Experimental design

Designing experiments for reliable results

Experimental design

- The logical construction of an experiment
- The most important part of your statistical education!
  - If you design an experiment well you will get reliable, clear results, and will be able to find a suitable way to statistically analyze the data
  - If you design an experiment poorly no amount of statistical wizardry will save you!
- It's very important to know how to design experiments that yield reliable results
- It's important to be able to recognize design flaws in the studies you read

The goal of an experiment

- Testing for cause/effect relationships is done with **manipulative** experiments
  - Make a change in the predictor variable
  - Observe responses
- To accomplish this goal, we want to:
  - Obtain clean, clear results → minimum of noise in the data, maximum statistical power (minimize Type II errors)
  - Avoid wasted time, effort, and money → faster progress
  - Isolate factors to test → what we *think* we are testing is what we *actually* are testing

The Experimental Ideal

- The Experimental Ideal = controlled, manipulative experiments in which we:
  - Hold everything constant except the one variable we want to test
  - Vary that single predictor of interest, measure the response
- Done correctly, there will be only one *possible* explanation for an observed change in the response variable
  - Or rather, only one possible explanation aside from a false positive, Type I error...can’t eliminate that possibility

The enemy: confounded variables

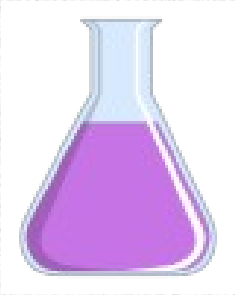
- When more than one predictor variable could explain a response the predictors are **confounded**
- Confounded predictor variables are a failure to isolate variables, and lead to unclear, uncertain results
- An apparent treatment effect that is actually due to a confounded variable is called a **spurious** result

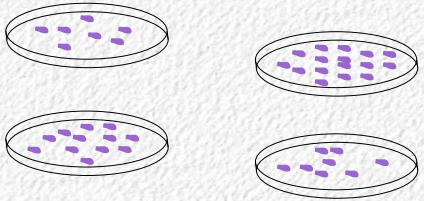
Example: studying the effects of an antibiotic on bacterial growth

- Bacteria, *E. coli*, grown in cultures in the lab
- What's the best way to test for an effect of antibiotic?

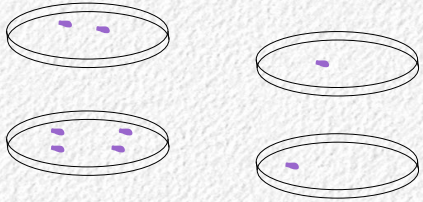
Good design?

No antibiotic



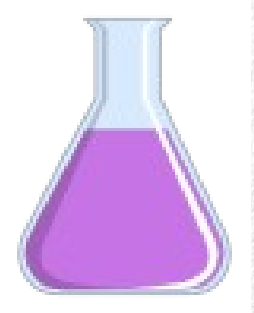


Antibiotic added

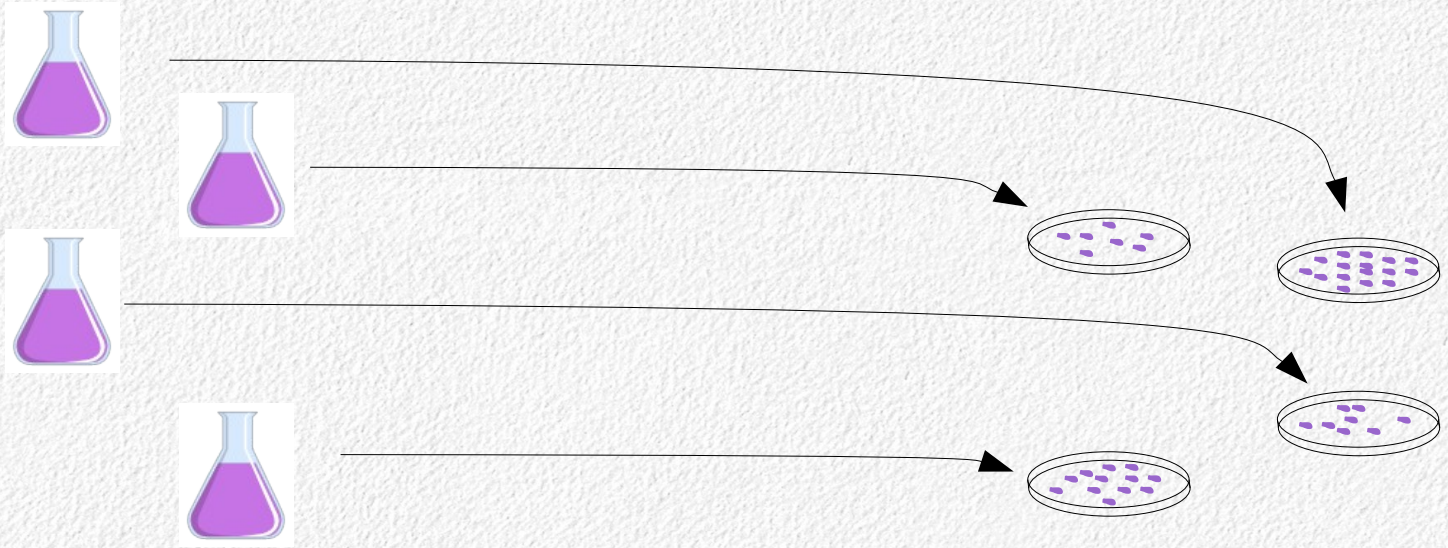


Replication

- Replication → confidence in the consistency, repeatability of a result
- Replication is done at the level of application of the treatment
  - Each independent application of the treatment is a “replicate”
- Antibiotic added to an entire flask, with responses measured for multiple samples from each flask = pseudoreplication
  - Sometimes done intentionally as “technical replicates” - used to assess repeatability, consistency of measurements
  - True independent application of treatments are “biological replicates”
- How would we truly replicate the *E. coli* experiment?

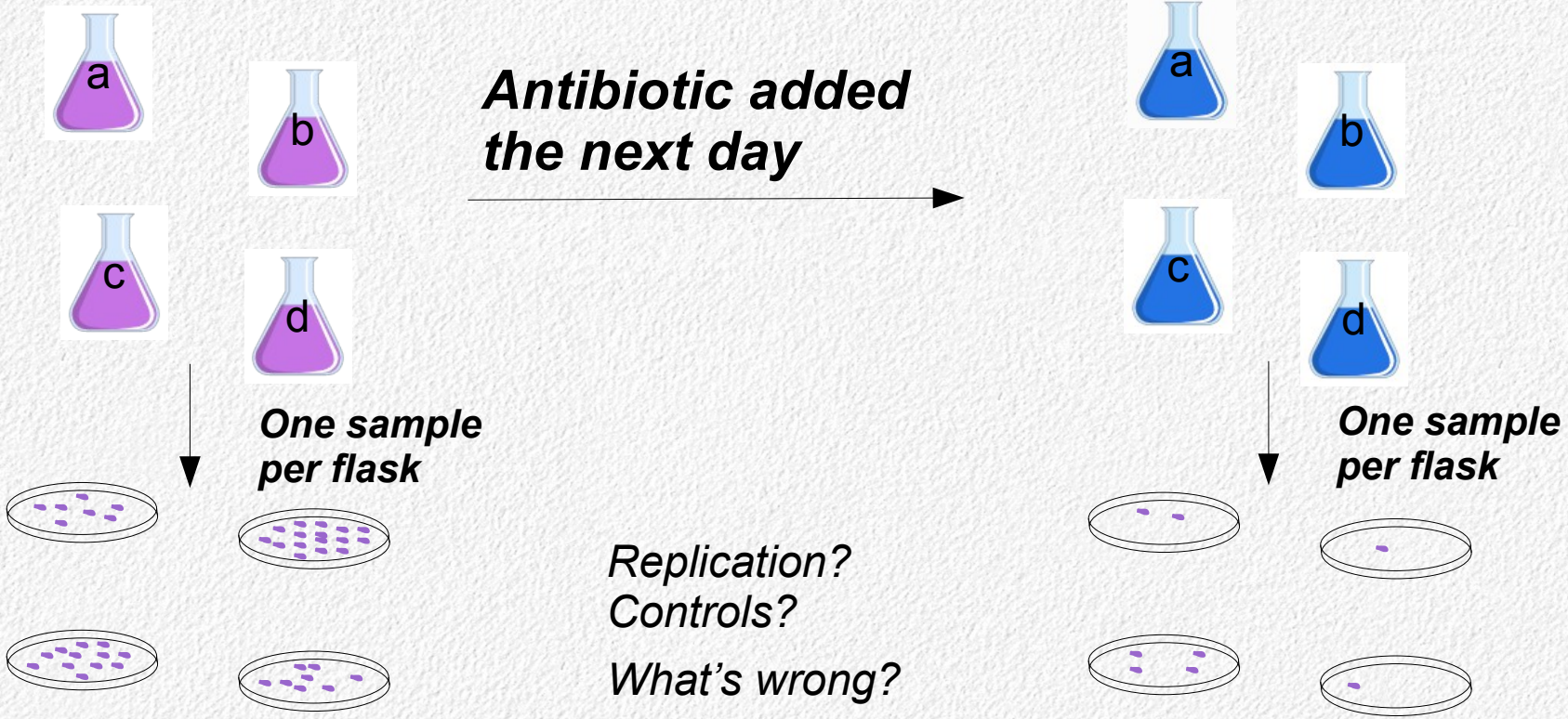


Replication – four separate cultures per group

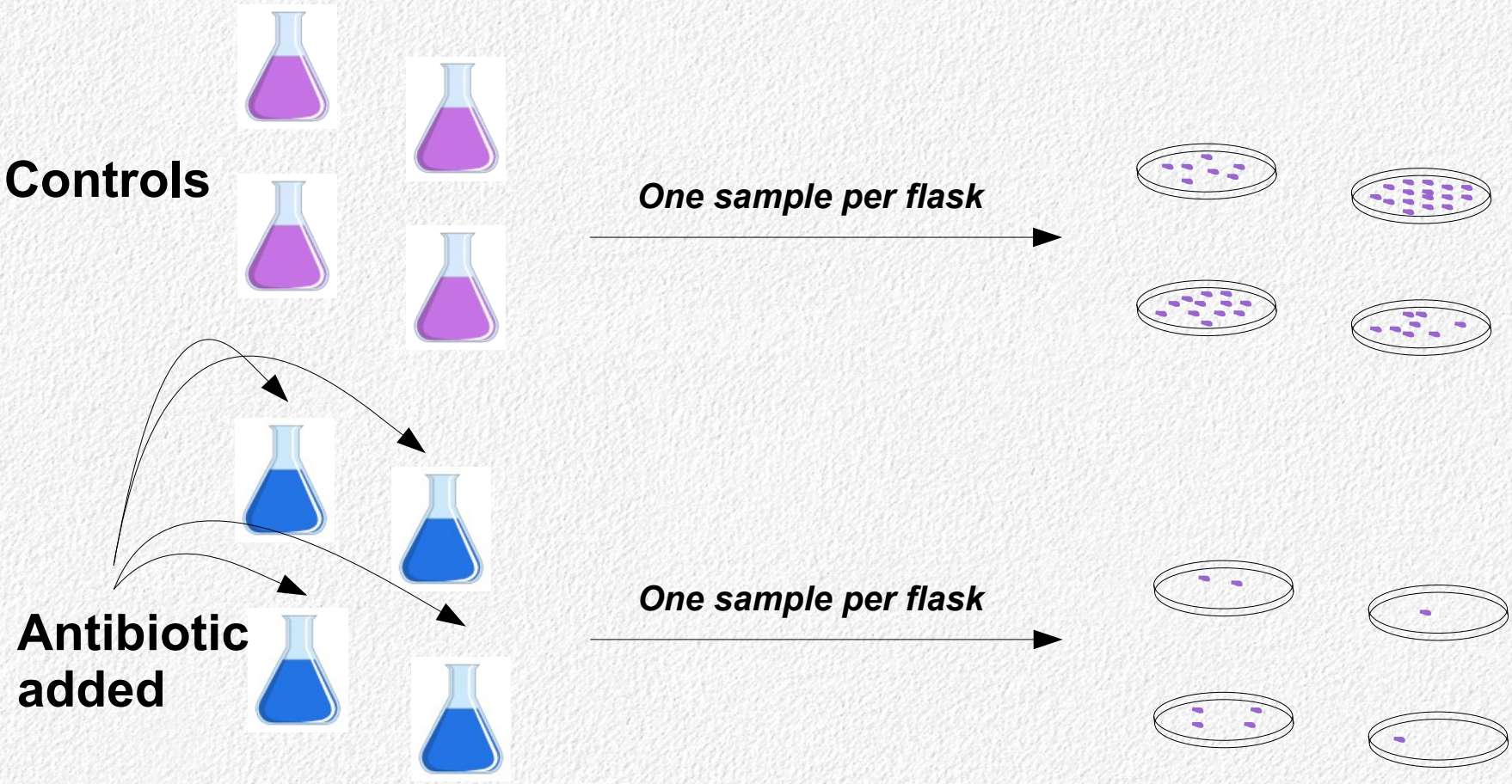


How about this?

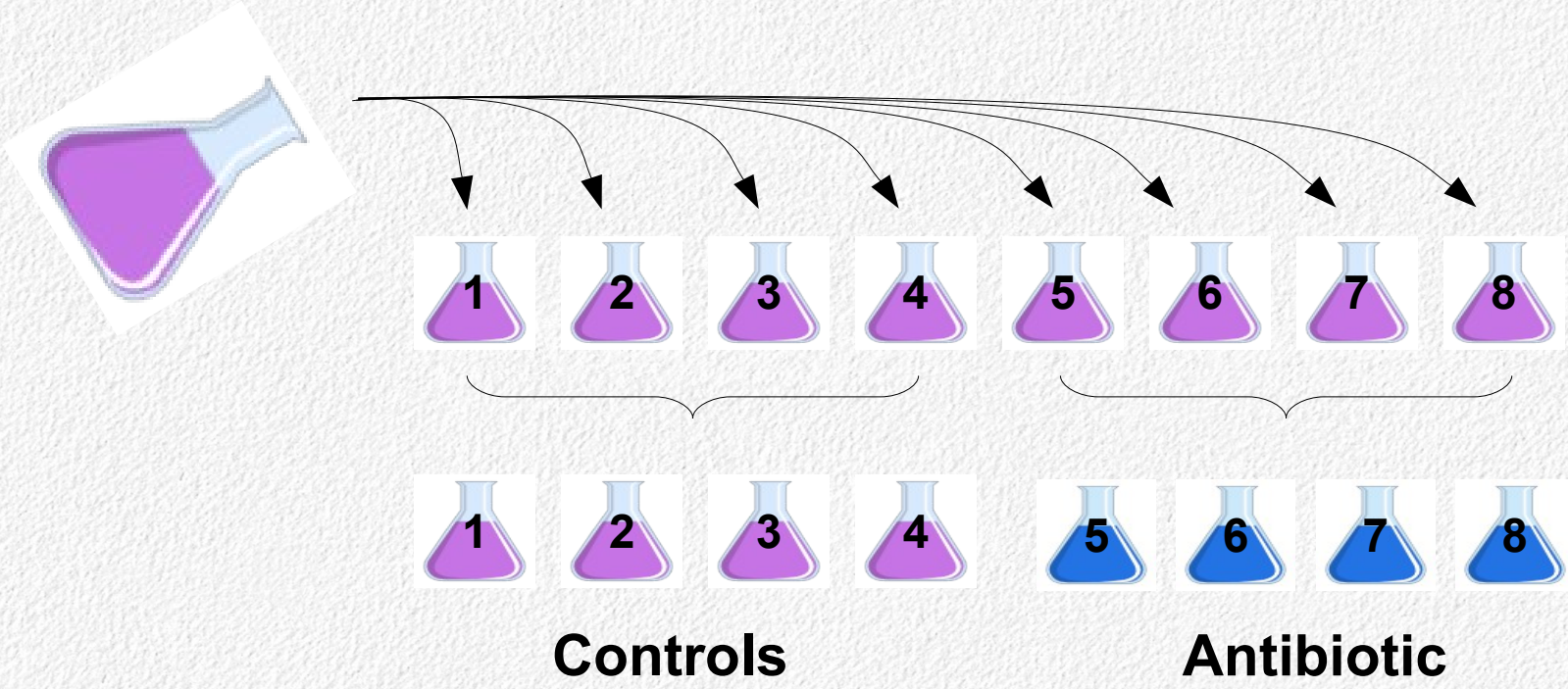
Controls



Simultaneous, independent controls are better



Assigning subjects to treatment groups

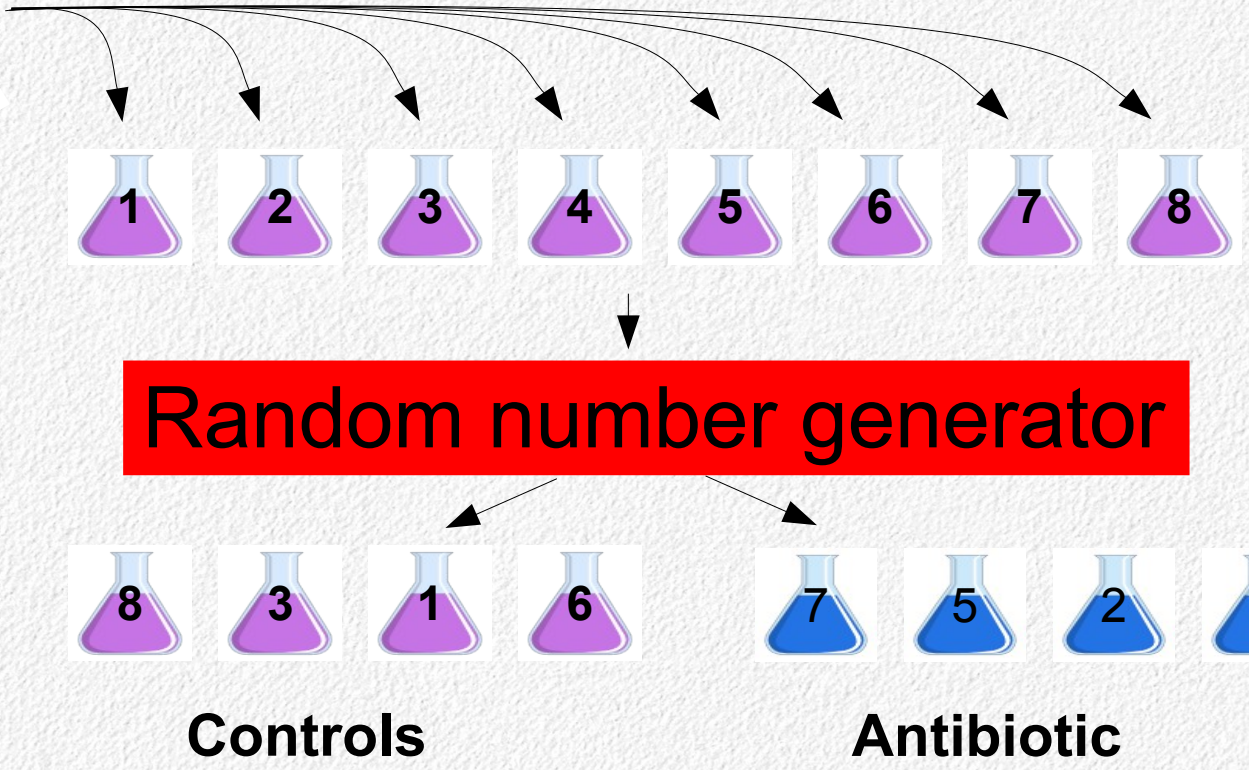


Problem?

Random assignment

- Selection bias → individuals chosen for treatment group are not the same (on average) as those in control, even before treatment is applied
- Examples:
  - First 10 mice caught → treatment, next 10 mice caught → control
  - First 4 flasks decanted from stock culture → control, next 4 → treatment
  - Asking for (human) volunteers to be in treatment group
- Best method for avoiding it is to randomly assign subjects to treatment groups
  - Works even with variables we don’t know about

Random assignment of subjects to treatment groups

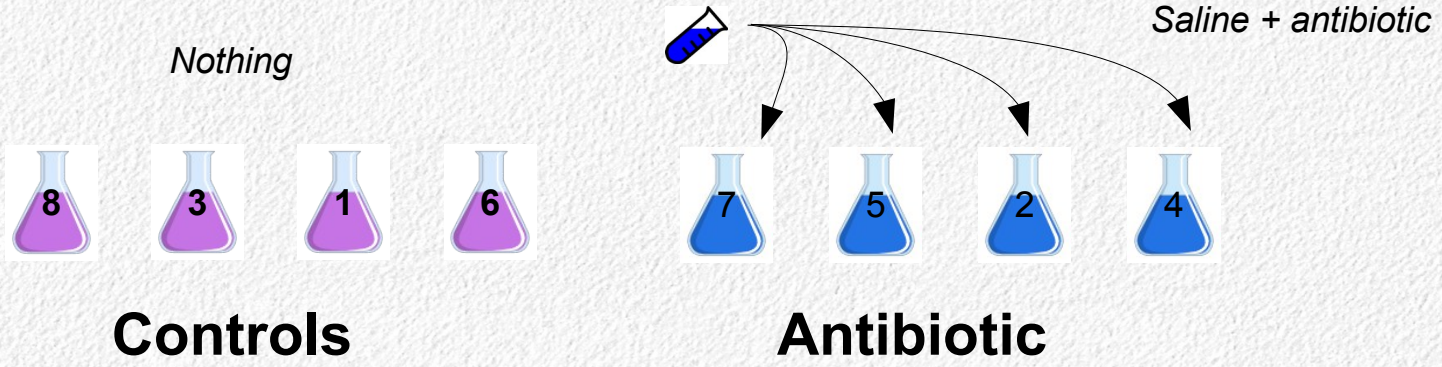


Isolating factors

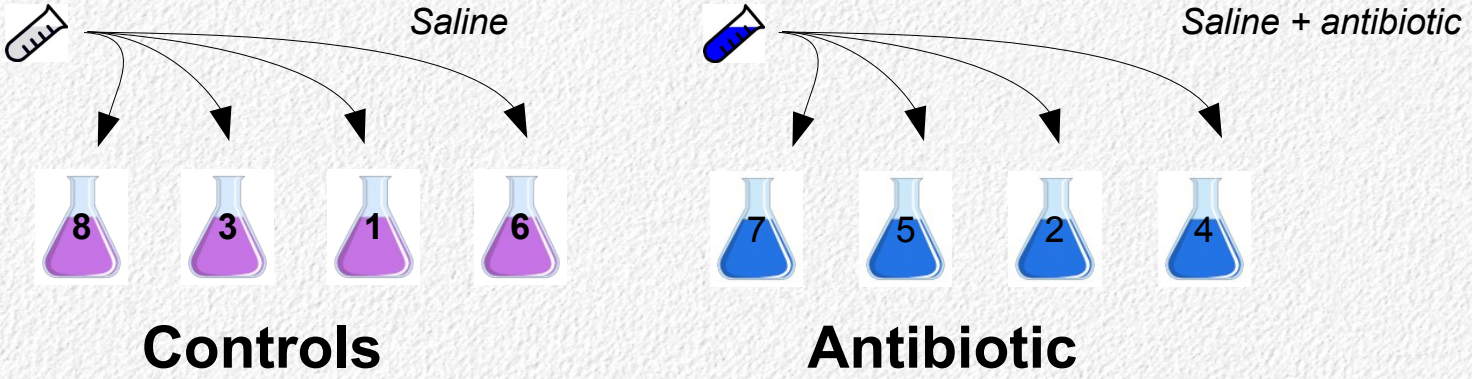
- What is the problem with:
  - Testing the effect of caffeine on plant growth by watering treatment plants with coffee, control plants with water?
  - Testing the effect of surgical ligation of the superficial mammary artery on chest pain by ligating treatment patients, and doing no surgery of any kind on controls?
  - Testing the effect of a headache medicine by giving the medicine to treatment patients, and giving nothing to controls?
- To isolate the effect of the treatment, controls have to be identical in *every possible way* except for the treatment

Which would be better?

This...



...or this?



Placebos, sham treatments, blinding

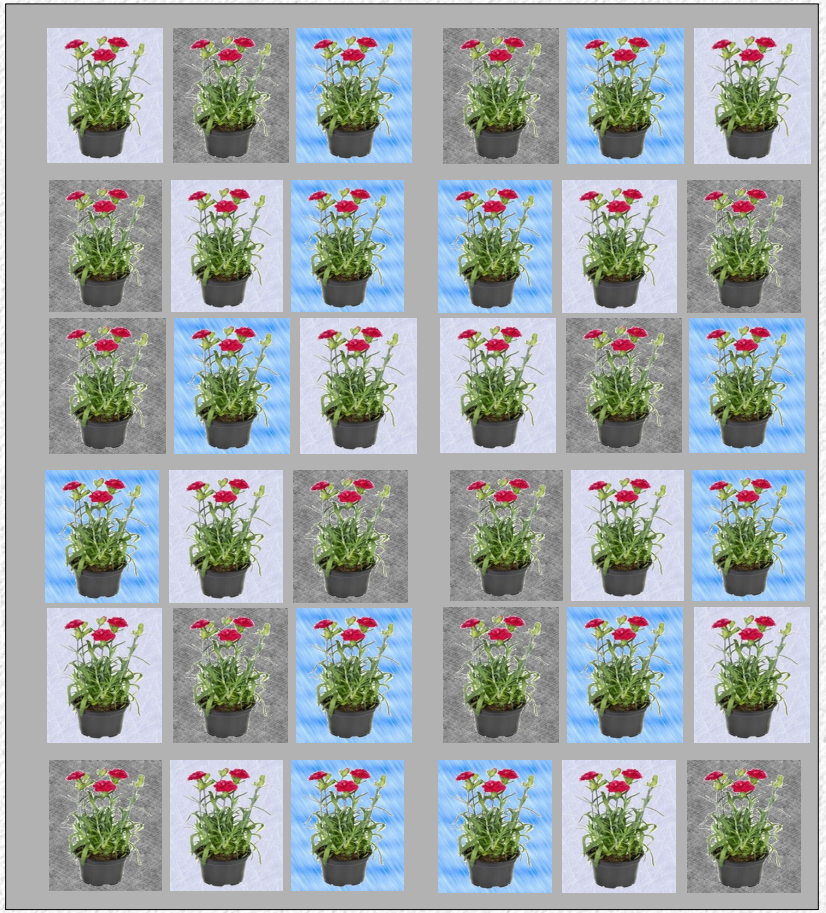
- Subjective response variables are problematic – the scientist's expectations can affect their recording of responses
  - Solution: have a different people apply the treatments and record the data – the recording person doesn't know what the treatments are
  - This is called **blinding** of the researcher
- If the subjects are also people, their expectation of getting better (in the treatment group) or not (in the control group) can affect their state
  - Particularly problematic in pain studies, or any case in which patient reporting of improvement is needed → subjectivity
  - Solution: blind the patient as well
- Double-blinded studies are the gold standard in clinical trials

Blocking

- Blocking is a design method, and an accompanying statistical method
  - Common for an experimental design to have an accompanying ANOVA design used to analyze the data properly
- Blocking allows for us to use statistical elimination with categorical variables
  - Factors that are nuisance variables are the blocks
- Done correctly, blocks are independent of (orthogonal to) the scientifically interesting variables (and each other)
  - Statistically reduce the amount of random, unexplained variation
  - Don’t interfere with measurement of the treatment effect

Example: carnations experiment

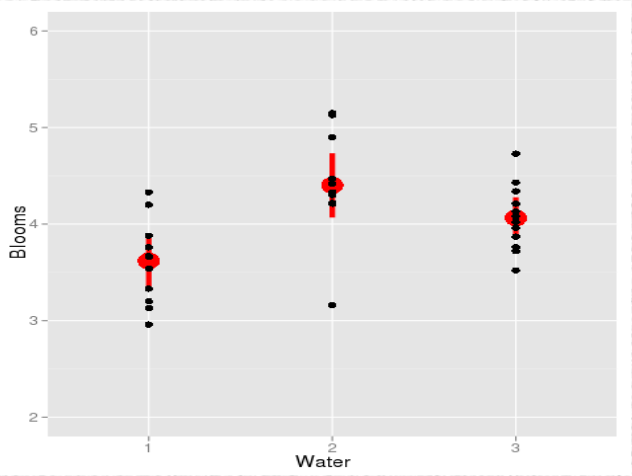
- A carnation grower is interested in effects of watering (predictor) on the number of blooms (response) put on by carnations
  - Predictor is categorical, with three levels: 1=low, 2=medium, 3=high
  - Replication: 12 plants grown at each water level
- Possible confounded effects
  - Shade – different areas have different levels of sunlight
  - Bed – different beds available to do the experiment
- Experimental ideal solution: standardize
  - Grow all at the same level of shading
  - Use only one flower bed



Holding shade and bed constant

```
Response: blooms
          Df Sum Sq Mean Sq F value    Pr(>F)
Water      2 3.6977  1.8488  9.9262 0.0004216 ***
Residuals 33 6.1465  0.1863
Total     35 9.8442

R-squared = 0.3756
```



- Water affects number of blooms
- Using one bed, one shade level → low overall variance, clean result, low p-value

What if you can’t hold everything constant?

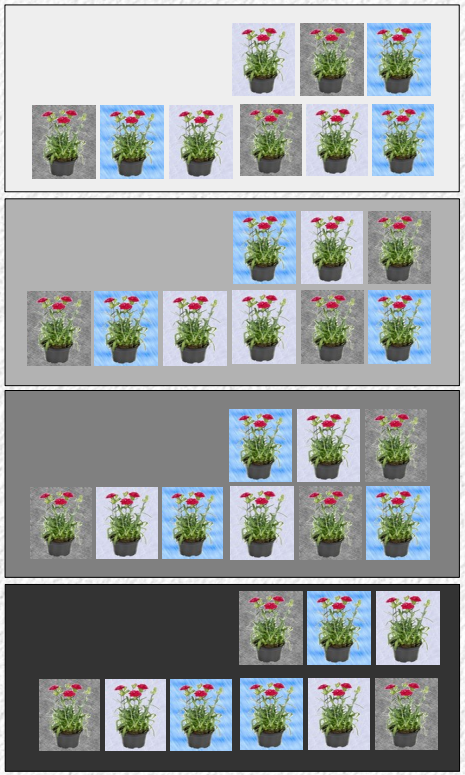
- Can’t always because:
  - Not enough space with constant levels of shade for all 36 plants → can’t hold shade constant
  - Not enough space in a single bed → can’t hold bed constant
- Since we know these variables could be important, we can block on them (by design), and account for them (in analysis)

The Randomized Complete Block design

- The Randomized Complete Block design – used to account for nuisance variables that can't be held constant
- Blocks and treatment variables are independent (orthogonal)
  - Each treatment level appears in each block = complete
  - Randomly assigning subjects to the needed combination of treatment and block = randomized
  - Number of replicates in each cell (i.e. combination of treatment and block) is the same = balanced
- If the blocks are orthogonal to the treatment, they can't possibly give the appearance of a treatment effect → no confounding!

Example: block on shade, use a single bed

- Shade is the blocking variable
- Every combination of water level and shade level is used = complete
- Pots randomly assigned to treatment / block combinations = randomized
- Equal numbers of each combination of shade and water level (3) = balanced
  - Equal numbers of each shade level (9)
  - Equal numbers of each water level (12)
  - Equal numbers of each combination (3)
- Intersperse the treatments within a block to avoid position effects



Accounting for blocks in your analysis

- The block design prevents confounding even if the block is not accounted for statistically in your analysis
- But, to get the full benefit of blocking you need to account for the variation in the data between blocks statistically
- **Partition the variance** in number of blooms explained into effects of:
  - Water treatment – SS accounted for by water
  - Shade level – SS accounted for by shade level
- Orthogonal designs mean that:
  - The SS for water treatment will be identical whether shade level is included or not
  - No correlation between predictors, so Type I and Type II SS are the same (order doesn’t matter)

	WATER			
	1	2	3	
	4.35	4.46	4.11	
	3.28	4.36	4.36	
	3.81	4.36	4.49	
	3.31	4.57	4.11	
	3.67	4.67	4.36	
	4.1	5.49	5.07	
	3.6	4.35	3.45	
	3.13	5.06	4.36	
	3.81	4.23	3.65	
	4.11	4.68	3.86	
	3.54	4.25	3.91	
	3.32	2.94	3.65	Grand mean
Water means	3.67	4.45	4.12	4.08

Remember: ANOVA calculations

$$SS\ Total = \sum (Obs._i - Grand\ mean)^2 = 11.47$$

$$SS\ Water = 12 \sum (Water\ mean_i - Grand\ mean)^2 = 3.70$$

$$SS\ Resid. = SS\ Total - SS\ Water = 7.77$$

*Water means are **marginal means** = means in the margins of the table*

- Water means average across the shade levels

Explained variation is based on differences between means at each water level and grand mean

Anything not explained by water level is part of the residual SS

ANOVA for water

```
Response: blooms
          Df  Sum Sq Mean Sq F value    Pr(>F)
Water      2  3.6977  1.8488  7.84590 0.0016310
Residuals 33  7.7763  0.2356
Total     35 11.4740

R-squared = 0.4663
```

- Water is significant
- Shade is orthogonal with water by design, can't be the reason for the apparent effect of water
- But, we aren't accounting for the effect of shade in our analysis – we can do better

	WATER			
SHADE	1	2	3	Shade means
1	4.35	4.46	4.11	
	3.28	4.36	4.36	4.18
	3.81	4.36	4.49	
2	3.31	4.57	4.11	
	3.67	4.67	4.36	4.37
	4.1	5.49	5.07	
3	3.6	4.35	3.45	
	3.13	5.06	4.36	3.96
	3.81	4.23	3.65	
4	4.11	4.68	3.86	
	3.54	4.25	3.91	3.81
	3.32	2.94	3.65	
Water means	3.67	4.45	4.12	Grand mean 4.08

Accounting for shade

$$SS\ Total = \sum (Obs._i - Grand\ mean)^2 = 11.47$$

$$SS\ Water = 12 \sum (Water\ mean_i - Grand\ mean)^2 = 3.70$$

$$SS\ Shade = 9 \sum (Shade\ mean_i - Grand\ mean)^2 = 1.65$$

$$SS\ Resid. = SS\ Total - SS\ Shade - SS\ Water = 6.12$$

Now have marginal means for:

- Shade - average across the water levels

- Water - average across the shade levels

SS for shade is subtracted from the residual SS, so residual SS is smaller

Accounting for shade – ANOVA table

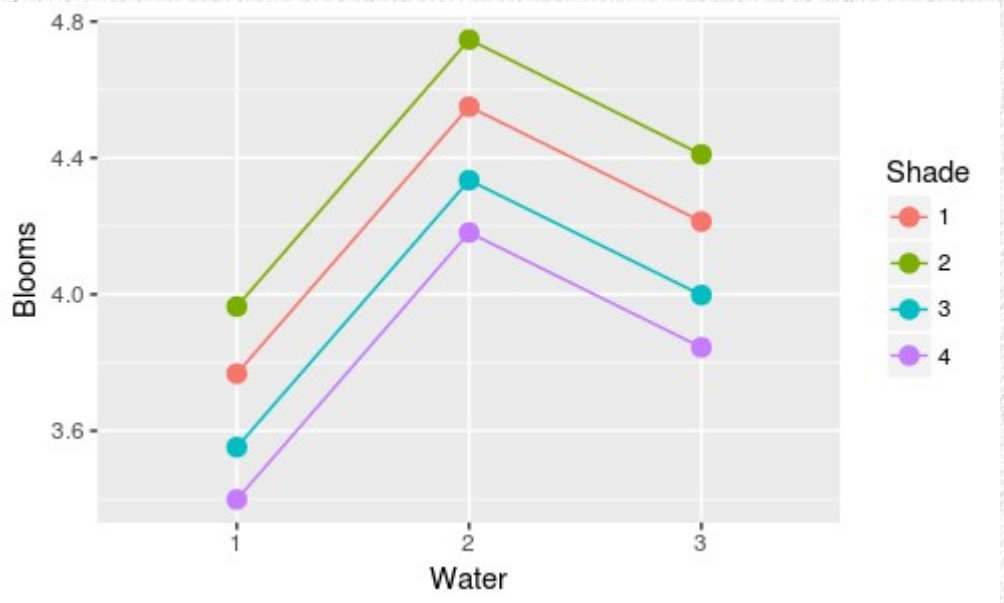
```
Response: blooms
          Df  Sum Sq Mean Sq F value    Pr(>F)
Water      2  3.6977  1.8488  9.0575 0.0008368 ***
Shade      3  1.6527  0.5509  2.6988 0.0633999 .
Residuals 30  6.1236  0.2041
Total     35 11.4740

R-squared = 0.4663
```

- Shade accounts for some variation in blooms, not significant
- F for water is $MS_{water}/MS_{residual}$
- F for shade is $MS_{shade}/MS_{residual}$

The GLM model blocking on shade

$$Blooms = 3.7675 + \begin{matrix} 0.7825\, Water_2 \\ 0.4458\, Water_3 \end{matrix} + \begin{matrix} 0.1967\, Shade_2 \\ -0.2156\, Shade_3 \\ -0.3689\, Shade_4 \end{matrix}$$



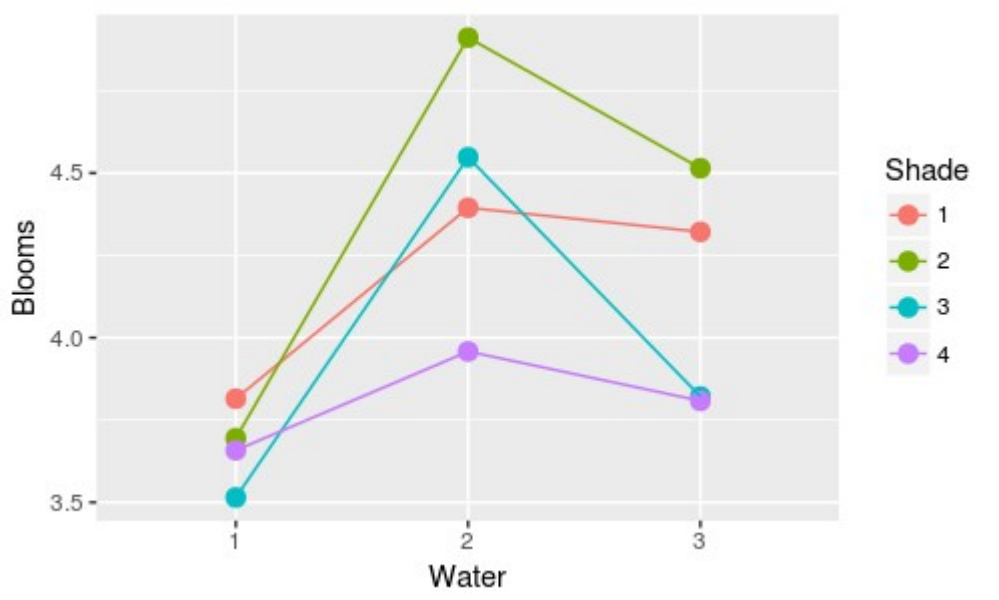
What is the intercept?

What are the slope coefficients?

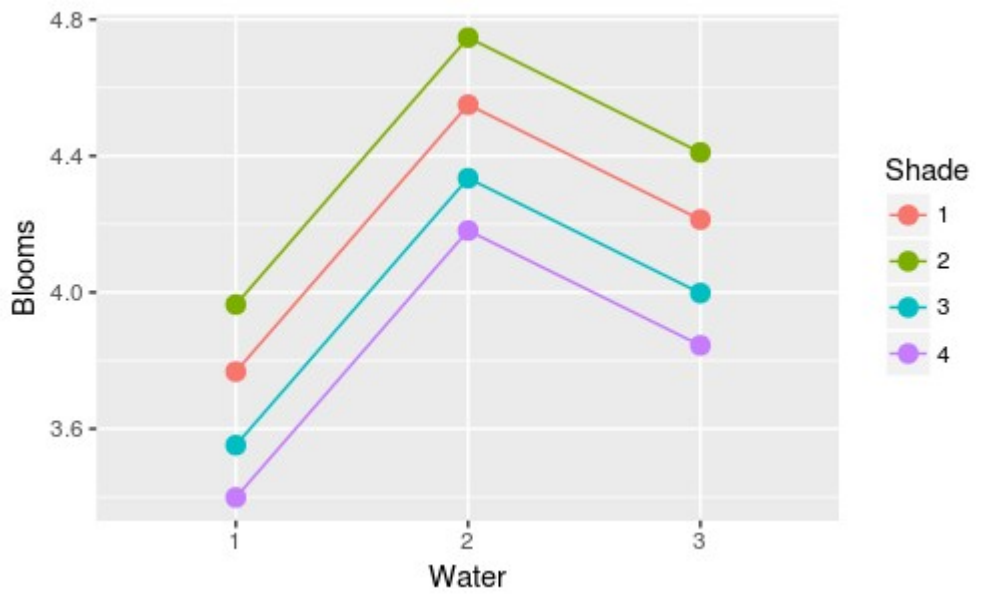
Graph shows predicted means for combinations of shade and water based on the model

Block analysis does not use means for combinations of water and shade!

The data



The GLM

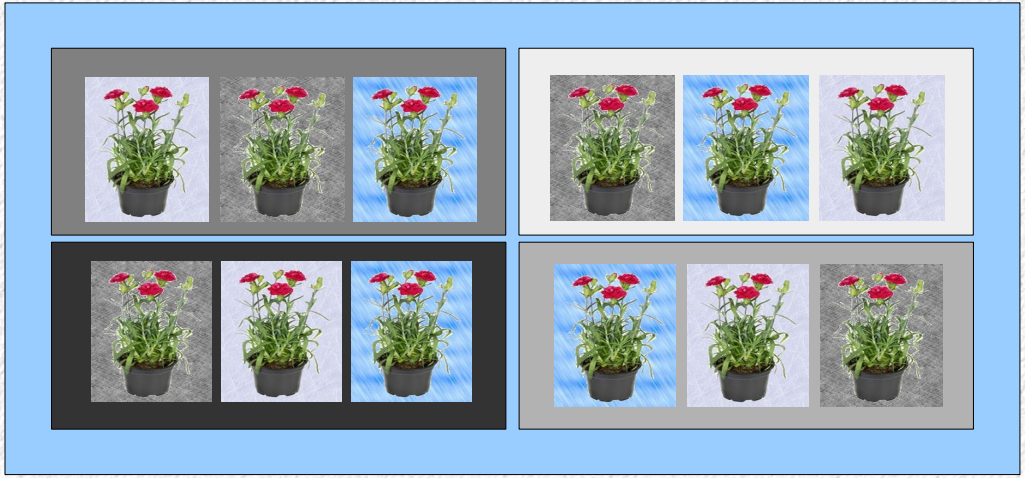


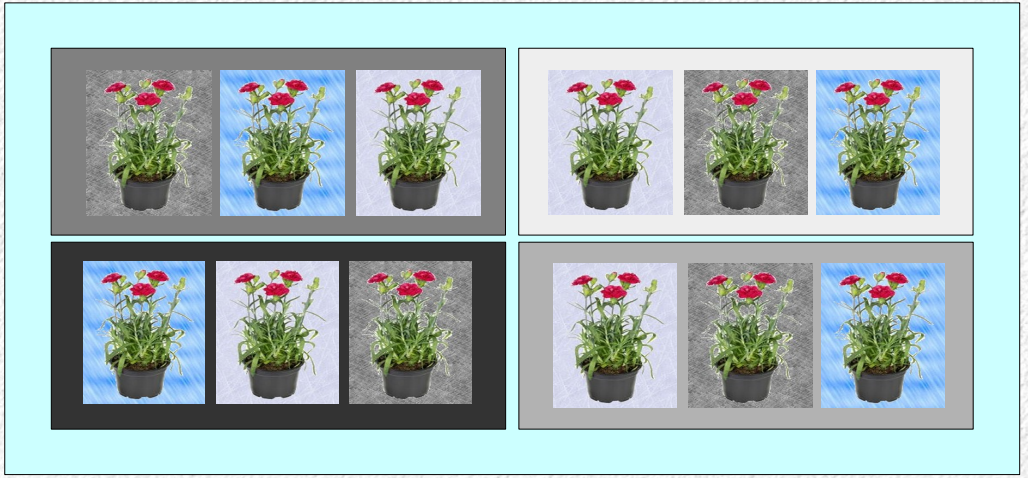
Left is group means, right is the predicted values for each group – not the same!

Model assumes additive, independent marginal effects → parallel lines

Not what the data shows! To model non-parallel lines need interactions – later

Block on shade and bed





Every combination of bed, shade, and water appears once

Most complicated we can make this design!

To block on anything else would require more replicates, or fewer levels for treatment or block

Water, shade and bed

WATER				
Bed 1	Mean	4.09		
SHADE	1	2	3	Shade 1 mean
1	4.36	4.47	4.12	4.13
2	3.32	4.58	4.12	
3	3.61	4.36	3.46	
4	4.12	4.69	3.87	
Bed 2	Mean	4.41		
SHADE	1	2	3	Shade 2 mean
1	3.61	4.69	4.69	4.32
2	4.00	5.00	4.69	
3	3.46	5.39	4.69	
4	3.87	4.58	4.24	
Bed 3	Mean	3.59		
SHADE	1	2	3	Shade 3 mean
1	3.32	3.87	4.00	3.91
2	3.61	5.00	4.58	
3	3.32	3.74	3.16	Shade 4 mean
4	2.83	2.45	3.16	3.76
Water means	3.62	4.40	4.07	**Grand mean 4.08**

$$SS\ Total = \sum (Obs._i - Grand\ mean)^2 = 15.60$$

$$SS\ Shade = 9 \sum (Shade\ mean_i - Grand\ mean)^2 = 1.65$$

$$SS\ Water = 12 \sum (Water\ mean_i - Grand\ mean)^2 = 3.70$$

$$SS\ Bed = 12 \sum (Bed\ mean_i - Grand\ mean)^2 = 4.13$$

$$SS\ Resid. = SS\ Total - SS\ Shade - SS\ Water - SS\ Bed = 6.12$$

Marginal means – shade means across beds and water levels

Bed means across shade and water levels

Water means across beds and shade levels

Effects of water, shade and bed

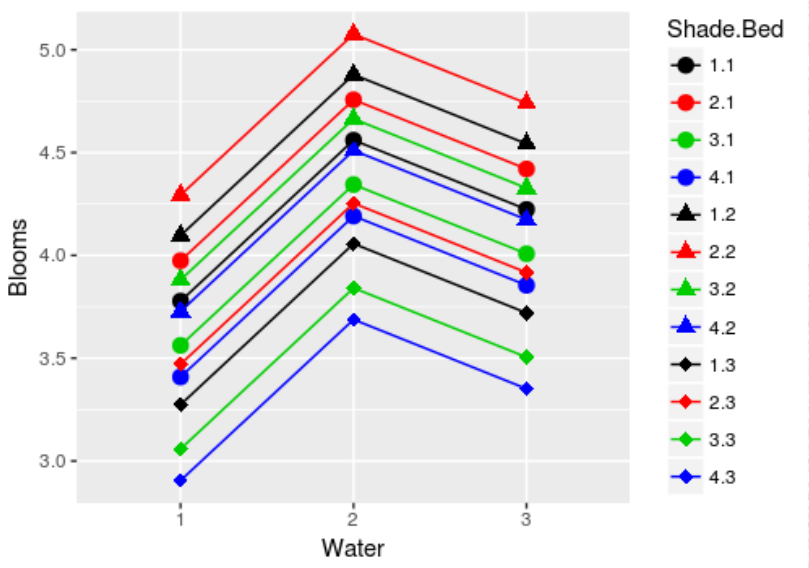
```
Response: Blooms
          Df  Sum Sq Mean Sq F value    Pr(>F)
Bed        2  4.1269  2.0634  9.4350 0.0007374
Shade      3  1.6527  0.5509  2.5189 0.0783731
Water      2  3.6977  1.8488  8.4537 0.0013420
Residuals 28  6.1236  0.2187
Total     35 15.6009

R-squared = 0.6075
```

- Greatest total variation in number of blooms
  - Variation due to water
  - Variation due to shade
  - Variation due to bed
- Highest model R^2
- Most informative – we now know the effect of water, shade, and flower bed on blooms in carnations

The model of water, shade, and bed

$$Blooms = 3.78 + \begin{matrix} 0.78\,Water_2 \\ 0.45\,Water_3 \end{matrix} + \begin{matrix} 0.20\,Shade_2 \\ -0.22\,Shade_3 \\ -0.37\,Shade_4 \end{matrix} + \begin{matrix} 0.32\,Bed_2 \\ -0.50\,Bed_3 \end{matrix}$$



Intercept?

Slopes?

Which is predicted to be the best combination of shade, bed, water?

Block designs that are balanced are orthogonal

```
      shade
water 1 2 3 4
    1 3 3 3 3
    2 3 3 3 3
    3 3 3 3 3
```

Equal number of replicates for each cell = balanced design

Type I SS - order doesn't matter

```
Response: blooms
          Df  Sum Sq Mean Sq F value   Pr(>F)
shade      3  1.6465 0.54882  1.6064 0.208609
water      2  3.7153 1.85767  5.4373 0.009661
Residuals 30 10.2496 0.34165
```

```
Response: blooms
          Df  Sum Sq Mean Sq F value   Pr(>F)
water      2  3.7153 1.85767  5.4373 0.009661
shade      3  1.6465 0.54882  1.6064 0.208609
Residuals 30 10.2496 0.34165
```

Block designs that are unbalanced are not orthogonal

```
      shade
water 1 2 3 4
    1 2 3 3 3
    2 3 3 3 3
    3 3 3 3 3
```

Unequal number of replicates for each cell = unbalanced design

The more unbalanced the design, the less independent the predictors become

Type I SS - order matters now

Response: blooms

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
shade	3	1.5958	0.53193	1.5819	0.215069
water	2	4.1524	2.07622	6.1747	0.005834
Residuals	29	9.7512	0.33625		

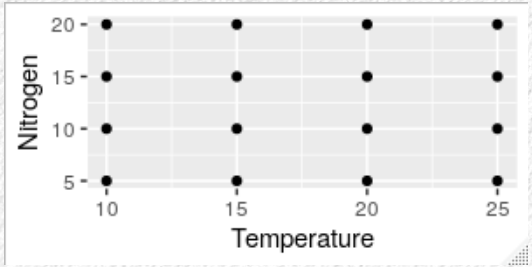
Response: blooms

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
water	2	4.2023	2.10117	6.2489	0.005539
shade	3	1.5459	0.51529	1.5325	0.227099
Residuals	29	9.7512	0.33625		

Orthogonal predictors in regression

- Correlation has to be exactly 0
- Can be achieved by design, if you have...
  - Evenly spaced predictor values
  - Equal number of measurements at each combination of values

...the correlation is 0, and predictors are orthogonal



Temperature	Nitrogen	Response
10	5	…
15	5	…
20	5	…
25	5	…
10	10	…
15	10	…
20	10	…
25	10	…
10	15	…
15	15	…
20	15	…
25	15	…
10	20	…
15	20	…
20	20	…
25	20	…

Summary: sources of confounding, design solutions

Source of confounding	Problem	Design solution
Individual, random variation	Any difference between individuals can look like a treatment effect	Replication
Spontaneous improvement, change over time	Changes due to factors other than the treatment	Simultaneous controls
Selection biases	Assigning subjects to treatment and control that are already different before experiment	Random assignment of subjects to treatment groups
Differences due to conditions other than application of the treatment	Treatment variable is not isolated as only possible cause of a change	Sham treatment, placebos, (double) blinding
Unavoidable differences due to equipment, observer, environmental gradients, etc.	Added statistical noise, possible confounding, spurious results	Blocking, accounting for blocks in analysis

Effect size

- Effect size = size of response produced by a predictor variable
  - Bigger effects are easier to detect with smaller n
- Measures of effect size often expressed as a signal to noise ratio (amount of effect of treatment / random variation)
  - Ex. Cohen's d = $(\bar{x}_1 - \bar{x}_2)/ s$ (good for t-tests)
  - Ex. Eta squared η^2 = FSS/TSS
  - Ex. Partial η^2 = FSS/(FSS+RSS)

  } (good for GLM's)
- Effect size is a property of our study systems (which we don't control) and experimental design choices (which we do)

Increasing **effect size** through experimental design

- Increasing the signal
  - Use a bigger dose of the treatment
    - Avoid over-dosing (not so much water that you drown the flowers)
    - Avoid dosages that are unreasonable (not more water than you can afford)
  - Find study areas that are sufficiently different in the predictor variable (rainfall, nitrogen deposition, elevation, etc.)
- Decrease the noise
  - Record data carefully, and with adequate precision
  - Standardize research protocols
  - Use homogeneous experimental subjects (or, measure and account for heterogeneities)