Experimental design and statistics

Michael FW Festing
michaelfesting@aol.com
www.3Rs-Reduction.co.uk

How do we acquire knowledge?

Observation
Cross-sectional
Cohort studies

} Correlation but not causation

Randomised experiments } Causation

Survey of 271 randomly selected papers using animals

- □ 87% did not report random allocation of subjects to treatments
- □ 86% did not report “blinding” where it seemed to be appropriate
- □ 100% failed to justify the sample sizes used
- □ 5% did not clearly state the purpose of the study
- □ 6% did not indicate how many separate experiments were done
- □ 13% did not identify the experimental unit
- □ 26% failed to state the sex of the animals
- □ 24% reported neither age not weight of animals
- □ 4% did not mention the number of animals used
- □ 35% which reported numbers used, these differed in the materials and methods and the results sections

Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS.One.* 2009; **4:** e7824.

Statistics and design

- Experimental design:
  - Controlling variability
  - Requires a good knowledge of biology and sources of biological variation
- Statistics:
  - Deals with the variation which was not controlled by the design
  - At elementary level requires an understanding of data and statistical software
  - At advanced level requires a good understanding of mathematics

Types of controlled experiment

- Pilot study
  - Logistics and preliminary information
- Exploratory experiment
  - To provide data to generate hypotheses
  - May “work” or “not work”
  - Often many outcomes
  - Statistical analysis may be problematical (many characters measured, data snooping).
- Confirmatory experiment
  - Simple formal hypothesis stated *a priori.* p-values must be correct
- Experiments to estimate relationships between variables (regression and correlation)

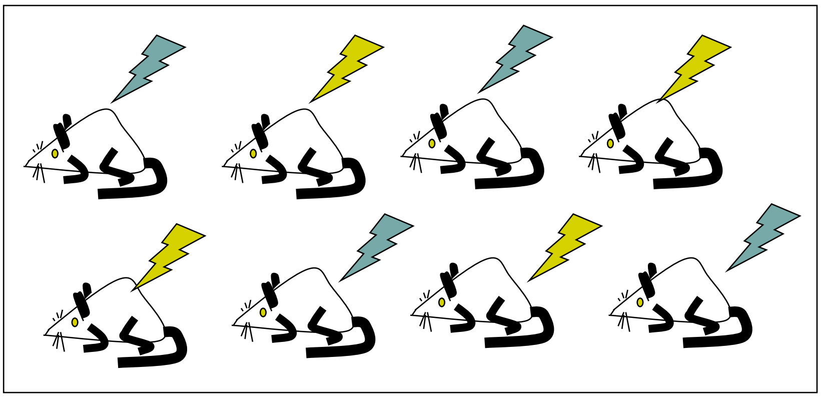
A well designed experiment

- Absence of bias
  - Experimental unit, randomisation, blinding
- High power
  - Low noise (uniform material, blocking, covariance)
  - High signal (sensitive subjects, high dose)
  - Large sample size
- Wide range of applicability
  - Replicate over other factors (e.g. sex, strain): factorial designs
- (Simplicity)
- (Amenable to a statistical analysis)

Experimental Unit

The smallest division of the experimental material such that any two experimental units can receive different treatments

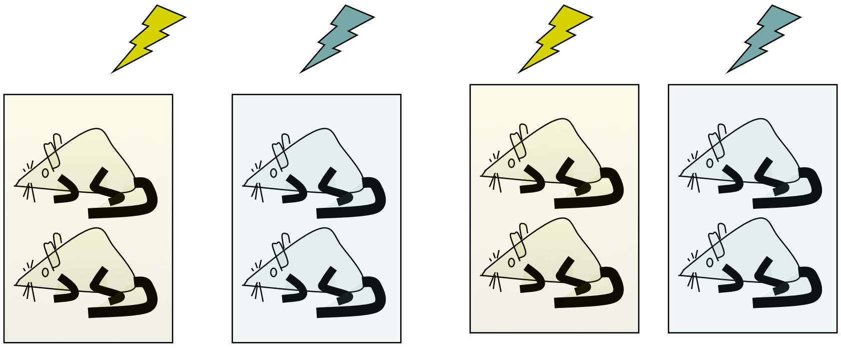
The animal as the experimental unit



Animals individually treated. May be individually housed or grouped

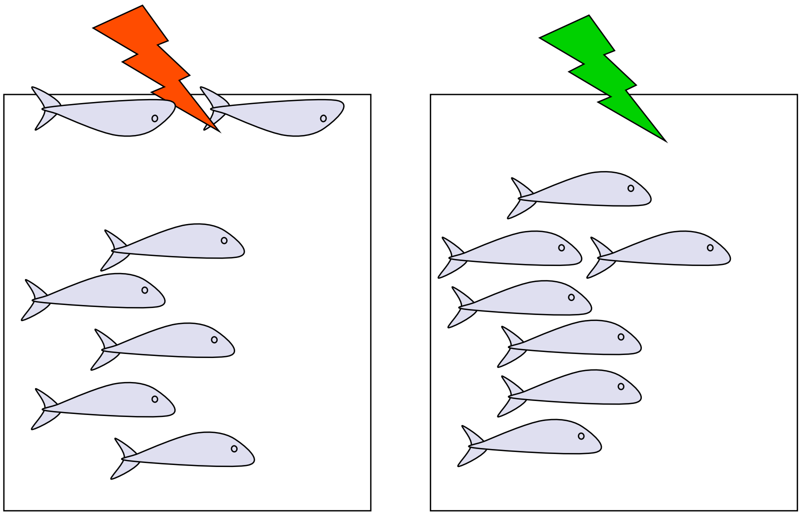
N=8

What is the Experimental Unit?

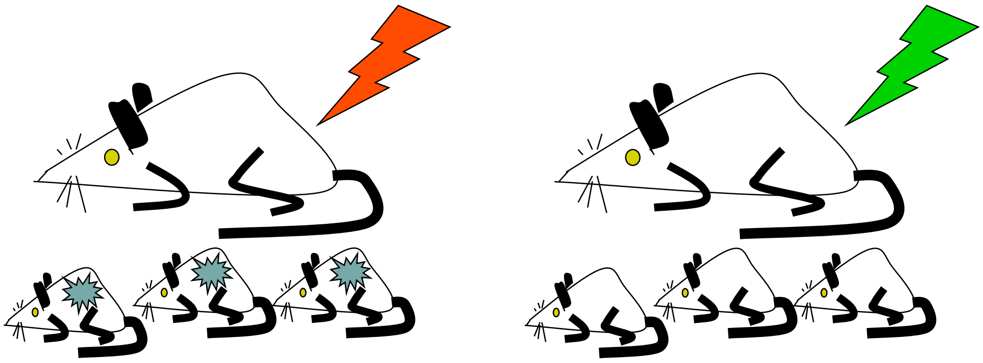


Treatment in water or diet.

Two tanks of fish

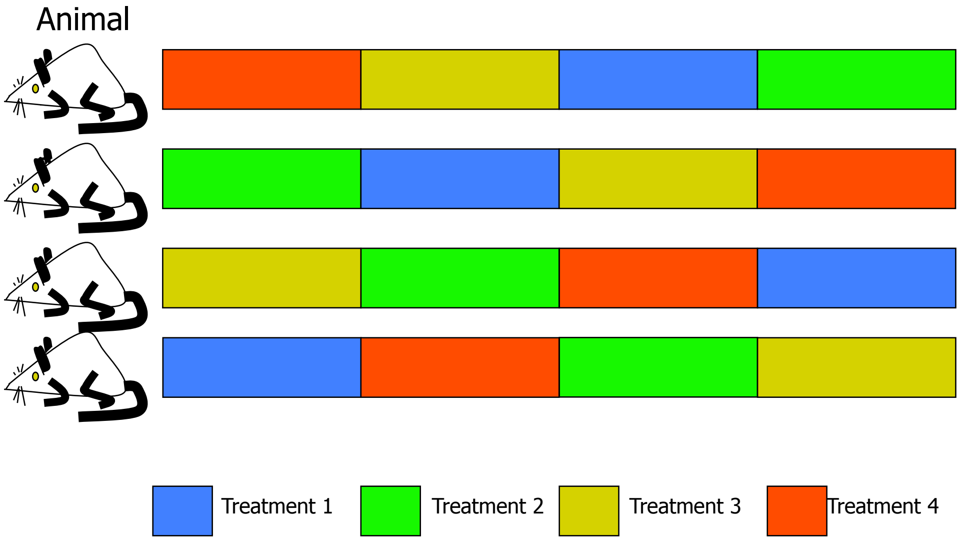


Teratology: mother treated, young measured



Animals given four treatments sequentially.

What is the experimental unit?



An animal for a period of time,   N=16

What is the experimental unit?

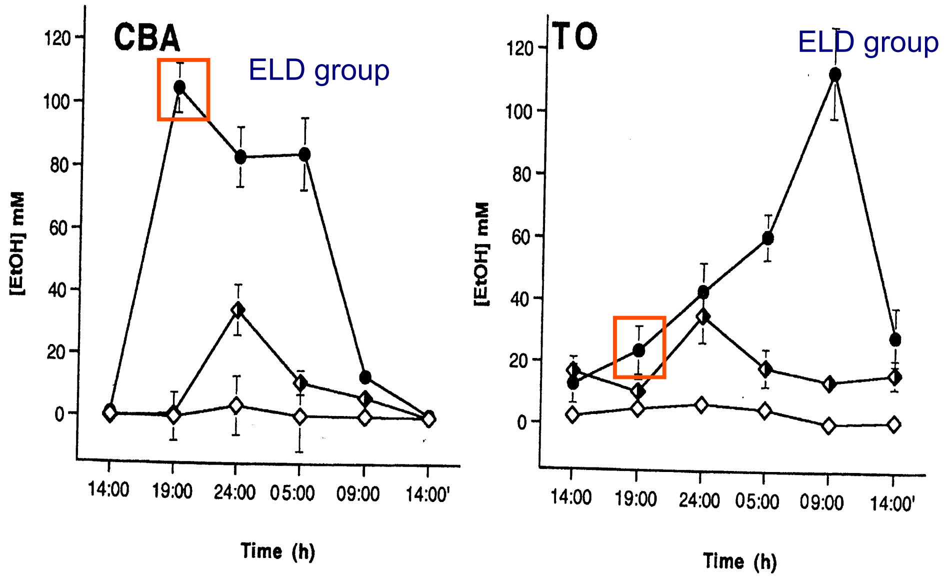
Humans who suffer from depression seem to be more sensitive to pain. An investigator wants to know if this is also the case in rats.

WKY rats are used as a model of depression
Wistar rats are not depressive.

So he obtains 10 rats of each strain, houses them two per cage for three weeks and tests them using a standard test of pain threshold.

What is the experimental unit in this experiment?

Failure to identify the experimental unit correctly (aim to look at strain differences in diurnal pattern of blood alcohol)



ELD group

ELD group

Single cage of 8 mice killed at each time point (36x8=288 mice in total)

Randomisation of 12 animals to three treatments (A-C) using EXCEL

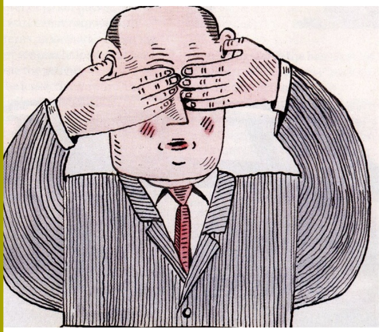
1. The treatment designations A-C were put in the first column, 4 subjects per treatment
2. A random number was put in the second one (preferably as “values”)
3. The columns were then sorted on the random number column to give column 3 in random order. The animal numbers are then added
4. In this case the first three animals will be assigned to A, the 4th. To C etc.

Original	=rand()	Sorted on =rand()		Animal number
A	0.527	A	0.067	1
A	0.100	A	0.100	2
A	0.067	A	0.122	3
A	0.122	C	0.210	4
B	0.665	B	0.248	5
B	0.875	C	0.265	6
B	0.478	B	0.478	7
B	0.248	A	0.527	8
C	0.210	C	0.628	9
C	0.628	B	0.665	10
C	0.265	B	0.875	11
C	0.895	C	0.895	12

Sometimes a random order doesn’t look very random, such as when the first three animals (here) all receive treatment A.
But use this sort of method and you won’t go far wrong.

Experimental units need to be randomised to treatments then blinded to help avoid bias

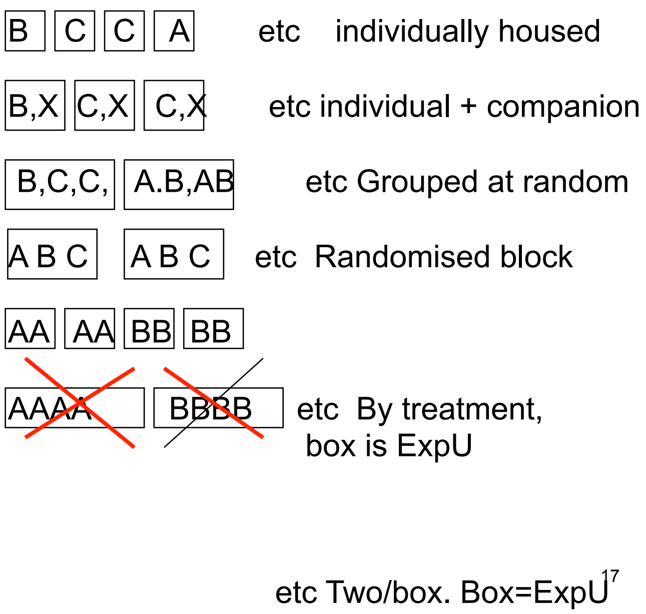
Animal	Treatment
1	B
2	B
3	B
4	D
5	C
6	A
7	A
8	D
9	D
10	C
11	A
12	C



Randomisation, blinding and cage assignment

Randomized	Mouse
B	1
C	2
C	3
A	4
B	5
A	6
B	7
A	8
C	9
B	10
C	11
A	12

Cages



B C C A etc individually housed

B,X C,X C,X etc individual + companion

B,C,C, A.B,AB etc Grouped at random

A B C A B C etc Randomised block

AA AA BB BB

AAAA BBBB etc By treatment, box is ExpU

etc Two/box. Box=ExpU

Failure to randomise and/or blind leads to more "positive" results

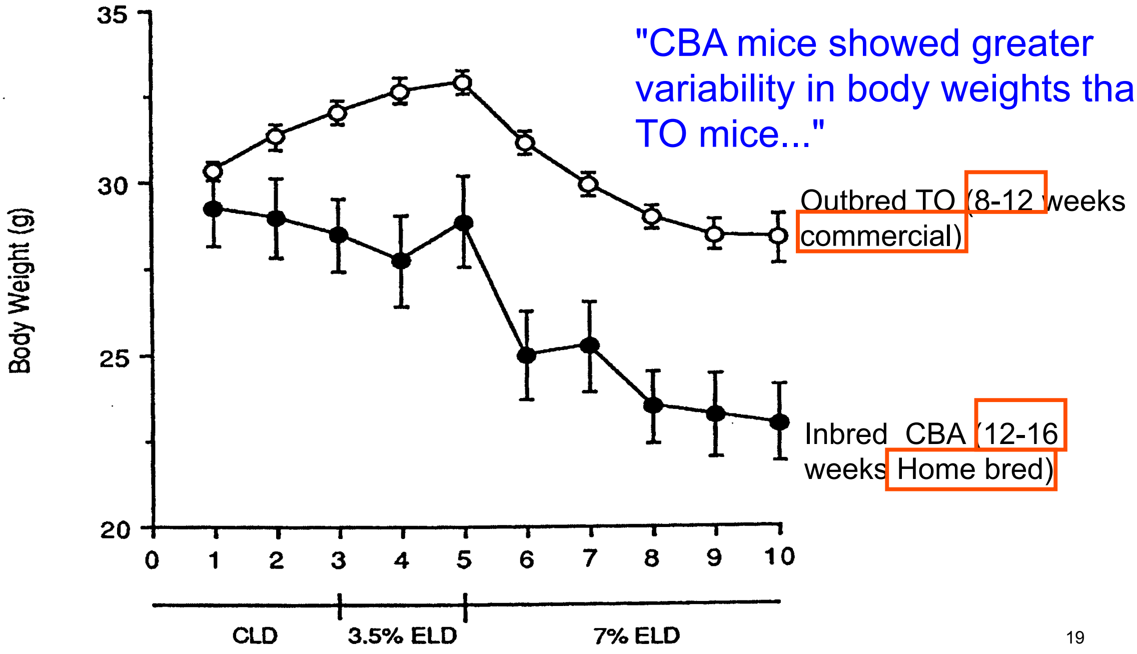
Blind/not blind	odds ratio	3.4 (95% CI 1.7-6.9)
Random/not random	odds ratio	3.2 (95% CI 1.3-7.7)
Blind Random/ not blind random	odds ratio	5.2 (95% CI 2.0-13.5)

290 animal studies scored for blinding, randomisation and positive/ negative outcome, as defined by authors

Bebarta et al 2003 Acad. emerg. med. 10:684-687

“Classification variables” (e.g. strain, sex) can not be randomised so special care is needed to ensure comparability

Six cages of 7-9 mice of each strain: error bars are SEMs



"CBA mice showed greater variability in body weights than TO mice..."

A well designed experiment

- Absence of bias
  - Experimental unit, randomisation, blinding
- High power
  - Low noise (uniform material, blocking, covariance)
  - High signal (sensitive subjects, high dose)
  - Large sample size
- Wide range of applicability
  - Replicate over other factors (e.g. sex, strain): factorial designs
- Simplicity
- Amenable to a statistical analysis

Controlling variability:Genetic Stocks of Laboratory Animals

- Outbred stocks
  - e.g. Swiss mice, Wistar Rats
- Inbred strains
  - e.g. BALB/c, F344
- Mutants and polymorphisms
  - e.g. *Foxn1^{nu}*, *Foxn1^{rnu}*
- Transgenic strains
  - e.g. TG.AC, BigBlue

Exercise 1A

You want to test a compound to see if it will delay rejection of transplanted hearts.

The experiment involves heart grafts between a donor and recipient rat.
All rats have heart grafts (but retain the own heart)
Half receive the test compound, half receive the vehicle

The following rat strains are available: Outbred Wistar and Sprague-Dawley and inbred ACI, F344 and LEW.

Which strains will you use as donor and recipient, and why?

Exercise 1B

You know it is not acutely toxic but need to do a long-term toxicity study with control and treated rats.

A toxicologist points out that you wish to model humans who are genetically heterogeneous. He suggests that you use outbred genetically heterogeneous Sprague-Dawley rats, the strategy used by virtually all toxicologists.

Do you decide to accept or reject his advice. Give your reasons

Exercises 1A and 1B

Questions:

A: does your new drug prolong graft survival?

Heart transplant. Choose donor and recipient rats from:
Outbred: Wistar, Sprague-Dawley,
Inbred: ACI, LEW, F344

B is the compound chronically toxic?

Toxicity test. Aim is to model humans. Outbred Sprague-Dawley rats suggested. Accept or reject this advice?

Variable results with heart transplants

"We transplanted hearts of young ... Fishers into ... recipient Sprague-Dawleys. An outbred strain was selected since such animals are usually heartier and easier to handle...

We are puzzled by our results....palpable heart beats were evident in the saline group long after acute rejections...were expected...Results in the experimental groups varied considerably..."

Choice of outbred stocks

"..it is more correct to test on a random-bred stock on the grounds that it is more likely that at least a few individuals will respond to the administration of an active agent in a group which is genetically heterogeneous"

Arcos JC, Argus MF, Wolf G, eds. (1968) Chemical induction of cancer. 491pp, London, Academic Press.

The problem with genetic heterogeneity
A “completely randomized” design

Control	Treated
Beagle	Goat
Chicken	Pig
Mouse	Crow
Horse	Frog
Gerbil	Hamster
Guinea-pig	Quail
Lion	Beaver
Duck	Cat

A matched pairs (randomized block) design

Control	Treated
Beagle	Beagle
Mouse	Mouse
Horse	Horse
Gerbil	Gerbil
Guinea-pig	Guinea-pig
Lion	Lion
Duck	Duck
Rabbit	Rabbit

A randomized block design

Control	Treated
A/J	A/J
A2G	A2G
BALB/c	BALB/c
CBA	CBA
C3H	C3H
C57BL/6	C57BL/6
DBA/2	DBA/2
NIH	NIH

There could be more than two treatment groups

A randomised block design

Strain	Control	Treated
A/J	22.8	21.8
A2G	24.0	23.2
BALB/c	22.3	21.5
CBA	20.6	20.5
C3H	24.0	23.9
C57BL/6	24.8	24.7
DBA/2	22.4	21.7
NIH	29.6	30.0
Mean	23.8	23.4
SD	2.7	3.0

How should this be statistically analysed?

Statistical analysis

Strain	Control	Treated	Control-treated
A/J	22.8	21.8	1.0
A2G	24.0	23.2	0.8
BALB/c	22.3	21.5	0.8
CBA	20.6	20.5	0.1
C3H	24.0	23.9	0.1
C57BL/6	24.8	24.7	0.1
DBA/2	22.4	21.7	0.7
NIH	29.6	30.0	-0.4
Mean	23.8	23.4	0.4
SD	2.7	3.0	0.5

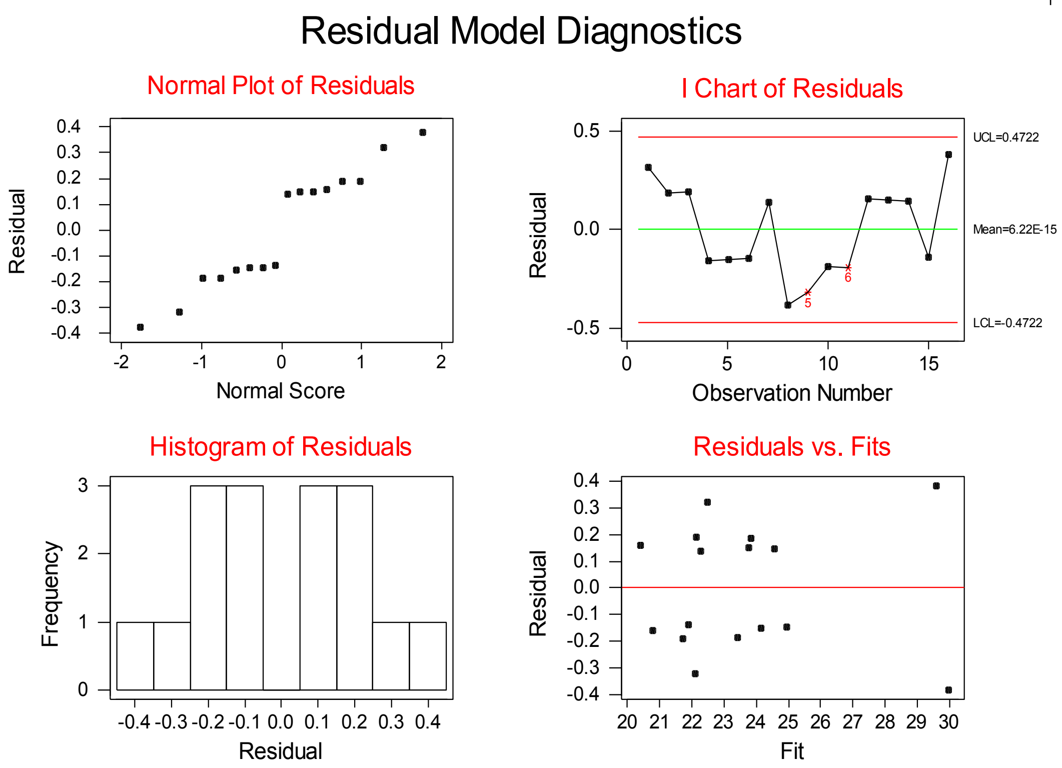
Paired t-test

```
One-Sample T: Difference

Test of mu = 0 vs mu not = 0

Variable        N     Mean     StDev   SE Mean
Difference      8    0.387    0.482     0.171

Variable          95.0% CI             T      P
Difference    (  -0.017,   0.791)    2.27  0.058
```

Two-way ANOVA without interaction for a randomised block design

Analysis of Variance for Weight

```
Source       DF        SS        MS        F       P
Strains       7   111.717    15.960   137.18   0.000
Treatmen      1     0.599     0.599     5.15   0.058
Error         7     0.814     0.116
Total        15   113.131
```



Statistical analysis should fit the purpose of the study

A Completely Randomised Design

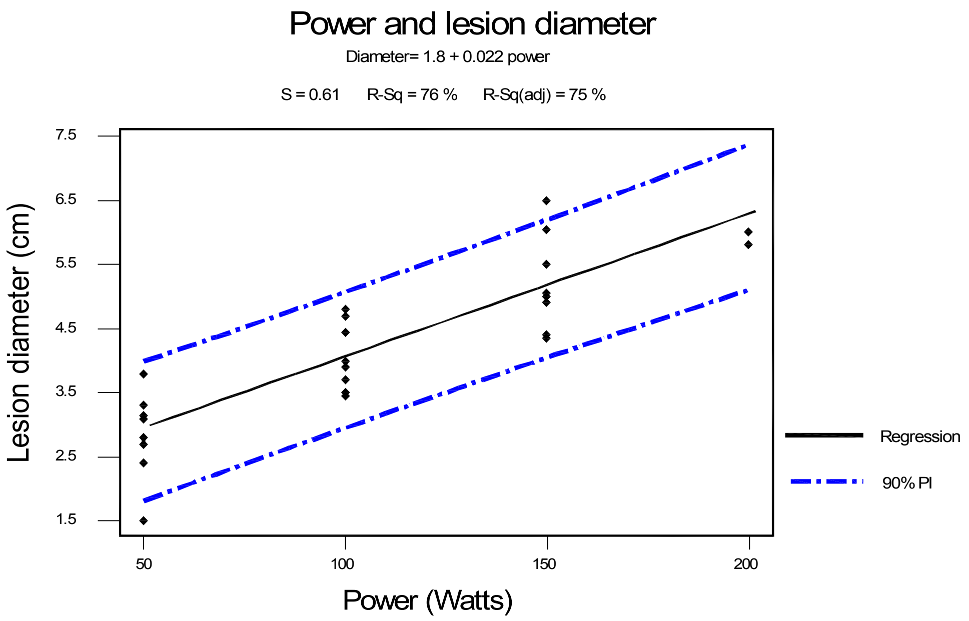
Experimental unit??

Lesion diameter following microwave treatment of pig liver

Power (watts)										Mean
50	3.3	3.2	2.8	2.8	2.4	2.7	3.2	3.8	1.5	2.9
100	4.7	4.0	3.5	4.4	3.9	4.8	4.4	3.7	4.0	4.2
150	5.5	5.0	4.4	4.5	6.0	6.5	5.0	5.0		5.3
200	5.8	6.0								5.9

Lesion diameter clearly increases with power, but aim is to quantify this

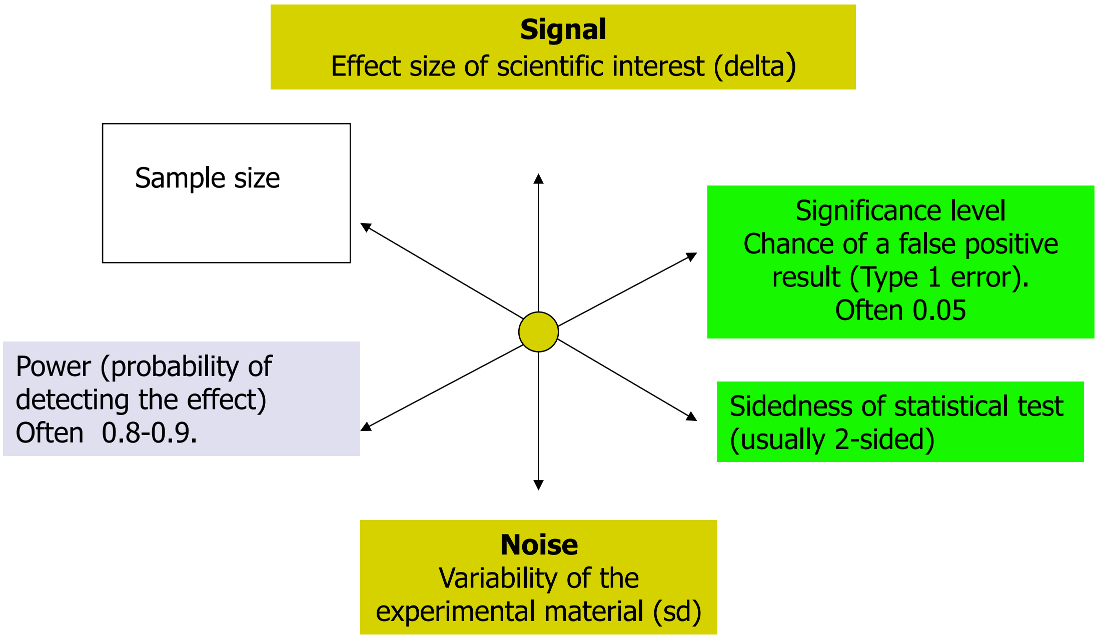
Estimation versus hypothesis testing



Sample size

- Power analysis (particularly for clinical trials)
  - Useful; for simple, expensive experiments
  - Difficult fo complex experiments with many groups
  - Need separate calculations for each character
  - Requires estimate of standard deviation
  - Requires estimate of effect size of scientific importance
- Resource equation (when power analysis not possible)
  - Easy to use even for complex designs with many characters
  - Does not require estimate of standard deviation
  - Crude compared with power analysis

Sample size: Power analysis



Signal
Effect size of scientific interest (delta)

Sample size

Significance level
Chance of a false positive result (Type 1 error).
Often 0.05

Power (probability of detecting the effect)
Often 0.8-0.9.

Sidedness of statistical test (usually 2-sided)

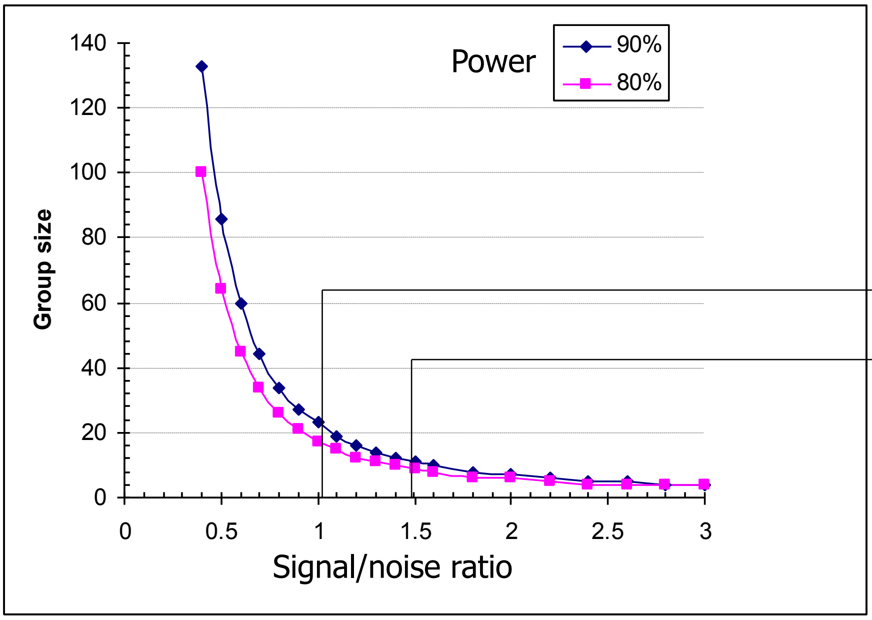
Noise
Variability of the experimental material (sd)

Calculation of sample size using R

```
> power.t.test(sig.level=0.05, power=0.9,delta=6,sd=6)

     Two-sample t test power calculation

              n = 22.0211
          delta = 6
             sd = 6
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

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

(two-sided by default)

Group size and Signal/noise ratio



As a rough guide:

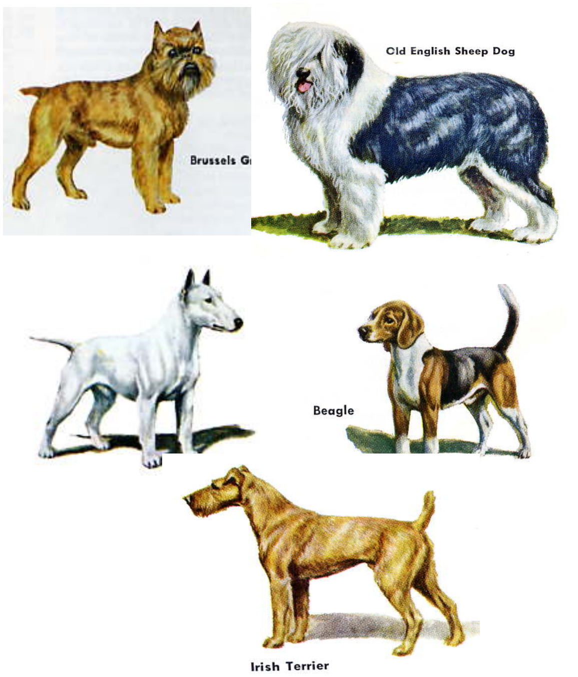
20 ExpUs/group will detect an effect size of one SD

10 ExpUs/group will detect an effect size of 1.5 SDs

Assuming 2-sample, 2 sided t-test and 5% significance level

Comparison of two anaesthetics for dogs under clinical conditions

(Vet. Anaesthes. Analges.)



Unsexed healthy clinic dogs,

- Weight 3.8 to 42.6 kg.
- Systolic BP 141 (SD **36**) mm Hg

Assume:

- **a 20 mmHg** difference between groups is of clinical importance,
- a significance level of α**=0.05**
- a **power=90%**
- a **2-sided t-test**

Signal/Noise ratio 20/36 = 0.56
(standardised effect size)

$\delta = |\mu_1 - \mu_2|/\sigma$

Required sample size 68/group

Dog example: random dogs

Command in the R statistical package. Default is 2-sided

```
power.t.test(delta=0.56, sd=1, power=.9, sig.level=0.05)
```

```
     Two-sample t test power calculation

              n = 67.98649
          delta = 0.56
             sd = 1
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

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

A second paper described:

- Male Beagles weight 17-23 kg
- mean BP 108 (SD **9**) mm Hg.
- Want to detect **20**mm difference between groups (as before)

With the same assumptions as previous slide:

Signal/noise ratio = 20/9 = 2.22

Required sample size 6/group

Summary for two sources of dogs: aim is to be able to detect a 20mmHg change in blood pressure

Type of dog	SDev	Signal/noise	Sample size/gp(1)	%Power (n=8) (2)
Random dogs	36	0.56	68	18
Male beagles	9	2.22	6	98

(1) Sample size: 90% power
(2) Power, Sample size 8/group

Assumes α=5%, 2-sided t-test and effect size 20mmHg

Hexobarbital Sleeping time in mice: inbreds are more uniform and strains differ

Strain	"N"	Mean	SD	Signal/noise	Needed*	Power**
A/N	25	48	4	1.0	23	86
BALB/c	63	41	2	2.0	7	>99
C57BL/HeN	29	33	3	1.3	13	98
C3HB/He	30	22	3	1.3	13	98
SWR/HeN	38	18	4	1.0	23	86
CFW	47	48	12	0.3	191	17
Swiss	47	43	15	0.26	297	13

Mean of SDs: inbreds = 3.2, outbreds = 13.5, p=<0.001

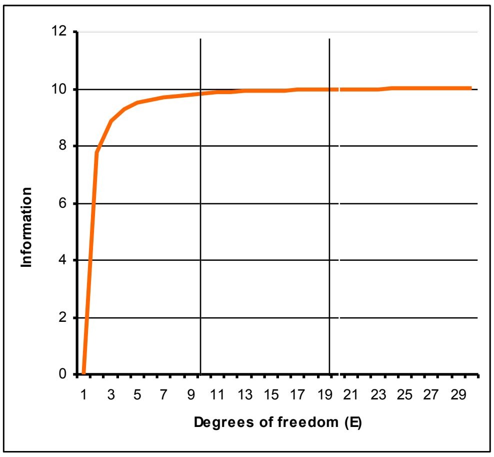
* Power analysis: number needed in a two-sample t-test to detect a 4 min. change in the mean (2-sided) with α=0.05 and a power of 90%
** power of an experiment to detect a 4 min. change in the mean if the sample size is fixed at 20 mice/group

Data from Jay 1955 Proc Soc. Exp Biol Med 90:378

NB. This is based on differences in the SDs. Strains will also differ in sensitivity, as shown in the means, but this can not be predicted

The resource equation

A power analysis is not always possible.
1. If lots of characters
2. No estimate of the standard deviation,
3. Impossible to specify an effect size of scientific importance
4. Complex designs



So use the Resource Equation method.
(Law of diminishing returns)

E= (Total number of experimental units)-(number of treatment groups)

E should be between 10 and 20

The randomised block design: for controlling noise and splitting up the experiment

Treaments A, B & C, within a block subjects are matched

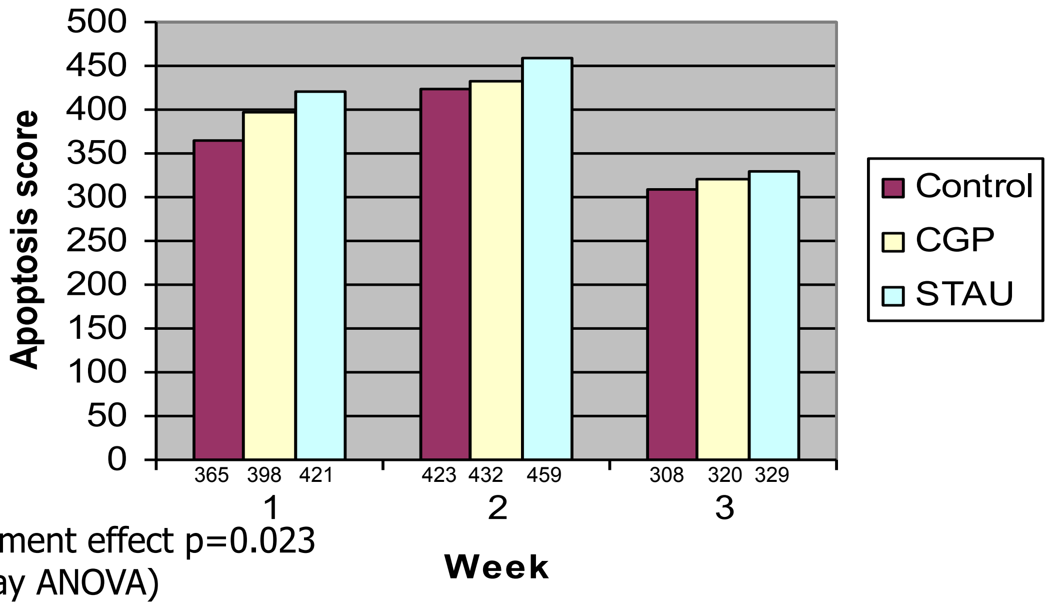
B	C	A	B1
A	C	B	B2
B	A	C	B3
A	C	B	B4
B	C	A	B5

Blocking
- Randomisation is *within-block*
- Multiple differences between blocks

Use when:
- Heterogeneous age/weight
- Different shelves/rooms
- Natural structure (litters)
- Experiment split in time

E= 15-3 = 12

A randomised block experiment



Treatment effect p=0.023
(2-way ANOVA)

Randomised BlocK ANOVA

```
Correct 2-way Analysis of Variance for Apop

Source          DF     SS      MS         F        P
Block            2  21764   10882    114.82    0.000
Treat            2   2129    1064     11.23    0.023
Error            4    379      94
Total            8  24272
```

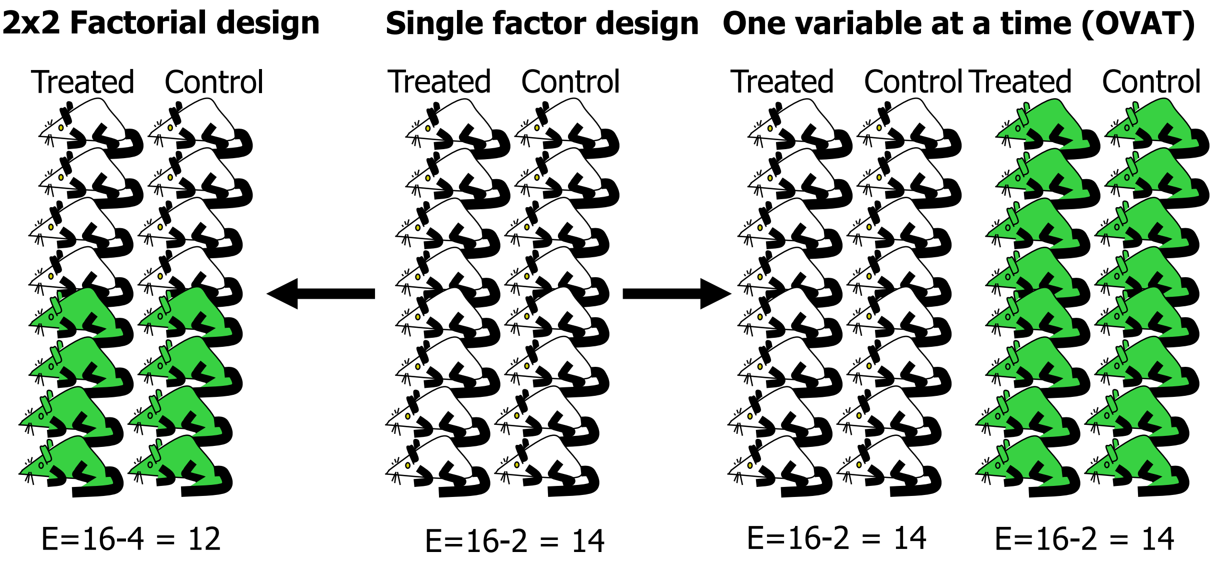
Variance

Post-hoc comparisons required to indicate which means differ.

A well designed experiment

- Absence of bias
  - Experimental unit, randomisation, blinding
- High power
  - Low noise (uniform material, blocking, covariance)
  - High signal (sensitive subjects, high dose)
  - Large sample size
- Wide range of applicability
  - Replicate over other factors (e.g. sex, strain): factorial designs
- Simplicity
- Amenable to a statistical analysis

Factorial designs



2x2 Factorial design

Treated Control

E=16-4 = 12

Single factor design

Treated Control

E=16-2 = 14

One variable at a time (OVAT)

Treated Control Treated Control

E=16-2 = 14 E=16-2 = 14

Factorial designs

(*By using a factorial design*)".... an experimental investigation, at the same time as it is made more comprehensive, may also be made more efficient if by more efficient we mean that more knowledge and a higher degree of precision are obtainable by the same number of observations."

R.A. Fisher, 1960

Effect of chloramphenicol on RBC counts (2000μg/kg)

Want to know:

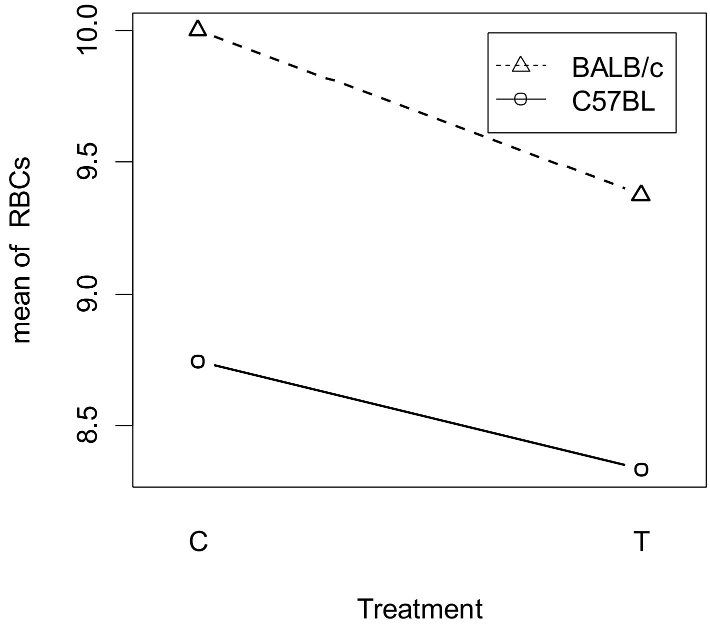
1. Does treatment have an effect on RBC counts
2. Do strains differ in RBC counts
3. Do strains differ in their response (interaction)

Strain	Control	Treated	Strain means
BALB/c	10.10	8.95	
	10.08	8.45	
	9.73	8.68	
	10.09	8.89	9.37
C57BL	9.60	8.82	
	9.56	8.24	
	9.14	8.18	
	9.20	8.10	8.86
Treat. Mean	9.69	8.54	

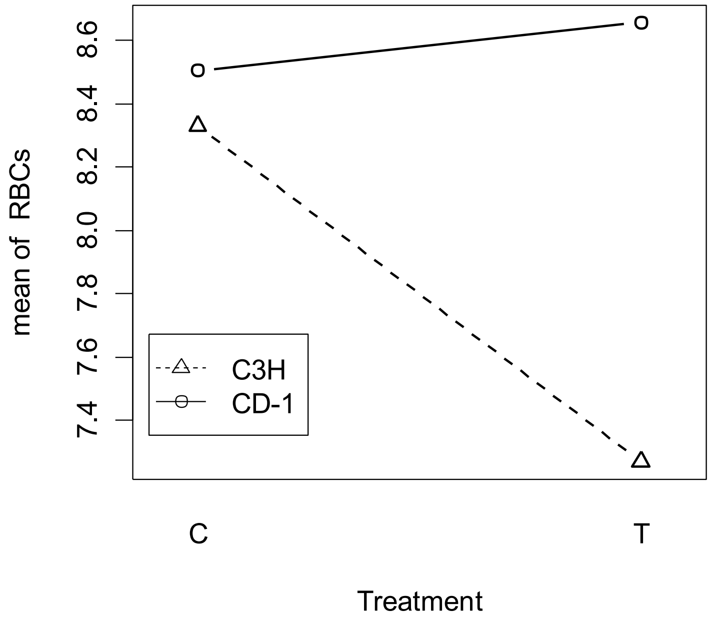
2-way ANOVA with interaction

```
Analysis of Variance Table

Response: RBCs
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Treatment         1 1.0661  1.0661 17.1512  0.001367 **
Strain            1 5.2785  5.2785 84.9232 8.595e-07
***
Treatment:Strain  1 0.0473  0.0473  0.7611  0.400108
Residuals        12 0.7459  0.0622
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
>
```

No interaction



Interaction



Factorial designs

- Very common in biomedical research
- Often incorrectly analysed
- Can have any number of factors and any number of levels of each factor
- 2^n designs can be used to study many factors simultaneously

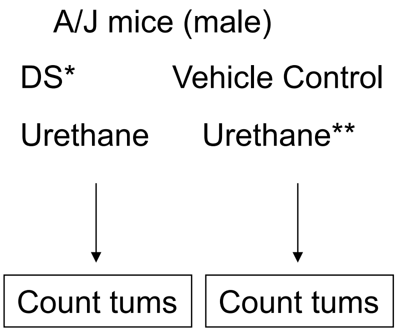
Factorial designs are often incorrectly analysed

Number of studies	513
Factorial designs	153 (30%)
Number analysed correctly	78 (50%)

Niewenhuis et al (2011) Nature Neurosci. 14:1105

A factorial experiment

Could garlic (diallyl sulphide, DS) help to prevent cancer?



A/J mice (male)

DS* Vehicle Control

Urethane Urethane**

Count tums Count tums

*By gavage 0.2mg/g body wt. for 3 days prior to and 3 days following carcinogen treatment
** 1mg/g by i.p. injection

A possible design

A/J male mice	
Vehicle +Urethane	DS + Urethane
10	10

Resource equation
E=20-2 = 18

Power analysis:
A/J mice get about 20 tumors/mouse with a SD of 6 tumors.
10 mice/group should have about an 85% chance of detecting a 1.4 SD decline (8.4 tumors) with a 5% significance level.

Add females: a 2^2 factorial

	Vehicle + Urethane	DS + Urethane
A/J males	5	5
A/J females	5	5

E=20-4 = 16

Plus two carcinogens: a 2^3 factorial

	Vehicle + Urethane	Vehicle + 3MC	DS + Urethane	DS + 3MC
A/J males	3	3	3	3
A/J females	3	3	3	3

E= 24-8= 16

Add another strain: a 2^4 factorial

	Vehicle + Urethane	Vehicle + 3MC	DS + Urethane	DS + 3MC
A/J males	2	2	2	2
A/J females	2	2	2	2
NIH males	2	2	2	2
NIH females	2	2	2	2

E= 32-16= 16. Note each main effect has 16 animals/group

Statistical analysis

Data to be analysed as a 2^4 factorial design using an analysis of variance.

Tumuor counts tend to have a poisson distribution so counts transformed to a square root.

We need to look at assumptions for parametric tests, and for outliers

Data into computer one row per subject

Strain	antiox	inhib	carc	root
1	1	1	1	3.31662
1	1	1	1	3.74166
1	1	1	1	3.74166
1	1	1	2	4.89898
1	1	1	2	4.79583
1	1	1	2	5.74456
1	1	2	1	2.82843
1	1	2	1	3.87298
1	1	2	1	2.64575
1	1	2	2	3.74166
1	1	2	2	4.12311
1	1	2	2	3.74166

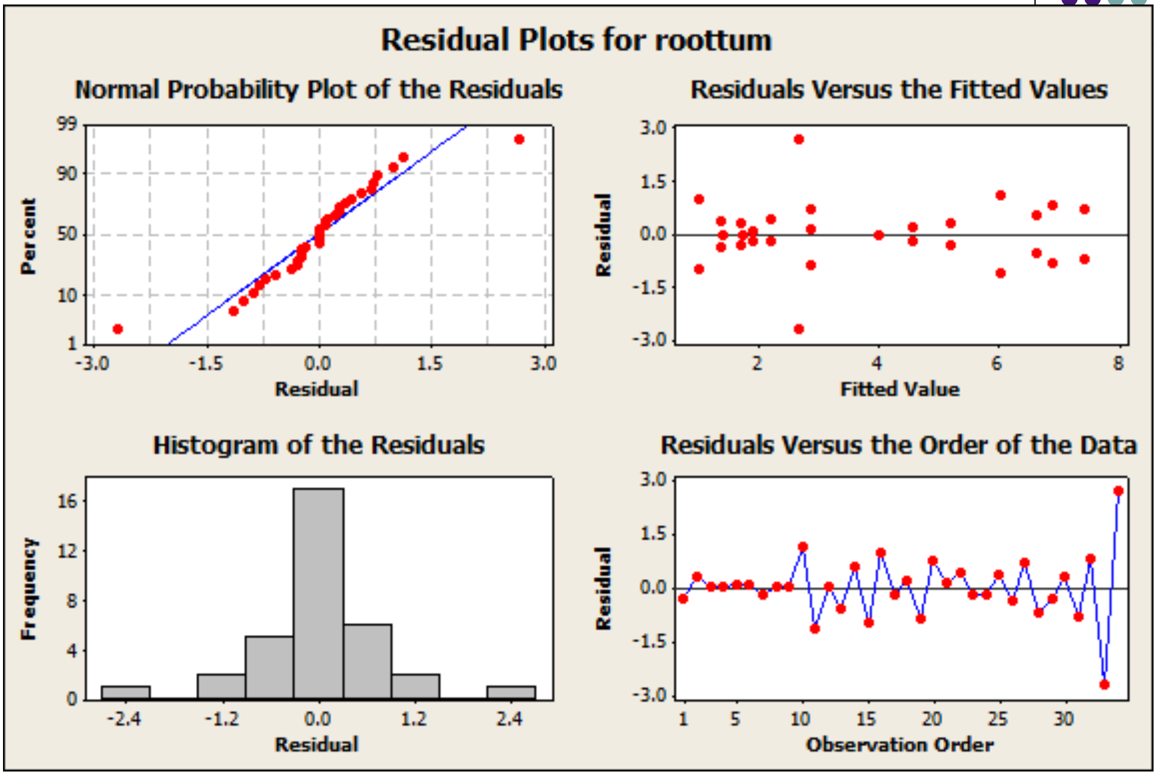
Etc, etc

```
General Linear Model: roottum versus Strains, Sexes, Carcs, Antioxs

Factor   Type   Levels  Values
Strains  fixed       2  A, N
Sexes    fixed       2  F, M
Carcs    fixed       2  3MC, Urethane
Antioxs  fixed       2  No, Yes


Analysis of Variance for roottum, using Adjusted SS for Tests

Source                       DF   Seq SS   Adj SS   Adj MS      F       P
Strains                       1   56.401   72.457   72.457  53.73  0.000***
Sexes                         1    0.436    0.411    0.411   0.30  0.588
Carcs                         1   19.122   12.304   12.304   9.12  0.007***
Antioxs                       1   17.559    9.562    9.562   7.09  0.016*
Strains*Sexes                 1    0.379    0.088    0.088   0.06  0.802
Strains*Carcs                 1   33.965   32.912   32.912  24.41  0.000***
Strains*Antioxs               1    0.794    0.689    0.689   0.51  0.484
Sexes*Carcs                   1    3.065    3.672    3.672   2.72  0.116     X
Sexes*Antioxs                 1    0.640    0.461    0.461   0.34  0.566
Carcs*Antioxs                 1   13.271   12.685   12.685   9.41  0.007***
Strains*Sexes*Carcs           1    0.480    0.299    0.299   0.22  0.644
Strains*Sexes*Antioxs         1    0.554    0.726    0.726   0.54  0.472
Strains*Carcs*Antioxs         1    0.212    0.242    0.242   0.18  0.677
Sexes*Carcs*Antioxs           1    0.350    0.260    0.260   0.19  0.666
Strains*Sexes*Carcs*Antioxs   1    0.918    0.918    0.918   0.68  0.420
Error                        18   24.273   24.273    1.349
Total                        33  172.418

X becomes significant if outliers removed

S = 1.16126   R-Sq = 85.92%   R-Sq(adj) = 74.19%
```

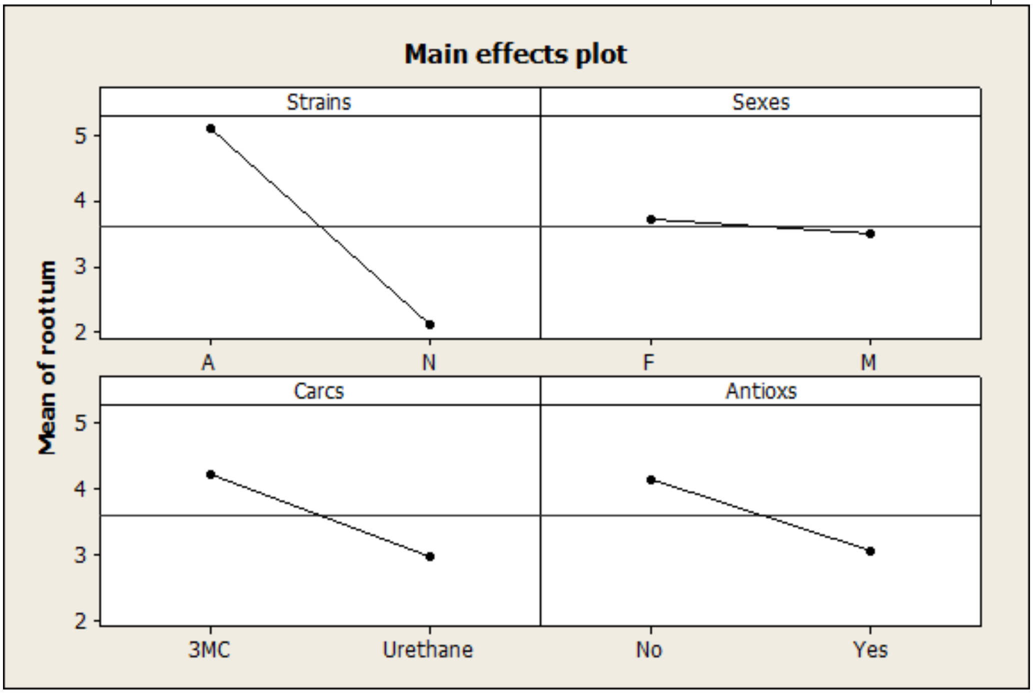
“Main effects”

Two-way interactions

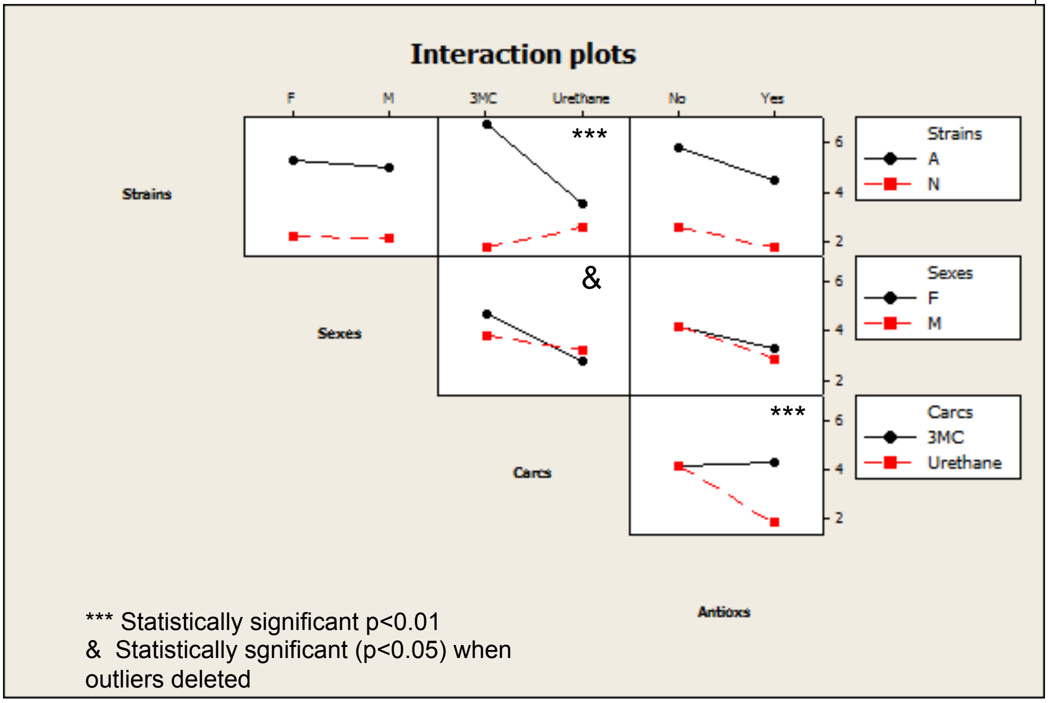
Higher interactions



Main effects



Two-way interactions



*** Statistically significant p<0.01
& Statistically sgnificant (p<0.05) when outliers deleted

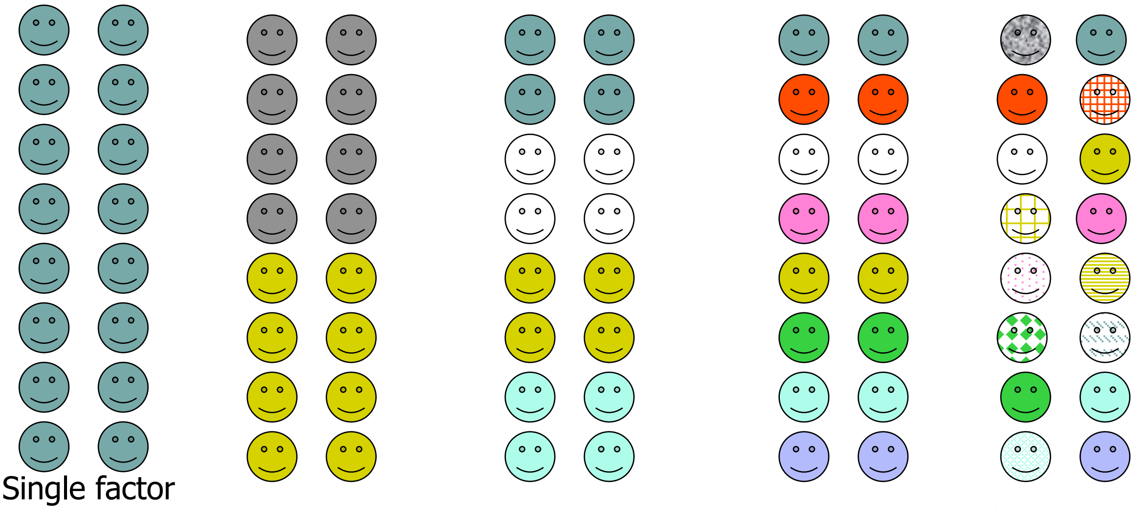
Conclusions: A factorial exploratory experiment

- As four separate experiments
  - 4 x 20 =80 mice
  - Each comparison 10 versus 10
  - No estimates of interactions
- As a 2^4 factorial
  - 32 mice used
  - Each main effect is 16 vs 16 mice
  - All interactions estimated
  - Some interactions important

Conclusions

- Well designed experiments save time, money and animals and improve the quality of the research
- To avoid bias
  - Identify correctly the experimental unit.
  - Assign these to treatments at random and measurements should be done in random order
  - Where possible investigators should be "blinded" using coded samples
- To maximise power (chance of detecting an effect)
  - Controlling all possible sources of variation, including genetic variation (using inbred strains)
  - Use randomised block designs to control time and space variation and split experiments into more manageable parts
  - Choose sensitive subjects (use factorial designs to find them)
  - Use an objective method of determining sample size (power analysis or resource equation)
- Explore the range of applicability using factorial designs (more information per experimental unit)
- "Gold Standard" and "ARRIVE" guidelines. Provide a checklist of what should be in your manuscript.

Factorial designs & group size



8

Trt. Ctrl.

Single factor
Inbred strain
E=14

8 or 4?

Trt. Ctrl.

2x2 Factorial
E=12

8 or 2?

Trt. Ctrl.

2x4 Factorial
E=8

8 or 1?

Trt. Ctrl.

Randomised block
E=7, special case

8 or ??

Trt. Ctrl.

Outbred
E=?