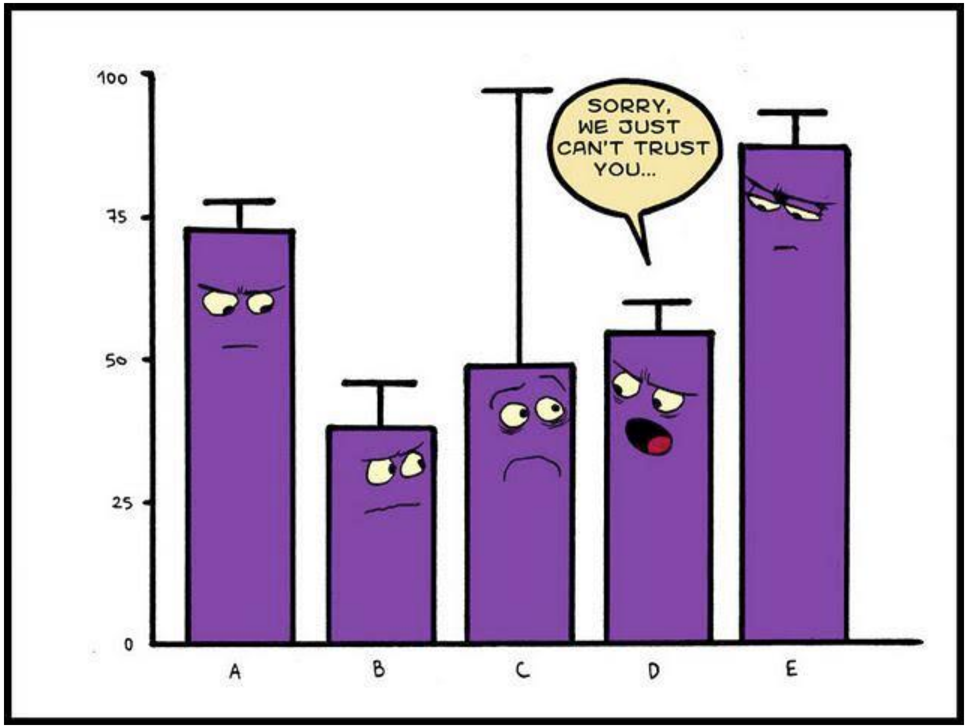
ANOVA Assumptions



“It is the mark of a truly intelligent person to be moved by statistics”
George Bernard Shaw (co-founder of the London School of Economics)

ANOVA Assumptions

1. The experimental errors of your data are normally distributed

2. Equal variances between treatments

   Homogeneity of variances
   Homoscedasticity

3. Independence of samples

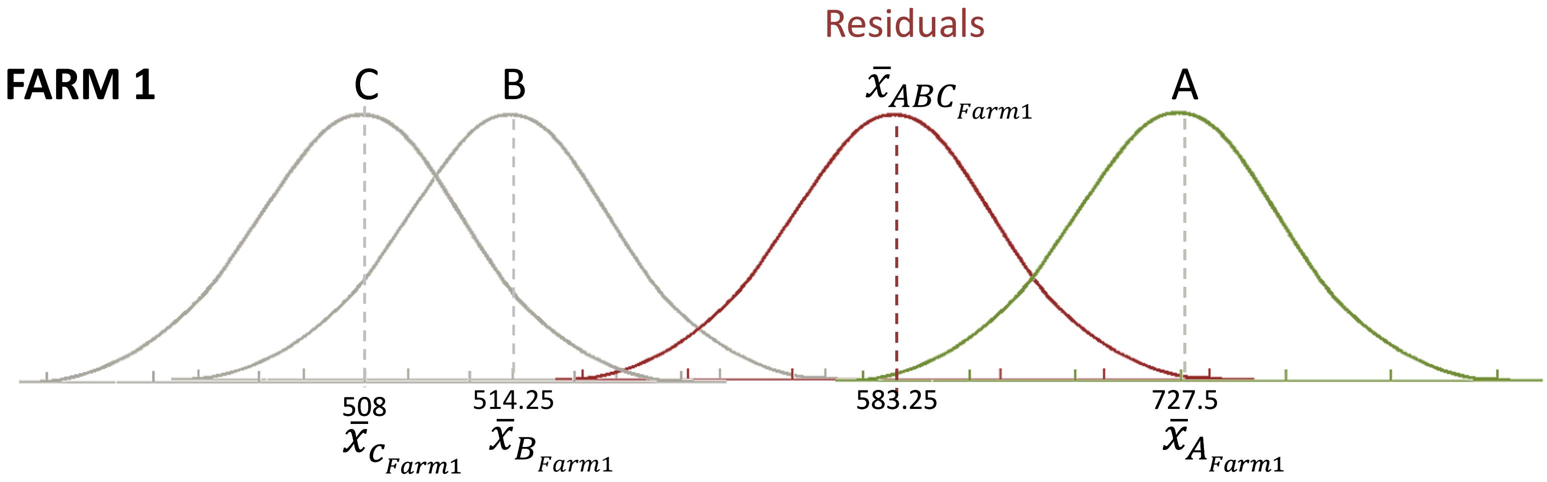
   Each sample is randomly selected and independent

Assumption #1: Experimental errors are normally distributed

"If I was to repeat my sample repeatedly and calculate the means, those means would be normally distributed."

Determine this by looking at the residuals of your sample:

residuals : subtract overall mean from the sample means



Calculate residuals in R:

```
res = residuals(lm(YIELD~VARIETY))
```

One-Way ANOVA

```
model=aov(YIELD~VARIETY) #Build a model with the normal ANOVA command
res=model$residuals #Create an object of the residuals of Y
```

Assumption #1: Experimental errors are normally distributed

Testing for Normality – **Shapiro Wilks Test**

Tests the hypotheses:

$$H_O: distribution\ of\ residuals = normal\ distribution$$
$$H_a: distribution\ of\ residuals \neq normal\ distribution$$

Non-Significant p-value = NORMAL distribution

$$\boxed{W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

a_i= constants generated from the means, variances and covariances of the order statistics of a sample of size n from a normal distribution (complex)

$x_{(i)}$= ordered sample values ($x_{(1)}$ is the smallest)

Small values of W are evidence of departure from normality

Shapiro-Wilks Test in R:

One-Way ANOVA

```
res = residuals(lm(YIELD~VARIETY))

model=aov(YIELD~VARIETY) #Build a model with the normal ANOVA command
res=model$residuals #Create an object of the residuals of Y

shapiro.test(res)
```

Assumption #1: Experimental errors are normally distributed

Alternative Tests

Shaprio-Wilks normality test – if your data is mainly unique values

D'Agostino-Pearson normality test – if you have lots of repeated values

Lilliefors normality test – mean and variance are unknown

Spiegelhalter's T' normality test – powerful non-normality is due to kurtosis, but bad if skewness is responsible

Assumption #1: **Experimental errors are normally distributed**

You may not need to worry about Normality?

> *"If I was to repeat my sample repeatedly and calculate the means, those means would be normally distributed."*
>
> Determine this by looking at the residuals of your sample

Central Limit Theorem:

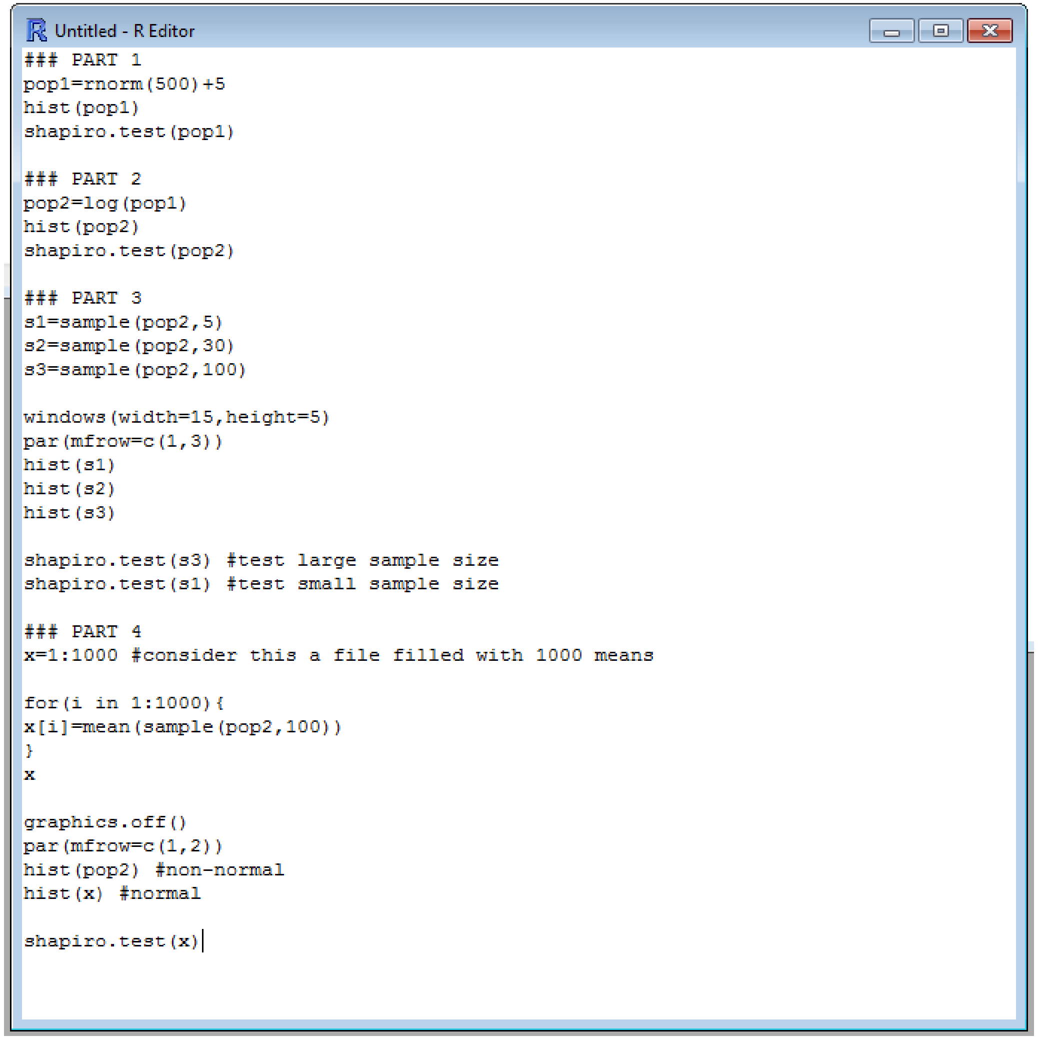
"Sample means tend to cluster around the central population value."

Therefore….
When sample size is large, the distribution of the sample means will always be large!

For large sample sizes testing for normality doesn't really work… best to just look at your data (*think histogram*)

Assumption #1: Experimental errors are normally distributed

You may not need to worry about Normality?



```
### PART 1
pop1=rnorm(500)+5
hist(pop1)
shapiro.test(pop1)

### PART 2
pop2=log(pop1)
hist(pop2)
shapiro.test(pop2)

### PART 3
s1=sample(pop2,5)
s2=sample(pop2,30)
s3=sample(pop2,100)

windows(width=15,height=5)
par(mfrow=c(1,3))
hist(s1)
hist(s2)
hist(s3)

shapiro.test(s3) #test large sample size
shapiro.test(s1) #test small sample size

### PART 4
x=1:1000 #consider this a file filled with 1000 means

for(i in 1:1000){
x[i]=mean(sample(pop2,100))
}
x

graphics.off()
par(mfrow=c(1,2))
hist(pop2) #non-normal
hist(x) #normal

shapiro.test(x)
```

For large N:

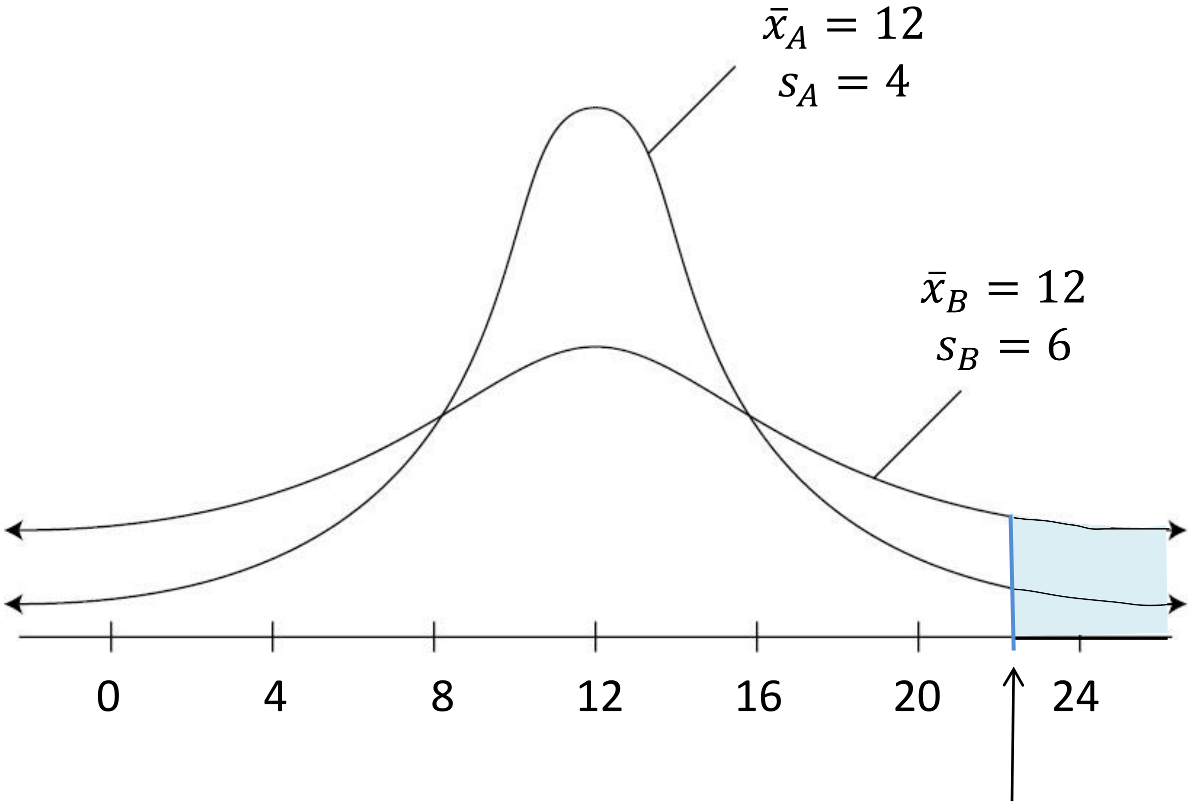
The assumption for Normality can be *relaxed*

ANOVA not really compromised if data is non-normal

Assumption of Normality is important when:

1. Very small N
2. Highly non-normal
3. Small effect size

Assumption #2: Equal variances between treatments



Let's say 5% of the A data fall above this threshold
But >5% of the B data fall above the same threshold

So with larger variances, you can expect a greater number of observations at the extremes of the distributions
This can have real implications on inferences we make from comparisons between groups

Assumption #2: Equal variances between treatments

Testing for Equal Variances – **Bartlett Test**

Tests the hypotheses: $H_O: variance_A = variance_B$

$H_a: variance_A \neq variance_B$

Non-Significant p-value = Equal variances

But you must compare the residuals for one treatment at a time (e.g. VARIETY and FARM)

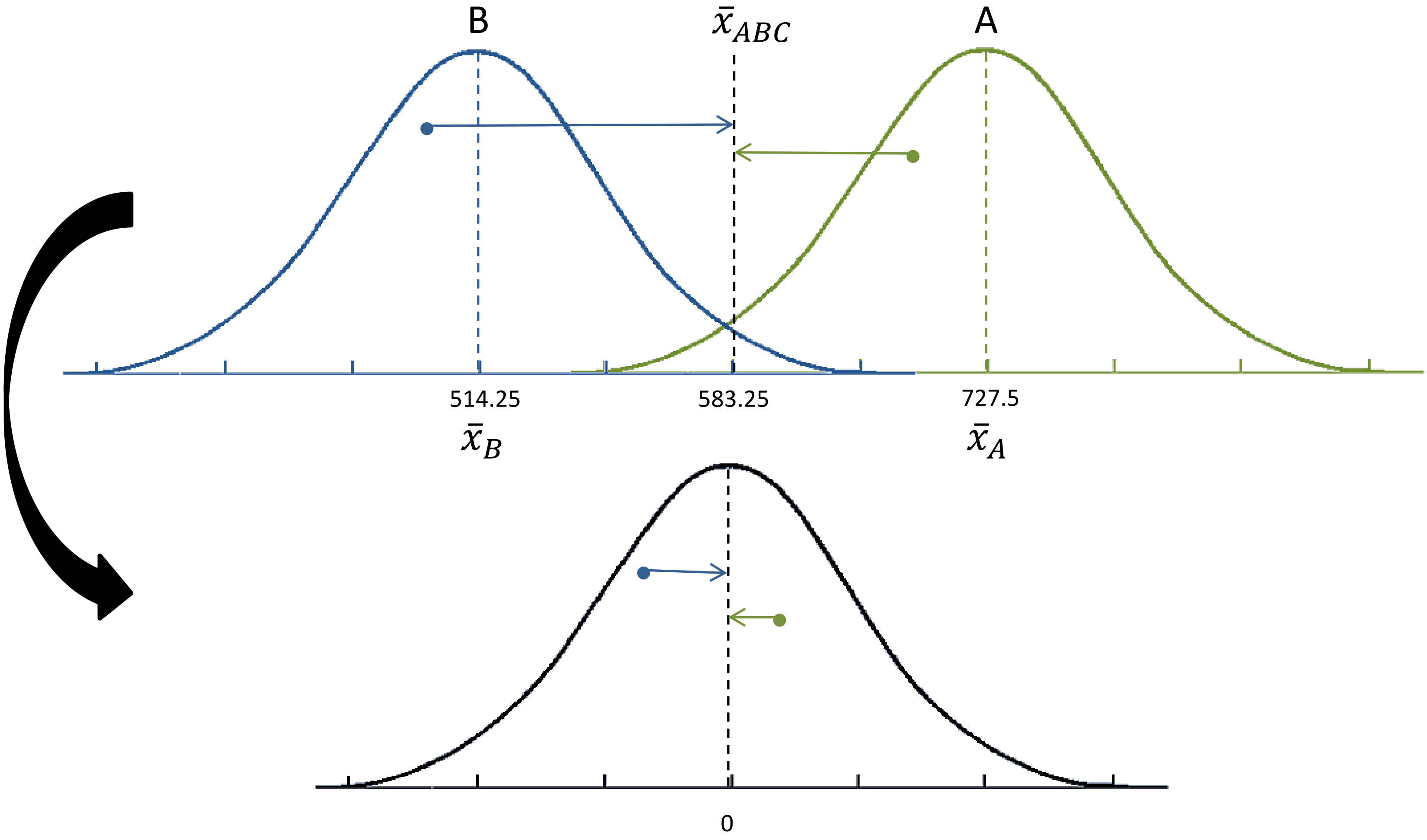
Bartlett Test in R:

```
bartlett.test(YIELD~VARIETY)
```

Assumption #2: Equal variances between treatments

Testing for Equal Variances – **Residual Plots**

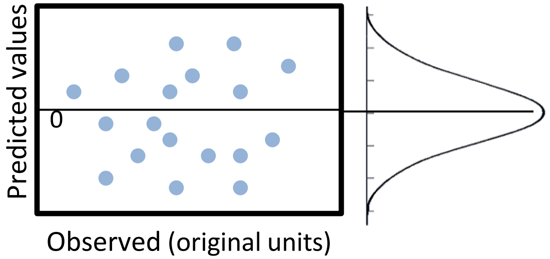
However, data residuals can also help us investigate whether variances are equal



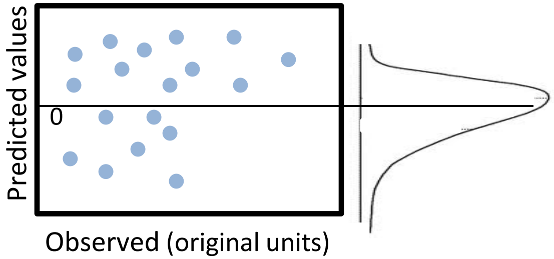
Assumption #2: Equal variances between treatments

Testing for Equal Variances – **Residual Plots**

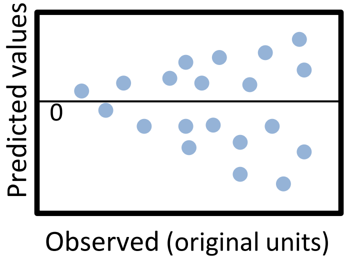
Residual plots in R (multiple plots):
`plot(lm(YIELD~VARIETY))` (2nd plot)



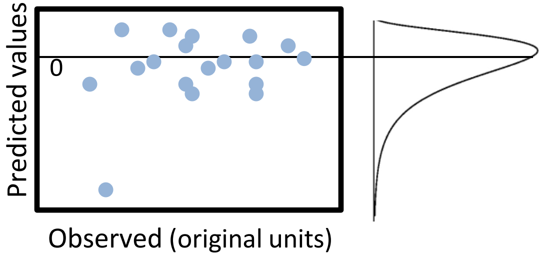
- NORMAL distribution: equal number of points along observed
- EQUAL variances: equal spread on either side of the $\text{mean}_{\text{predicted value}}=0$
- **Good to go!**



- NON-NORMAL distribution: unequal number of points along observed
- EQUAL variances: equal spread on either side of the $\text{mean}_{\text{predicted value}}=0$
- **Optional to fix**



- NORMAL/NON NORMAL: look at histogram or test
- UNEQUAL variances: cone shape – away from or towards zero
- **This needs to be fixed for ANOVA** (transformations)

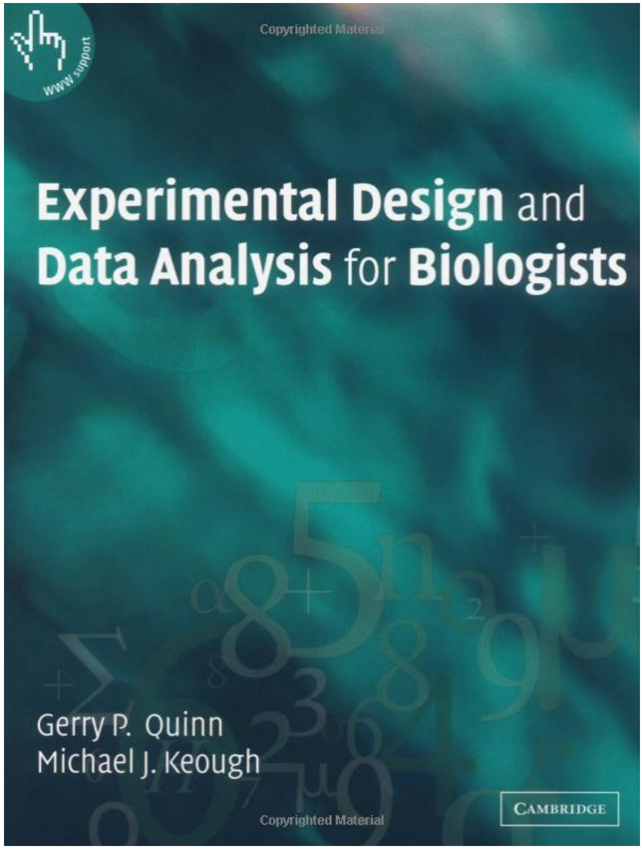


- OUTLIERS: points that deviate from the majority of data points
- **This needs to be fixed for ANOVA** (transformations or removal)

Assumption #3: Independence of samples

"Your samples have to come from a randomized or randomly sampled design."

Meaning rows in your data do NOT influence one another
Address this with experimental design (3 main things to consider)



Assumption #3: Independence of samples

Pseudoreplication

A particular combination of experimental design (or sampling) and statistical analysis which is inappropriate for testing the hypothesis of interest

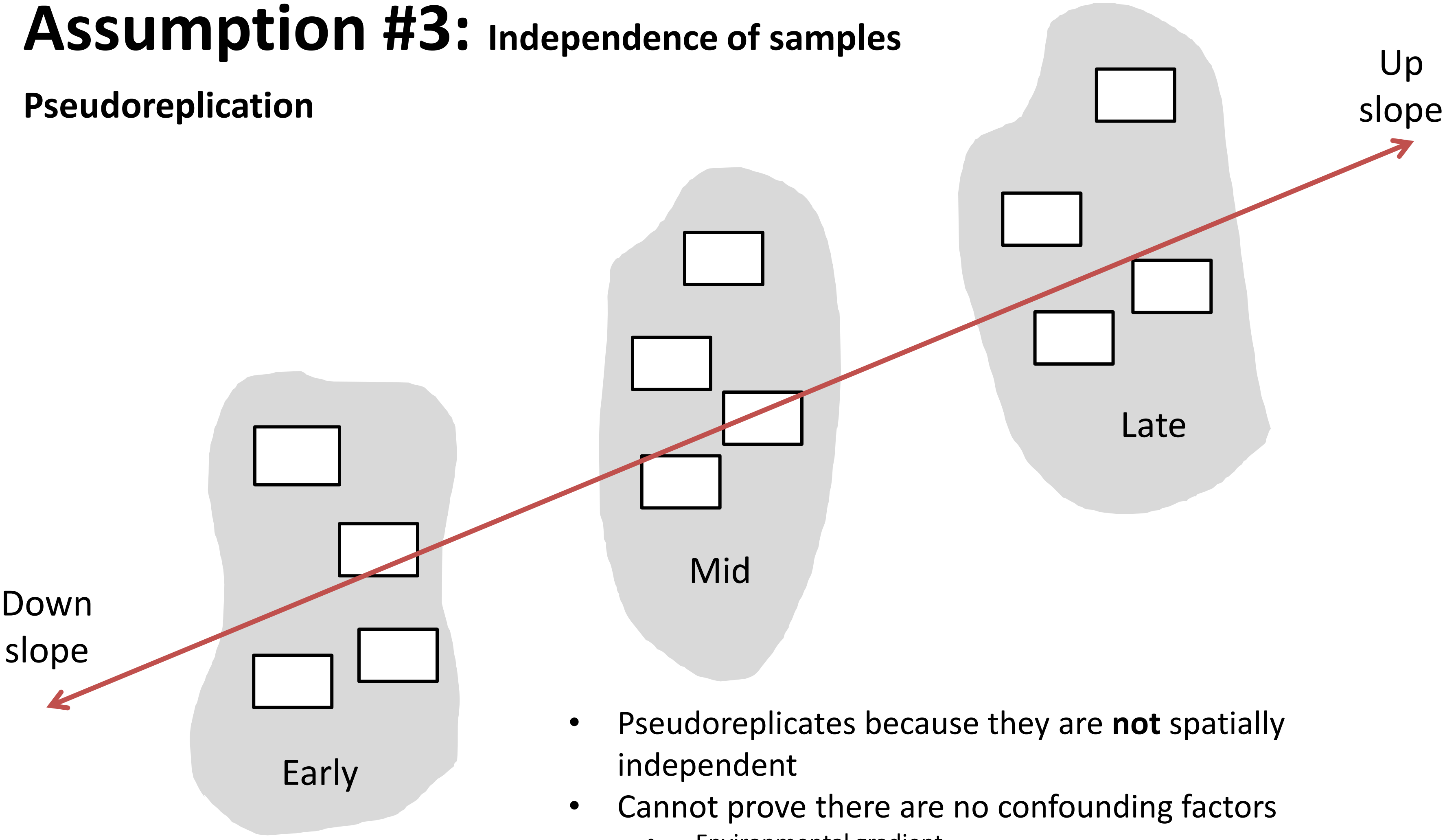
Occurs when a number of observations or the number of data points are treated inappropriately as independent replicates

Observations **may not** be independent if:

(1) repeated measurements are taken on the same subject
(2) observations are correlated in time
(3) observations are correlated in space.

Assumption #3: Independence of samples

Pseudoreplication

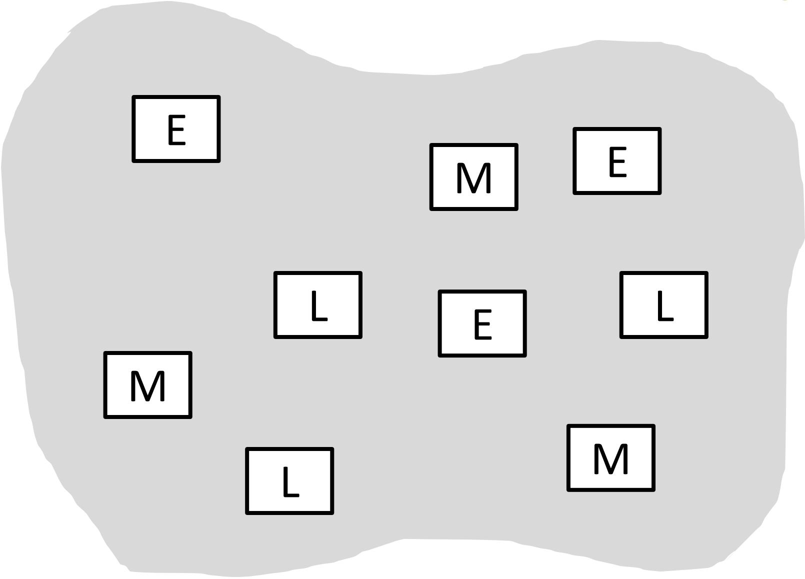


- Pseudoreplicates because they are **not** spatially independent
- Cannot prove there are no confounding factors
  - Environmental gradient
  - Topographical gradient
- Difficult to measure the variance between (signal) and the variance within (noise)
- Could measure alternative variables (treatements, covariates) – but often hard to do

Assumption #3: Independence of samples

Pseudoreplication

The right way to set up this experiment

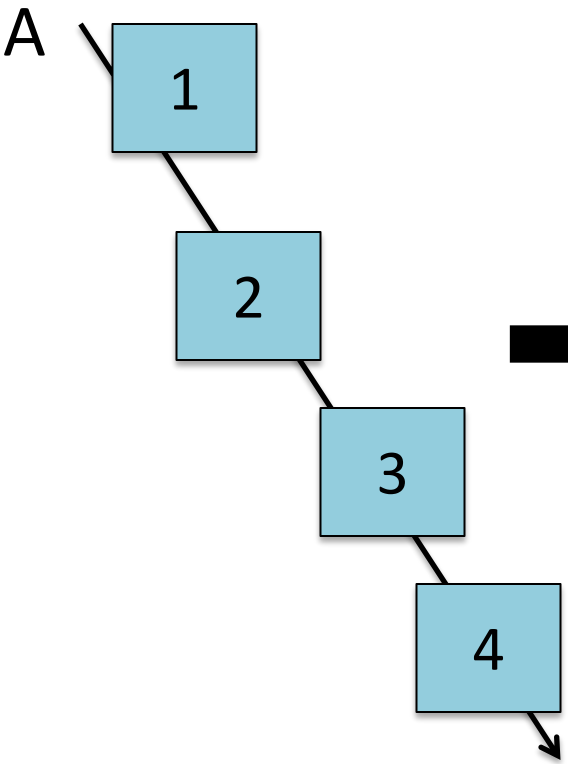


PlotID	Stand Stage	Sp1	Sp2	Sp3…
1	E			
2	M			
3	L			
4	M			
5…	L			

Best to avoid pseudoreplication and potential confounding factors by designing your experiment is a randomized design

Assumption #3: Independence of samples

Pseudoreplication – careful with experiments with transects



PlotID	TRANSECT	PLOT1	PLOT2	PLOT3	PLOT4	ALL PLOTS
1	A					
2	B					
3	C					
4	D					
5…	E…					

- Transects by definition are NOT spatially independent – potential for pseudoreplication
- Should therefore treat each transect as a sampling unit (row of data)
- Then you can compare between plots within transects and over the transect as a whole

Assumption #3: Independence of samples

Systematic arrangements

A	C	B
B	A	C
C	B	A

Systematic arrangement

- Poor practice
- “More random than randomized”
- Distinct pattern in how treatments are laid out
- If your treatments effect one another – the individual treatment effects could be masked or overinflated

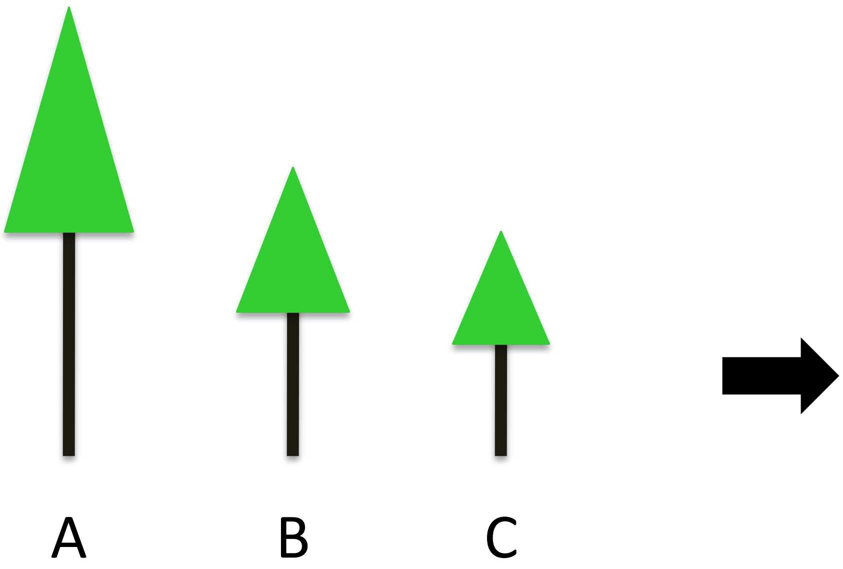
A	B	A
B	C	B
C	C	A

Randomized

- Good practice
- No distinct pattern in how treatments are laid out
- If your treatments effect is strong enough it will emerge as significant despite the leaching issue

Assumption #3: Independence of samples

Temporal Independence

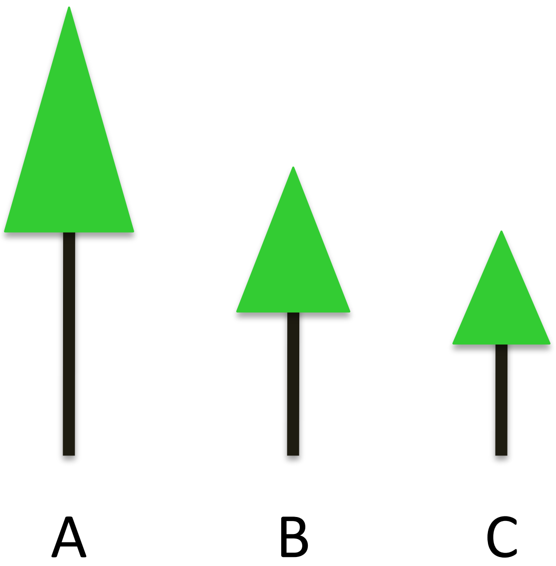


ID	VARIETY	YEAR	HT
1	A	1	17
2	A	2	18
3	A	3	19
4	B	1	12
5	B	2	14
6	B	3	13
7	C	1	7
8	C	2	8
9	C	3	9

- ANOVA assume each row of data you enter is an independent observation
- So if we run a simple ANOVA to determine the effect of VARIETY on HT we would me misinforming the analysis

Assumption #3: Independence of samples

Temporal Independence



The right way to set this data up to test the effect of VARIETY on HT

ID	VARIETY	YEAR	HT1	HT2	HT3
1	A	1	17	18	19
2	B	2	12	13	14
3	C	3	7	8	9

To Fix this problem:

1. You need multiple (independent) trees for each VARIETY to correctly answer this question
2. You would put HT in separate columns

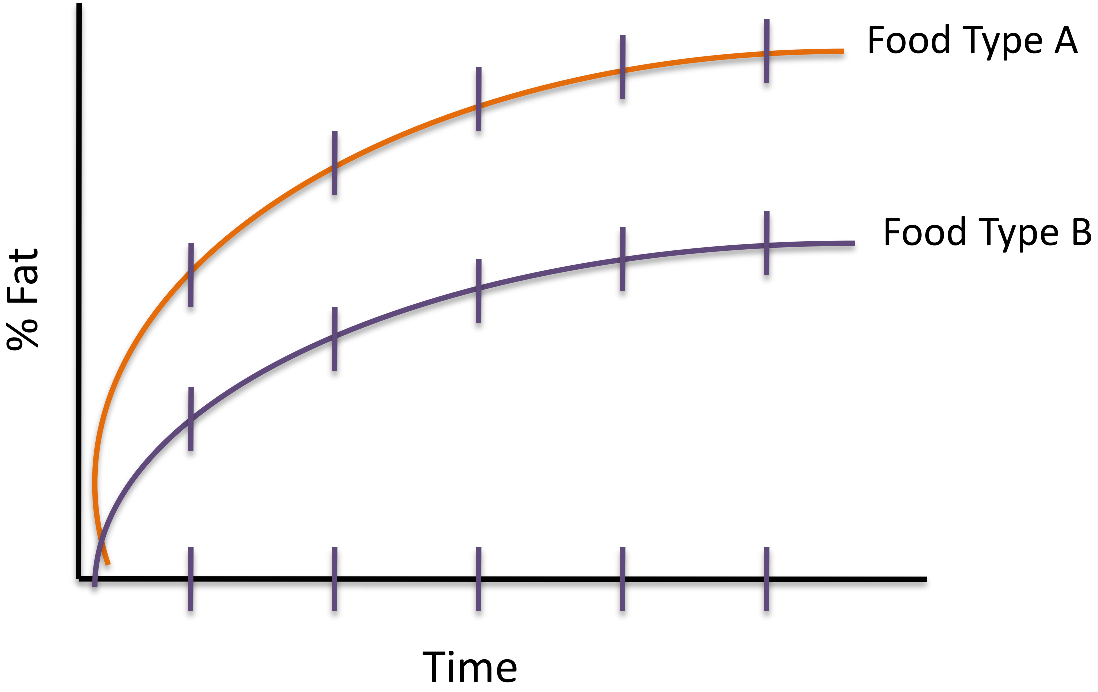
NOTE: ANOVA needs to have at least 1 degree of freedom – this means you need at least 2 reps per treatment to execute and ANOVA

Rule of Thumb: You need more rows then columns

Assumption #3: Independence of samples

Temporal Independence

Animal Science Example: Measuring how big the cows get over time on different food types



Rather then repeat ANOVAs

- Fit curves to data
- Important that measurements on all data are made at the same time
- Compare the coefficients of the curves with statistical tests

E.g. Are the slopes of the curves different?