Basic experiment design

C.O.S. Sorzano
coss@cnb.csic.es

National Center of Biotechnology (CSIC)

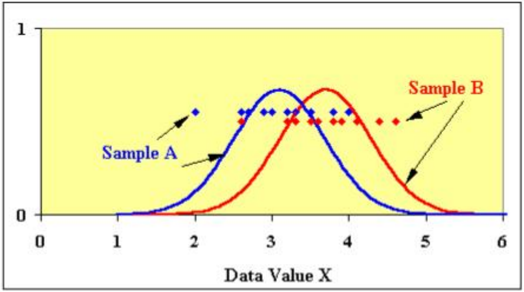
October 15, 2022

Outline

1. Basic designs
   - Completely Randomized Design (CRD)
   - Randomized Complete Block Design (RCBD)
   - Factorial design (FD)
   - Regression design (RD)
   - Conclusions

Objective

The objective today is to learn to use a tool much more powerful than the two independent groups Student's t-test (control and treatment).



Outline

1. Basic designs
    - Completely Randomized Design (CRD)
    - Randomized Complete Block Design (RCBD)
    - Factorial design (FD)
    - Regression design (RD)
    - Conclusions

Completely Randomized Design

Example 0

We are testing a new drug (X 325mg) for blood pressure versus a placebo on 1000 people. We divide the group of people in two equal groups of 500 people. Each person will be randomly assigned to the treatment or the placebo.

y_{11}	y_{21}
y_{12}	y_{22}
...	...
$y_{1,500}$	$y_{2,500}$

- $y_{1\cdot}, y_{2\cdot}$: Means of each one of the groups
- $y_{\cdot\cdot}$: Overall mean

Completely Randomized Design

The data (blood pressure) is supposed to be generated as

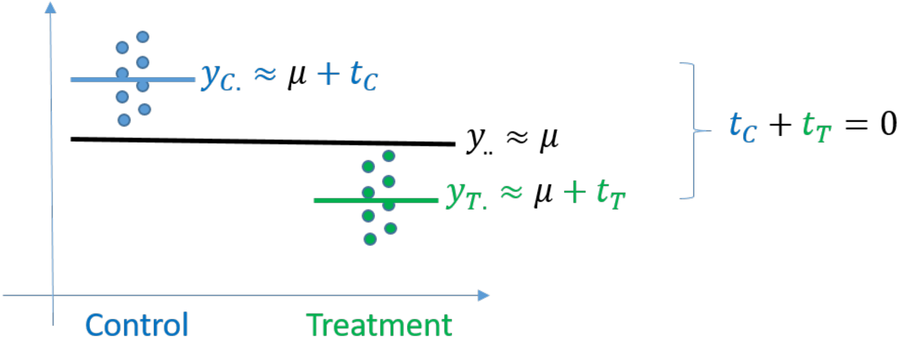
$$y_{jk} = \mu + t_j + \epsilon_{jk}$$

- μ is the average blood pressure of the whole population.
- t_1 and t_2 are the effects of the drug (t_1) and the placebo (t_2). It must be

$$\sum_j t_j = 0$$

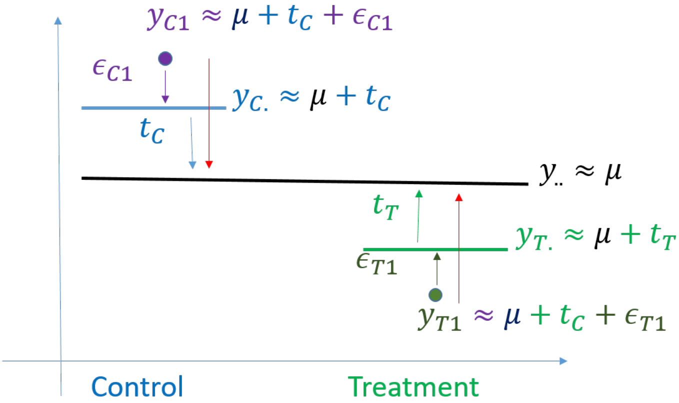
- y_{jk} is the measurement observed for the k-th individual who has been given treatment j.
- ϵ_{jk} is the part of the observed measurement that cannot be explained by the average and the treatment.

Completely Randomized Design



Completely Randomized Design

$$\underbrace{\sum_j \left(y_{C_j} - y_{..}\right)^2 + \sum_j \left(y_{Tj} - y_{..}\right)^2}_{\text{Total variation}} = \underbrace{\sum_j (t_C)^2 + \sum_j (t_T)^2}_{\text{Control/Treatment}} + \underbrace{\sum_j \left(\epsilon_{C_j}\right)^2 + \sum_j \left(\epsilon_{Tj}\right)^2}_{\text{Noise}}$$


$y_{C1} \approx \mu + t_C + \epsilon_{C1}$
ϵ_{C1}
$y_{C.} \approx \mu + t_C$
t_C
$y_{..} \approx \mu$
t_T
$y_{T.} \approx \mu + t_T$
ϵ_{T1}
$y_{T1} \approx \mu + t_C + \epsilon_{T1}$
Control
Treatment

Completely Randomized Design

Normally this is presented in a table

Source	Sum of Squares (SS)	Degrees of freedom (df)	Mean squares (MS=SS/df)
Treatments	$SS_T = \sum_{jk}(y_{j.} - y_{..})^2$	$t-1$	$MS_T = \frac{SS_T}{df_t}$
Residuals	$SS_\epsilon = \sum_{jk}(y_{jk} - y_{j.})^2$	$\sum_j(n_j - 1) = n - t$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	$SS = \sum_{jk}(y_{jk} - y_{..})^2$	$n-1$	

If the residuals are normally distributed, then the Linear Model checks whether the treatments have a significant contribution explaining the variance through a F-Snedecor statistic with $t-1$ and $\sum_j(n_j - 1)$ degrees of freedom.

$$F = \frac{MS_T}{MS_\epsilon}$$

Completely Randomized Design

Example 1

Let us assume that the table in our case is

Source	SS	df	MS=SS/df
Treatments	256.88	1	256.88
Residuals	13600.28	998	13.61
Total	13857.16	999	

Note

$$\begin{aligned} 13857.16 &= 256.88 + 13600.28 \\ 999 &= 1 + 998 \end{aligned}$$

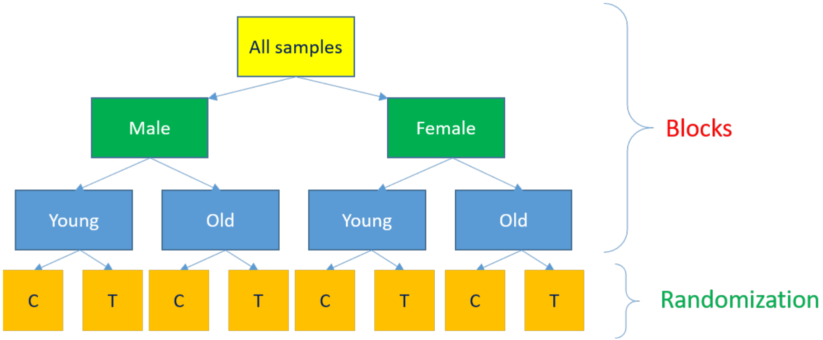
In this case

$$F = \frac{256.88}{13.61} = 18.87 \gg 3.85 = F_{0.95,1,998}$$

Outline

1. Basic designs
   - Completely Randomized Design (CRD)
   - Randomized Complete Block Design (RCBD)
   - Factorial design (FD)
   - Regression design (RD)
   - Conclusions

Randomized Complete Block Design



Randomized Complete Block Design

The data (blood pressure) is supposed to be generated as

$$y_{ijk} = \mu + b_i + t_j + \epsilon_{ijk}$$

- μ is the average blood pressure of the whole population.
- b_1 and b_2 are the differences in blood pressure between men (b_1) and women (b_2), the blocks. It must be
$$\sum_i b_i = 0$$
- t_1 and t_2 are the effects of the drug (t_1) and the placebo (t_2). It must be
$$\sum_j t_j = 0$$
- y_{ijk} is the measurement observed for the k-th individual of the i-th block who has been given treatment j.
- ϵ_{ijk} is the part of the observed measurement that cannot be explained by the average, block and treatment.

Randomized Complete Block Design

The table of the linear model becomes

Source	SS	df	MS=SS/df
Blocks	SS_B	$b-1$	$MS_B = \frac{SS_B}{df_B}$
Treatments	SS_T	$t-1$	$MS_T = \frac{SS_T}{df_T}$
Residuals	SS_ϵ	$n-b-t+1$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	SS	$n-1$	

If the residuals are Gaussian, we may test whether the contribution of the blocks or treatments are significant through the same F-Snedecor as before (pay attention to use the corresponding degrees of freedom).

Randomized Complete Block Design

Example 2

Let us assume that in our case it becomes

Source	SS	df	MS=SS/df
Blocks	1500.04	1	1500.04
Treatments	256.88	1	256.88
Residuals	12100.24	997	12.13
Total	13857.16	999	

Note

$$\begin{aligned} 13857.16 &= 1500.04 + 256.88 + 12100.24 \\ 999 &= 1 + 1 + 997 \end{aligned}$$

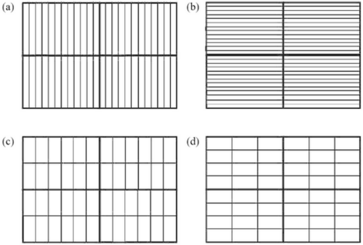
In this case

$$F = \frac{256.88}{12.13} = 21.17 \gg 3.85 = F_{0.95,1,997}$$

Randomized Complete Block Design

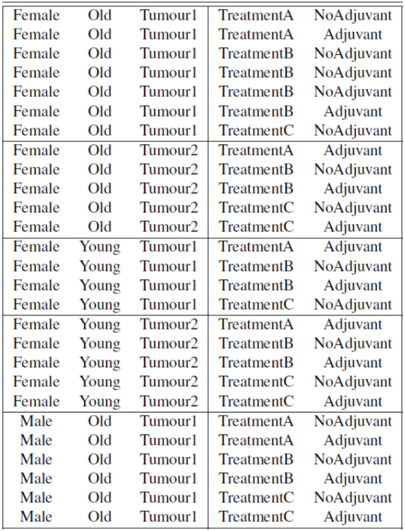
- If there are clear variables to block, they should be blocked. Litters are normally chosen as blocks (and birth weight as covariate), age, sex, researcher, week of the experiment, ...

- If there are no obvious blocking variables, but we may create blocks, we may do as an "insurance" against possible patterns not yet identified.



(e.g. 4 block, 12 treatments)

Randomized (In)Complete Block Design



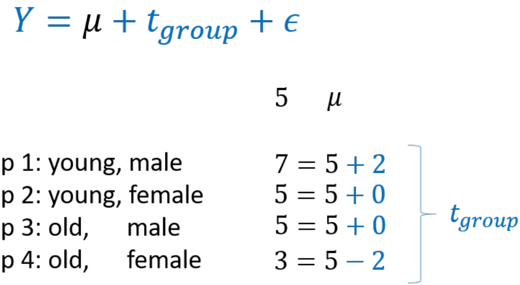
Blocks			Treatments	
Female	Old	Tumour1	TreatmentA	NoAdjuvant
Female	Old	Tumour1	TreatmentA	Adjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	Adjuvant
Female	Old	Tumour1	TreatmentC	NoAdjuvant
Female	Old	Tumour2	TreatmentA	Adjuvant
Female	Old	Tumour2	TreatmentB	NoAdjuvant
Female	Old	Tumour2	TreatmentB	Adjuvant
Female	Old	Tumour2	TreatmentC	NoAdjuvant
Female	Old	Tumour2	TreatmentC	Adjuvant
Female	Young	Tumour1	TreatmentA	Adjuvant
Female	Young	Tumour1	TreatmentB	NoAdjuvant
Female	Young	Tumour1	TreatmentB	Adjuvant
Female	Young	Tumour1	TreatmentC	NoAdjuvant
Female	Young	Tumour2	TreatmentA	Adjuvant
Female	Young	Tumour2	TreatmentB	NoAdjuvant
Female	Young	Tumour2	TreatmentB	Adjuvant
Female	Young	Tumour2	TreatmentC	NoAdjuvant
Female	Young	Tumour2	TreatmentC	Adjuvant
Male	Old	Tumour1	TreatmentA	NoAdjuvant
Male	Old	Tumour1	TreatmentA	Adjuvant
Male	Old	Tumour1	TreatmentB	NoAdjuvant
Male	Old	Tumour1	TreatmentB	Adjuvant
Male	Old	Tumour1	TreatmentC	NoAdjuvant
Male	Old	Tumour1	TreatmentC	Adjuvant

Outline

1. Basic designs
   - Completely Randomized Design (CRD)
   - Randomized Complete Block Design (RCBD)
   - Factorial design (FD)
   - Regression design (RD)
   - Conclusions

Factorial Design

We are measuring the effect of a treatment on a number of animals as a function of age and sex. These are called factors, and their different values are called levels. For each combination we have $N = 6$ animals. The numbers below show the average of each one of the groups.



$$Y = \mu + t_{group} + \epsilon$$

$$\begin{array}{lll} \text{All:} & 5 & \mu \end{array}$$

$$\left.\begin{array}{ll} \text{Group 1: young, male} & 7 = 5 + 2 \\ \text{Group 2: young, female} & 5 = 5 + 0 \\ \text{Group 3: old, male} & 5 = 5 + 0 \\ \text{Group 4: old, female} & 3 = 5 - 2 \end{array}\right\} t_{group}$$

Factorial Design

However, we could have analyzed the data differently gaining more insight into the influence of each factor. This kind of analysis is called main effects.

$$Y = \mu + t_{group} + \epsilon$$

All: $5 \quad \mu$

$$\left.\begin{array}{ll} \text{Group 1: young, male} & 7 = 5 + 2 \\ \text{Group 2: young, female} & 5 = 5 + 0 \\ \text{Group 3: old, male} & 5 = 5 + 0 \\ \text{Group 4: old, female} & 3 = 5 - 2 \end{array}\right\} t_{group}$$

$$Y = \mu + t_{age} + t_{sex} + \epsilon$$

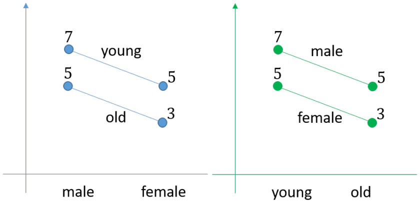
	male	female	
young	$7 = 5 + 1 + 1$	$5 = 5 + 1 - 1$	$t_{young} = 1$
old	$5 = 5 - 1 + 1$	$3 = 5 - 1 - 1$	$t_{old} = -1$
	$t_{male} = 1$	$t_{female} = -1$	$\mu = 5$

Factorial Design

We may arrange the response graphically. Note the fact that the two lines are parallel.

$$Y = \mu + t_{age} + t_{sex} + \epsilon$$

	male	female	
young	$7 = 5 + 1 + 1$	$5 = 5 + 1 - 1$	$6 = 5 + 1$
old	$5 = 5 - 1 + 1$	$3 = 5 - 1 - 1$	$4 = 5 - 1$
	$6 = 5 + 1$	$4 = 5 - 1$	$\mu = 5$

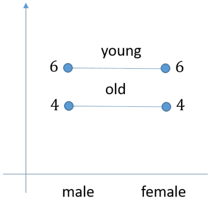


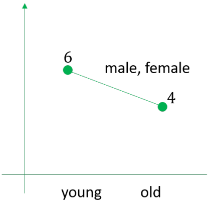
Factorial Design

In the following example, only one of the factors has an effect. The lines are still parallel or coincident.

$$Y = \mu + t_{age} + t_{sex} + \epsilon$$

	male	female	
young	$6 = 5 + 1 + 0$	$6 = 5 + 1 + 0$	$6 = 5 + 1$
old	$4 = 5 - 1 + 0$	$4 = 5 - 1 + 0$	$4 = 5 - 1$
	$5 = 5 + 0$	$5 = 5 + 0$	$\mu = 5$



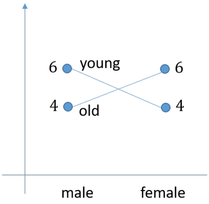


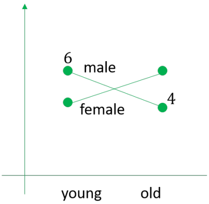
Factorial Design

Main effects alone are not able to explain the data. Lines are not parallel anymore.

$$Y = \mu + t_{age} + t_{sex} + \epsilon$$

	male	female	
young	$6 \neq 5 + 0 + 0$	$4 \neq 5 + 0 + 0$	$6 = 5 + 0$
old	$4 \neq 5 + 0 + 0$	$6 \neq 5 + 0 + 0$	$4 = 5 + 0$
	$5 = 5 + 0$	$5 = 5 + 0$	$\mu = 5$



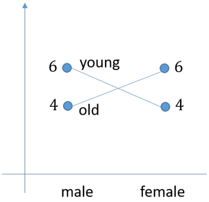


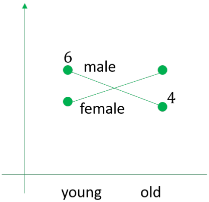
Factorial Design

We need to add interactions to be able to explain the data. **Interaction effects** exist when differences on one factor depend on the level you are on another factor. The interactions are between factors and not between levels.

$$Y = \mu + t_{age} + t_{sex} + t_{age.sex} + \epsilon$$

	male	female	
young	$6 = 5 + 0 + 0 + 1$	$4 = 5 + 0 + 0 - 1$	$6 = 5 + 0$
old	$4 = 5 + 0 + 0 - 1$	$6 = 5 + 0 + 0 + 1$	$4 = 5 + 0$
	$5 = 5 + 0$	$5 = 5 + 0$	$\mu = 5$





Factorial Design

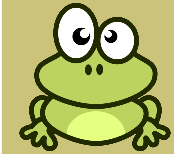
The analysis table may be represented as

Source	SS	df	MS=SS/df
P main effects	SS_P	$p-1$	$MS_P = \frac{SS_P}{df_P}$
Q main effects	SS_Q	$q-1$	$MS_Q = \frac{SS_Q}{df_Q}$
PQ interactions	SS_{PQ}	$(p-1)(q-1)$	$MS_{PQ} = \frac{SS_{PQ}}{df_{PQ}}$
Residuals	SS_ϵ	$n-pq$	$MS_\epsilon = \frac{SS_\epsilon}{df_\epsilon}$
Total	SS	$n-1$	

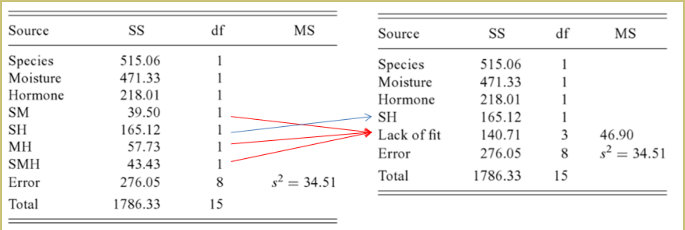
Factorial Design

Example 3

We are testing water uptake by amphibia. Frogs and toads (species factor S) are kept in most or dry conditions before the experiment (moisture factor M) and half of the animals are injected with a mammalian water balance hormone (hormone factor H). A full factorial experiment is performed with 2 animals per treatment combination (cell).



$$y_{ijkl} = \mu + s_i + m_j + h_k + (sm)_{ij} + (sh)_{ik} + (mh)_{jk} + \epsilon_{ijkl}$$



Source	SS	df	MS
Species	515.06	1	
Moisture	471.33	1	
Hormone	218.01	1	
SM	39.50	1	
SH	165.12	1	
MH	57.73	1	
SMH	43.43	1	
Error	276.05	8	$s^2 = 34.51$
Total	1786.33	15	

Source	SS	df	MS
Species	515.06	1	
Moisture	471.33	1	
Hormone	218.01	1	
SH	165.12	1	
Lack of fit	140.71	3	46.90
Error	276.05	8	$s^2 = 34.51$
Total	1786.33	15	

Factorial Design

Example 4

Assume that we have resources for 24 observations and we assume that there is no interaction between factors

$$y_{ijkl} = \mu + s_i + m_j + h_k + \epsilon_{ijkl}$$

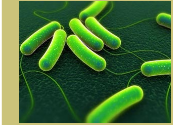
Three different experiment designs are considered:

1. One variable changes at a time
   - (Frogs,Dry,NoHormone) vs (Toad,Dry,NoHormone): 4 animals each
   - (Frogs,Dry,NoHormone) vs (Frogs,Wet,NoHormone): 4 animals each
   - (Frogs,Dry,NoHormone) vs (Frogs,Dry,Hormone): 4 animals each
2. Do not repeat (Frogs,Dry,NoHormone) in each comparison:
   - (Frogs,Dry,NoHormone): 6 animals
   - (Toads,Dry,NoHormone): 6 animals
   - (Frogs,Wet,NoHormone): 6 animals
   - (Frogs,Dry,Hormone): 6 animals
3. Factorial design (all possible combinations) with 3 animals each.

Factorial designs and single replicates

High-order interactions can be assimilated to the error, and single replicate factorial designs may be conceived.

Example 5



We are interested in the survival of *Salmonella typhimurium* under 3 experimental factors: 3 levels of sorbic acid (=Factor S), 6 levels of water activity (=Factor A), and 3 levels of pH (=Factor P). The data will be the log (density/ml) measured after 7 days after treatment started.

We have $3 \times 6 \times 3 = 54$ treatments, and we will use a single replicate for each treatment.

Factorial designs and single replicates

Example 5(continued)

The data analysis table would be

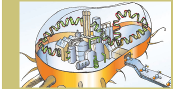
	SS	df	MS	F
Water activity (A)	81.57	5=(6-1)	16.31	$473 > F_{0.95,5,20}$
Sorbic acid (S)	2.76	2=(3-1)	1.38	$40 > F_{0.95,5,20}$
pH (P)	0.01	2=(3-1)	0.01	$0.2 < F_{0.95,2,20}$
AS	1.32	10=(6-1)(3-1)	0.13	$3.8 > F_{0.95,10,20}$
AP	0.45	10=(6-1)(3-1)	0.04	$1.3 < F_{0.95,10,20}$
SP	0.23	4=(3-1)(3-1)	0.06	$1.7 < F_{0.95,4,20}$
$ASP \approx$ **Error**	0.69	20=(6-1)(3-1)(3-1)	0.03	
Total	87.03	53		

The problem with single replicate, factorial designs is that 1) it is difficult to use blocking, 2) due to the lack of replication, there is no possibility to construct an unbiased estimate of the noise.

Fractional Factorial Design

Example 6

We are interested in a cell line as biologics bioreactor, and we want to optimize production. We have identified 7 variables we may control (temperature, humidity, pH, O_2 concentration, CO_2 concentration, glucose concentration, aminoacid concentration). For each variable we have 2 possible values. There are $2^7 = 128$ possible treatments, but we can only afford 64. We do not foresee 3rd order interactions or higher. Can we perform this experiment?



The number of degrees of freedom needed to identify main effects and 2nd order interactions is

	df
Main effects	7
2nd Order Interactions	21=C(7,2)=7!/(2!5!)

So we need 28 samples plus sufficient replication for estimating the error. For instance, if we perform 64 experiments, there would be 37 df for the noise.

Factorial Design

Advantages of factorial design:

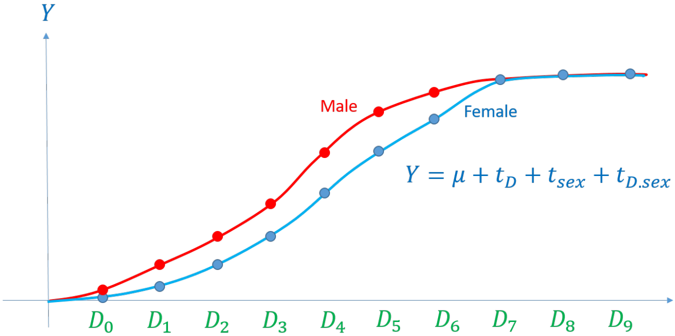
- Interactions between factors can be estimated and their significance tested.
- Wider validity of main effects: they have been tested in many different cases (e.g. the effect of moisture have been tested with frogs and toads, and with and without hormone)
- Several experiments are done simultaneously: the variance of pairwise comparisons is minimal, as shown in the following experiment

Factorial design: Hold all factors constant except ~~the one~~ those whose effects we are investigating.

Outline

1. Basic designs
   - Completely Randomized Design (CRD)
   - Randomized Complete Block Design (RCBD)
   - Factorial design (FD)
   - Regression design (RD)
   - Conclusions

Regression Design

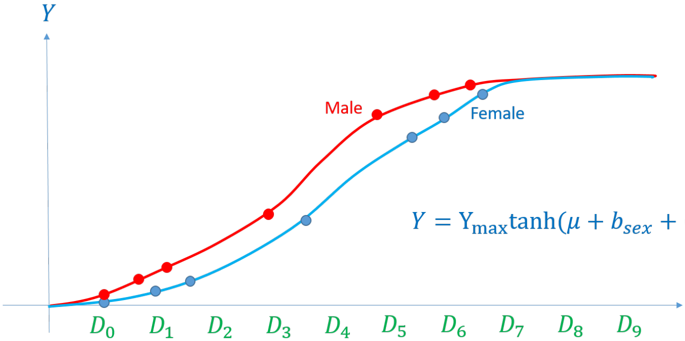


$$Y = \mu + t_D + t_{sex} + t_{D.sex}$$

$df = 19$

Doses can be analyzed as a 2-way ANOVA, although we will need more samples.

Regression Design



$$Y = \mathrm{Y}_{\mathrm{max}}\tanh(\mu + b_{sex} + (\beta + \beta_{sex} d)) \qquad df = 5$$

Doses can be analyzed as a regression, with fewer samples and located in different positions.

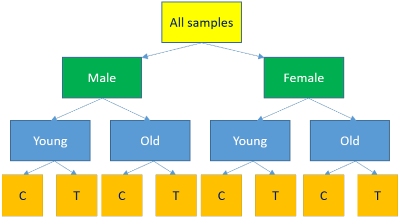
Outline

1. Basic designs
   - Completely Randomized Design (CRD)
   - Randomized Complete Block Design (RCBD)
   - Factorial design (FD)
   - Regression design (RD)
   - Conclusions

Conclusions

- The goal of our research is to show that there is a difference with respect to some factor.
- To be statistically significant this difference must be above the level of noise.
- Experimental design helps controlling the sources of variability.
- Always randomize at the level of blocks.
- Control what you can, block what you cannot, and randomize the rest.

Conclusions

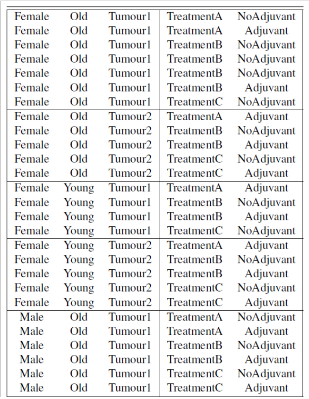


Blocks

+

Factors (Randomization)

Experiment Design



Blocks			Treatments	
Female	Old	Tumour1	TreatmentA	NoAdjuvant
Female	Old	Tumour1	TreatmentA	Adjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	NoAdjuvant
Female	Old	Tumour1	TreatmentB	Adjuvant
Female	Old	Tumour1	TreatmentC	NoAdjuvant
Female	Old	Tumour2	TreatmentA	Adjuvant
Female	Old	Tumour2	TreatmentB	NoAdjuvant
Female	Old	Tumour2	TreatmentB	Adjuvant
Female	Old	Tumour2	TreatmentC	NoAdjuvant
Female	Old	Tumour2	TreatmentC	Adjuvant
Female	Young	Tumour1	TreatmentA	Adjuvant
Female	Young	Tumour1	TreatmentB	NoAdjuvant
Female	Young	Tumour1	TreatmentB	Adjuvant
Female	Young	Tumour1	TreatmentC	NoAdjuvant
Female	Young	Tumour2	TreatmentA	Adjuvant
Female	Young	Tumour2	TreatmentB	NoAdjuvant
Female	Young	Tumour2	TreatmentB	Adjuvant
Female	Young	Tumour2	TreatmentC	NoAdjuvant
Female	Young	Tumour2	TreatmentC	Adjuvant
Male	Old	Tumour1	TreatmentA	NoAdjuvant
Male	Old	Tumour1	TreatmentA	Adjuvant
Male	Old	Tumour1	TreatmentB	NoAdjuvant
Male	Old	Tumour1	TreatmentB	Adjuvant
Male	Old	Tumour1	TreatmentC	NoAdjuvant
Male	Old	Tumour1	TreatmentC	Adjuvant

Outline

1. Basic designs
   - Completely Randomized Design (CRD)
   - Randomized Complete Block Design (RCBD)
   - Factorial design (FD)
   - Regression design (RD)
   - Conclusions