size calculations

Power analysis: A prediction about the success of a planned experiment, based on no data, or biased and unrepresentative data,

Power/sample size calculations

Power analysis: A prediction about the success of a planned experiment, based on no data, or biased and unrepresentative data, and proclaimed with unusually high confidence.

And the LORD spake, saying, “First shalt thou take out the Holy Pin, then shalt thou count to three, no more, no less. Three shall be the number thou shalt count, and the number of the counting shall be three. Four shalt thou not count, neither count thou two, excepting that thou then proceed to three. Five is right out. Once the number three, being the third number, be reached, then lobbest thou thy Holy Hand Grenade of Antioch towards thy foe...”

The resource equation: an alternative for exploratory studies

The advice: Have 10 to 20 samples to estimate the error variance (i.e. Residual df between 10 and 20).

The reason: Below 10, the error variance is poorly estimated, and above 20 resources have been wasted because additional questions could have been asked by including further variables.

You don't need to specify a primary outcome, an effect size, or the within-group variability.

Mead R, et al. (2012); Lazic SE (2016);

`http://isogenic.info/html/resource_equation.html`

Power/sample size calculations for confirmatory studies

1) A primary outcome and a main question

Power/sample size calculations for confirmatory studies

1) A primary outcome and a main question
2) Sample size (N): Is the number of experimental units.

Power/sample size calculations for confirmatory studies

1) A primary outcome and a main question
2) Sample size (N): Is the number of experimental units.
3) Effect size: E.g. difference between two means, correlation between two variables, difference between two proportions, ratio of the between-group to within-group variability (ANOVA). Usually the minimum effect you want to detect or the predicted size of the effect.

Power/sample size calculations for confirmatory studies

1) A primary outcome and a main question
2) Sample size (N): Is the number of experimental units.
3) Effect size: E.g. difference between two means, correlation between two variables, difference between two proportions, ratio of the between-group to within-group variability (ANOVA). Usually the minimum effect you want to detect or the predicted size of the effect.
4) Within-group variability (σ). Obtained from previous experiments, published studies, or an educated guess. It is a prediction about the variability that will be observed in the planned experiment.

Power/sample size calculations for confirmatory studies

1) A primary outcome and a main question
2) Sample size (N): Is the number of experimental units.
3) Effect size: E.g. difference between two means, correlation between two variables, difference between two proportions, ratio of the between-group to within-group variability (ANOVA). Usually the minimum effect you want to detect or the predicted size of the effect.
4) Within-group variability (σ). Obtained from previous experiments, published studies, or an educated guess. It is a prediction about the variability that will be observed in the planned experiment.
5) Power: Values of 0.8 or 0.9 are common.
6) Significance threshold (α): Usually set at 0.05.

If you can simulate it, you can analyse it

Posted on ASA Connect by David C. Norris, MD on a question about statistical problems commonly seen in research:

- If you're fundamentally attracted to Statistics as a means to support your viewpoint rather than challenge it, go read Richard Feynman's 1974 Caltech Commencement Address.
- If you can't simulate the DGP [data generating process] for your data, work with someone who can.

I venture to assert that these 2 bits of advice, if followed, would eliminate most of the errors cited so far.

References

1) Fields RD (2014). NIH policy: mandate goes too far. *Nature* 509(7505): 340.
2) Lazic SE (2018). Four simple ways to increase power without increasing the sample size. *Laboratory Animals* (in press). `https://peerj.com/preprints/`
3) Lazic SE, Clarke-Williams CJ, Munafo MR (2018). What exactly is *N* in cell culture and animal experiments? *PLoS Biology* (in press).
4) Lazic SE (2016). *Experimental Design for Laboratory Biologists Maximising Information and Improving Reproducibility*. Cambridge University Press: Cambridge, UK.
5) Mead R, Gilmour SG, Mead A (2012). *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. Cambridge University Press: Cambridge, UK.
6) Sheiner LB (1997). Learning versus confirming in clinical drug development. *Clin Pharmacol Ther* 61(3): 275291.