Regression, Anova , Experimental Design

Joseph Abraham

Lecture V RBP5793

Outline of Lecture

- Linear Regression
- Anova
- Basic Experimental Design

Statistical Reasoning I: Prof. John Mcready

Johns Hopkins School of Public Health Open Course Ware

Linear Regression I

A huge subject (may books and papers written). For us a preliminary step to understanding Anova.

Basic idea: we a set of **paired quantities** (x,y).

For each pair, the value of x is assumed to partially determine to value of y. Why partially ?

Linear Regression II

Imagine $y = 1,4 + 2\,x$ (exactly). Then we expect

x	y
2	5,4
1,88	5,16
2,12	5,64
1,39	4,18

Given x we have perfect knowledge of y.

Linear Regression III

Supose we see instead

x	y
2	4,4
1,88	3,78
2,12	5,44
1,39	4,02

Given x do we have perfect knowledge of y ?

Linear Regression III

Supose we see instead

x	y
2	4,4
1,88	3,78
2,12	5,44
1,39	4,02

Given x do we have perfect knowledge of y ?

Given x do we have some knowledge of y ?

Linear Regression III

Supose we see instead

x	y
2	4,4
1,88	3,78
2,12	5,44
1,39	4,02

Given x do we have perfect knowledge of y ?

Given x do we have some knowledge of y ?

For $x = 1,65$ $y = 4,91$? ,

Linear Regression III

Supose we see instead

x	y
2	4,4
1,88	3,78
2,12	5,44
1,39	4,02

Given x do we have perfect knowledge of y ?

Given x do we have some knowledge of y ?

For $x = 1,65$ $y = 4,91$? , $y = 4,68$

Linear Regression III

Supose we see instead

x	y
2	4,4
1,88	3,78
2,12	5,44
1,39	4,02

Given x do we have perfect knowledge of y ?

Given x do we have some knowledge of y ?

For $x = 1,65$ $y = 4,91$? , $y = 4,68$ y= 3,81 ?

We say that $y = 1,4 + 2\ x + error$

(some knowledge, not perfect)

Linear Regression IV

In this case we know $y = 1,4 + 2\,x$ works quite well.

In generally we have only pairs of values $(x_1, y_1), (x_2, y_2), \ldots$

We imagine that in reality $y = \alpha + \beta\,x + \textit{error}$

and we want to **determine** α and β.

Why is this a **statistical** problem ?

Linear Regression IV

In this case we know $y = 1,4 + 2\,x$ works quite well.

In generally we have only pairs of values $(x_1, y_1), (x_2, y_2), \ldots$

We imagine that in reality $y = \alpha + \beta\,x + \textit{error}$

and we want to **determine** α and β.

Why is this a **statistical** problem ?

Among α and β usually β is more important.

β tells us how y is affected by x. How to interpret α ?

Linear Regression V

α is a prediction for y when $x = 0$.

If β is zero we expect y independent of x

If β is different from zero then y varies with x

A large β value indicates a **small** change in x

leads to a **large** change in y. A small β value

indicates a **small** change in x leads to a **small** change in y

All this is true if the error is moderate. For large error

dependence of y on x is less even with larger β.

Final objective is a **straight line** fit between x and y.

Linear Regression VI

$y = \alpha + \beta x + error$

Why is this a **statistical** problem ?

Linear Regression VI

$y = \alpha + \beta x + error$

Why is this a **statistical** problem ?

The objective is to test if x and y are related.

For this objective, what is H_0 ?

Linear Regression VI

$y = \alpha + \beta x + error$

Why is this a **statistical** problem ?

The objective is to test if x and y are related.

For this objective, what is H_0 ?

error is for all the unknowns which also affect y

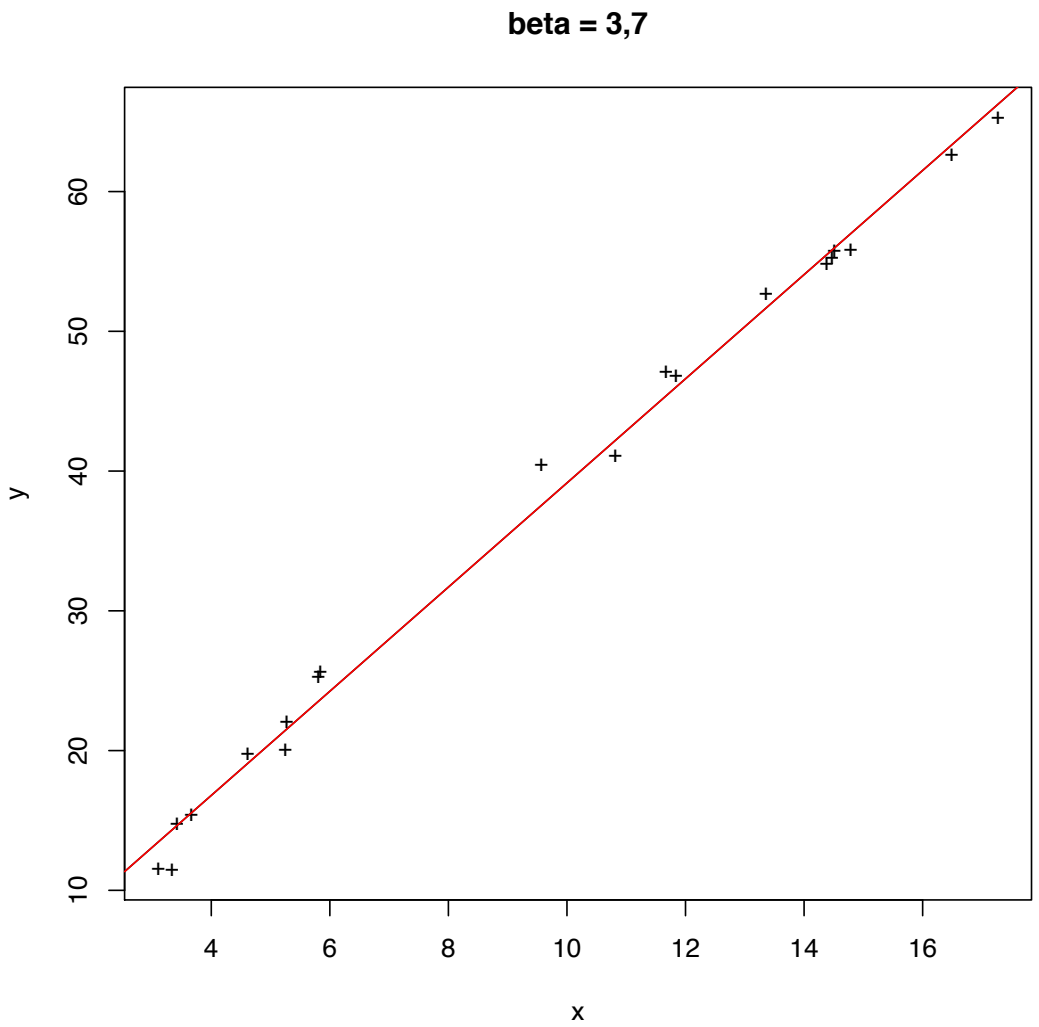
Assume a normal distribution for error ($\mu = 0$).

Variance determines the importance of the unknowns

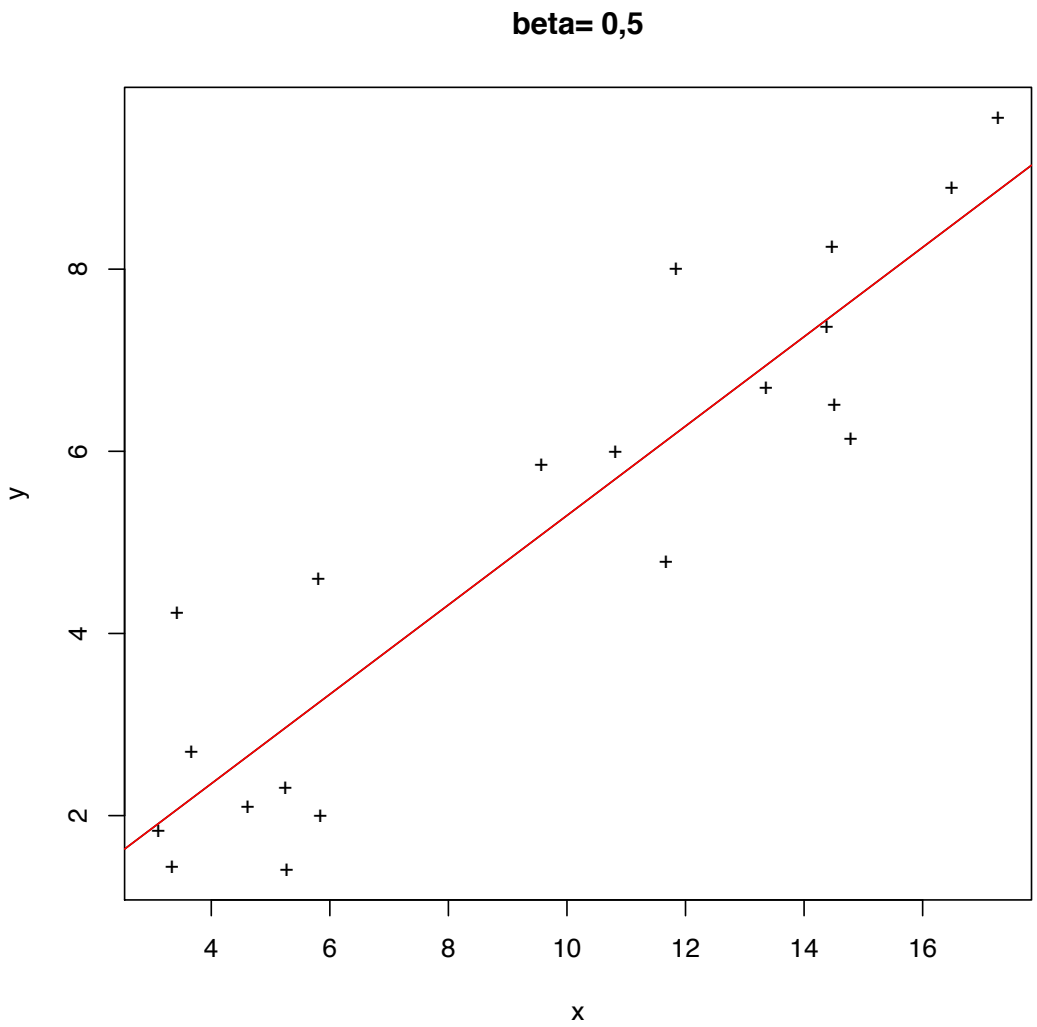
and other limitations of model (*eg.* $\beta x + \gamma x^2$)

and not just β.

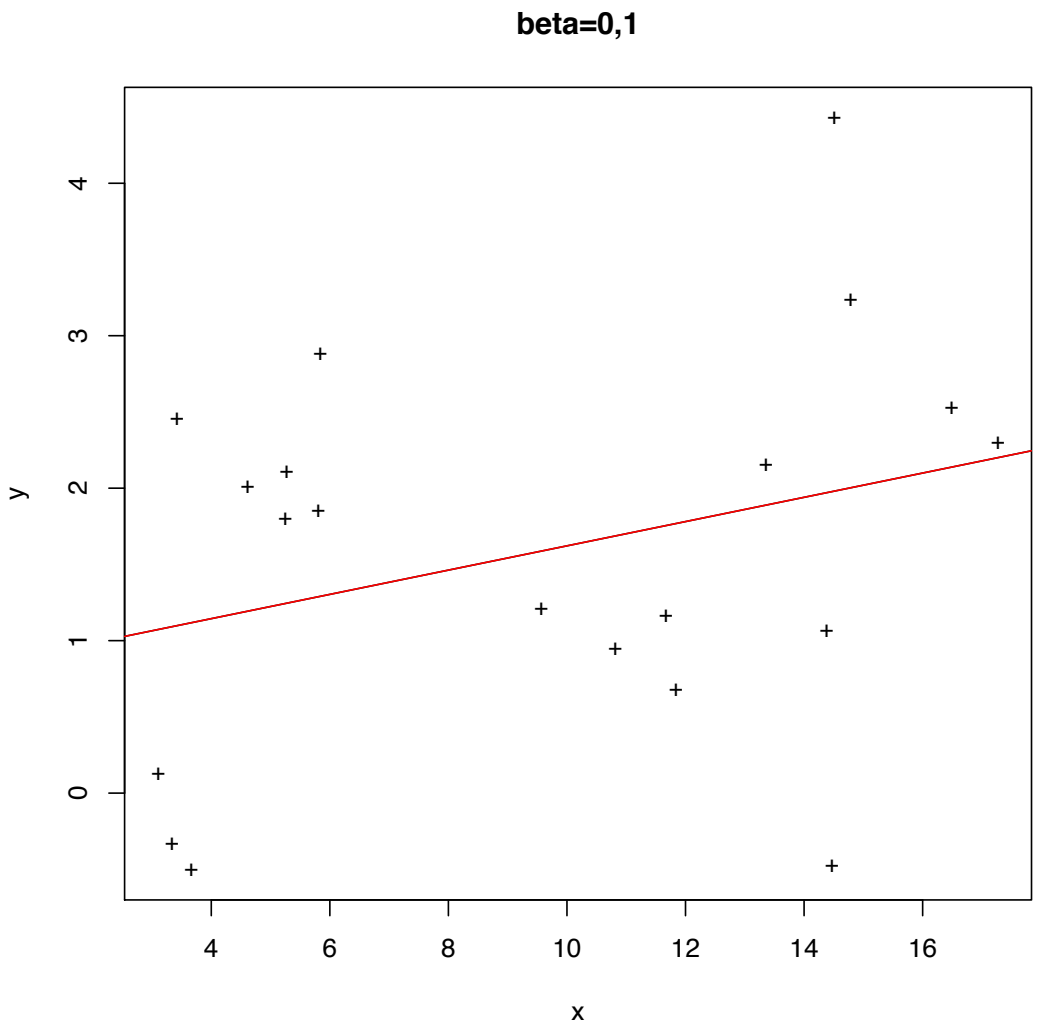
Linear Regression VII



Linear Regression VIII



Linear Regression IX



Linear Regression X

Twenty (x,y) pairs for each data set (α same).

β	p-value	Confidence Interval
1,5	$< 10^{-5}$	3,58 to 3,86
0,5	$< 10^{-5}$	0,383 to 0,598
0,1	0,187	-0.042 to 0,201

Which p-value is > 0,05 ?

Linear Regression X

Twenty (x,y) pairs for each data set (α same).

β	p-value	Confidence Interval
1,5	$< 10^{-5}$	3,58 to 3,86
0,5	$< 10^{-5}$	0,383 to 0,598
0,1	0,187	-0.042 to 0,201

Which p-value is > 0,05 ?

Which confidence Interval includes zero ?

Why does the same sample size not work in all 3 cases ?

Linear Regression X

Twenty (x,y) pairs for each data set (α same).

β	p-value	Confidence Interval
1,5	$< 10^{-5}$	3,58 to 3,86
0,5	$< 10^{-5}$	0,383 to 0,598
0,1	0,187	-0.042 to 0,201

Which p-value is > 0,05 ?

Which confidence Interval includes zero ?

Why does the same sample size not work in all 3 cases ?

How to improve for $\beta = 0,1$?

With 200 pairs

β	p-value	Confidence Interval
0,1	$< 0,01$	0,019 to 0,196

Linear Regression XI

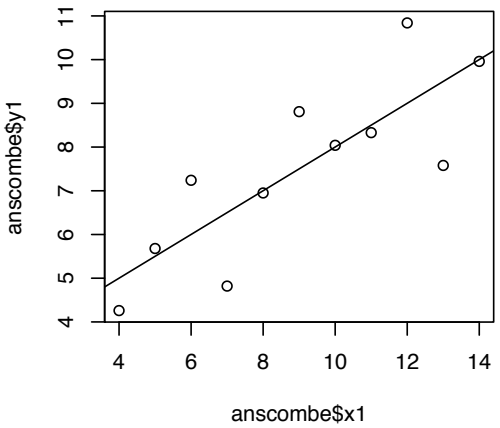
To check validity of linear regression not enough

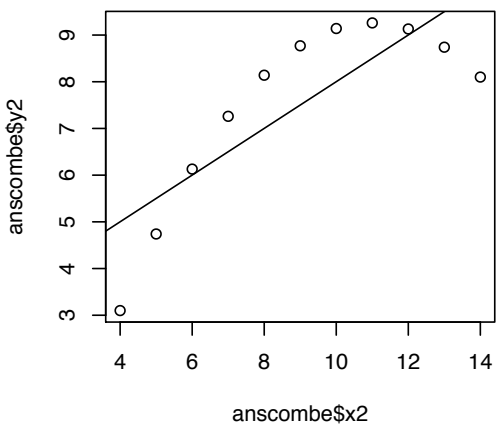
to test p-values and Confidence Intervals. Also need

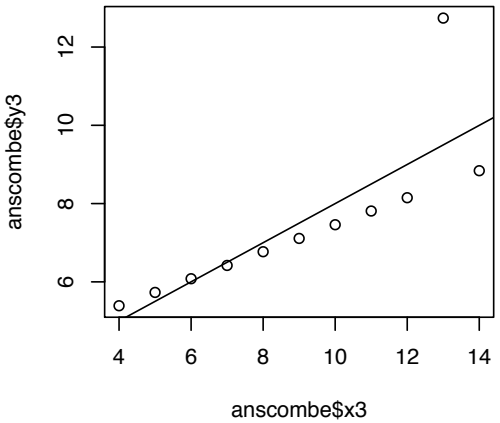
to see of the model is not obviously wrong

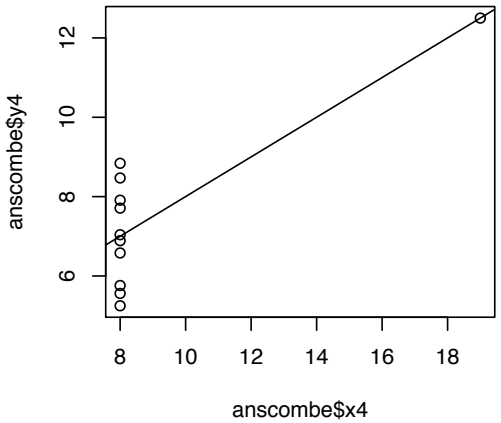
Some check can be done visually

Linear Regression XII



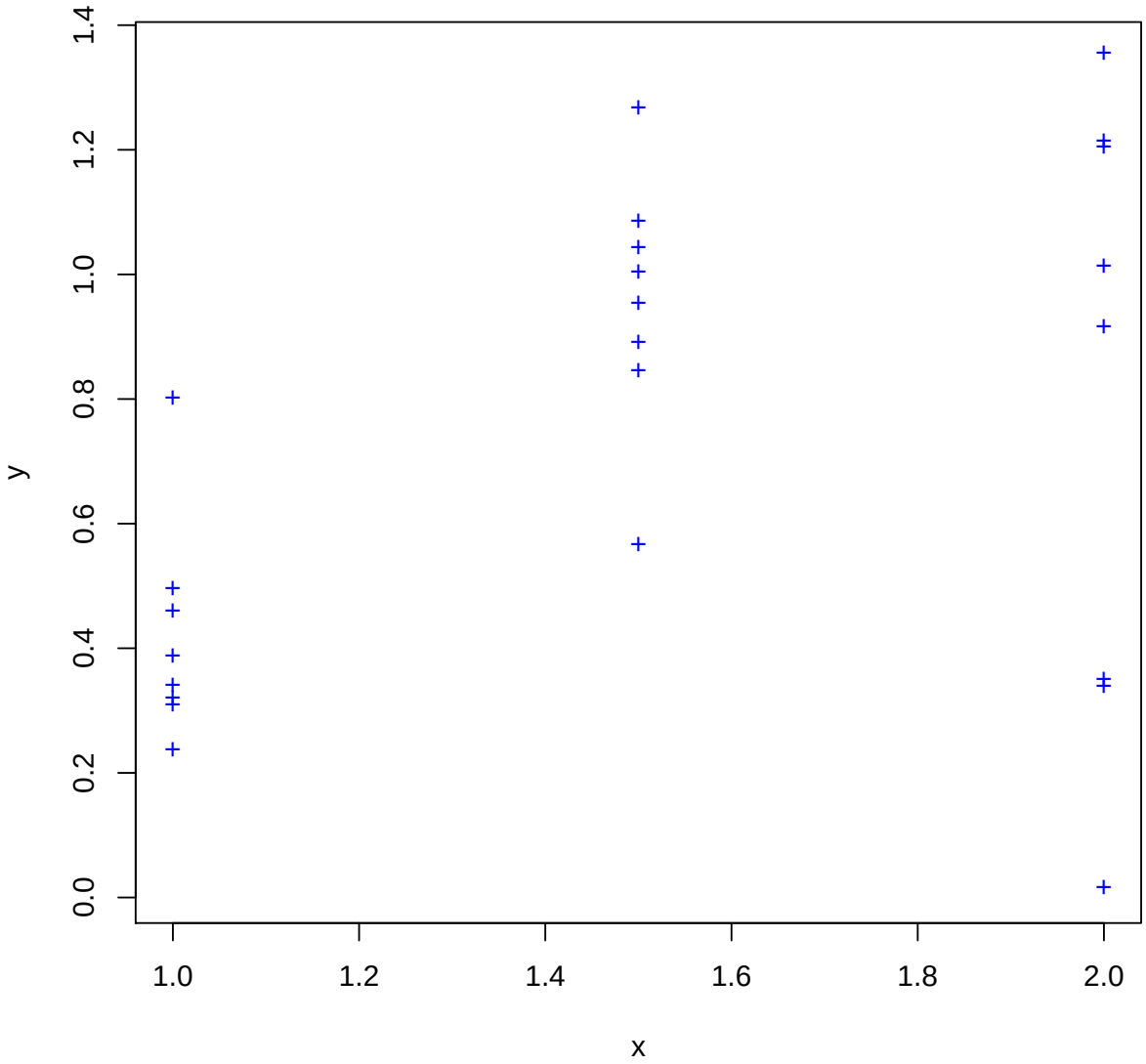






Anova I

Uptill now we assumed that in each pair (x, y) the x values were different. In some cases many x values may be the same, so that we have just a few distinct x values. This is one way to understand Anova. Date divides into a few distinct groups (treatments) which we need to compare. For two groups this is the $t - test$, for more than there are new features. What is H_0 ?


y
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.0
1.2
1.4
1.6
1.8
2.0

Anova III

H_0 is that all groups means are the same

In a gene expression experiment each gene has its own y. The groups are the same for all the genes. For each gene test for statistical significance. This is the idea for Affymetrix Agilent. many aspects similar for for RNA Seq. How to set this up ?

Experimental Design 0

To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

Ronald Fisher

Experimental Design I

Some terminology first:

Treatment: external condition whose effect we

want to study/compare (drug, hormone, temperature . . .).

Experimental Unit: Whatever receives the treatment.

Experimental Design I

Some terminology first:

Treatment: external condition whose effect we want to study/compare (drug, hormone, temperature . . .).

Experimental Unit: Whatever receives the treatment.

For comparing different treatments the experimenter assigns different treatments to experimental units in a **random** manner (**Randomization**).

Experimental Design II

To see what is not random imagine a clinical trial

to compare 2 drugs to reduce high blood pressure.

Idea is to compare the patients after one year of treatment.

If patients can choose which drug they want to try

why is this not a good thing ?

Experimental Design II

To see what is not random imagine a clinical trial

to compare 2 drugs to reduce high blood pressure.

Idea is to compare the patients after one year of treatment.

If patients can choose which drug they want to try

why is this not a good thing ? How does this affect

the conclusion for the comparison of the drugs ?

Experimental Design II

To see what is not random imagine a clinical trial

to compare 2 drugs to reduce high blood pressure.

Idea is to compare the patients after one year of treatment.

If patients can choose which drug they want to try

why is this not a good thing ? How does this affect

the conclusion for the comparison of the drugs ?

Why is random allocation better ?

Experimental Design III

Response Variable is what is observed after application of the treatments. For gene expression each gene supplies a response variable. For treatments for blood pressure what is the response variable ?

Experimental Design III

Response Variable is what is observed after application of the treatments. For gene expression each gene supplies a response variable. For treatments for blood pressure what is the response variable ? Sometimes we combine different treatment types, drugs and diets, fertilizer and genotypes.

Experimental Design IV

With multiple treatments can talk of

treatment factors (diet and drug, genotype and fertilizer, . . .)

Factors have different levels

drug1 & drug2, diet1, diet2, diet3 *etc.*

These are called treatment levels.

After randomization, the other key concept is **Replication.**

Replication is applying each treatment to multiple

different experimental units in a random manner.

Experimental Design V

We wish to compare two type of cattle feed (A & B) for the effect on growth. We consider two possible designs. In design 1 10 randomly selected cows in one farm will receive feed A and 10 randomly selected cows in another farm will receive feed B.

Experimental Design V

We wish to compare two type of cattle feed (A & B) for the effect on growth. We consider two possible designs. In design 1 10 randomly selected cows in one farm will receive feed A and 10 randomly selected cows in another farm will receive feed B. In design 2 5 cows in each farm receive feed A and 5 feed B.

Experimental Design V

We wish to compare two type of cattle feed (A & B) for the effect on growth. We consider two possible designs.

In design 1 10 randomly selected cows in one farm will receive feed A and 10 randomly selected cows in another farm will receive feed B. In design 2

5 cows in each farm receive feed A and 5 feed B.

In each design what are the treatments ?

Experimental Design V

We wish to compare two type of cattle feed (A & B) for the effect on growth. We consider two possible designs.

In design 1 10 randomly selected cows in one farm will receive feed A and 10 randomly selected cows in another farm will receive feed B. In design 2 5 cows in each farm receive feed A and 5 feed B.

In each design what are the treatments ?

In each design what are the experimental units ?

Which design has replication ?

Experimental Design V

We wish to compare two type of cattle feed (A & B) for the effect on growth. We consider two possible designs.

In design 1 10 randomly selected cows in one farm will receive feed A and 10 randomly selected cows in another farm will receive feed B. In design 2 5 cows in each farm receive feed A and 5 feed B.

In each design what are the treatments ?

In each design what are the experimental units ?

Which design has replication ? Which design is better ?

Experimental Design VI

In both designs, the cows are the **observational units**.

Design 2 is better. In Design 1 any change could be due to treatment or due to some difference in the farms.

Experimental Design VI

In both designs, the cows are the **observational units**.

Design 2 is better. In Design 1 any change could be

due to treatment or due to some difference in the farms.

The purpose of experimental design is to ensure

that observed differences are due to **treatment differences !**

Experimental Design VI

In both designs, the cows are the **observational units**.

Design 2 is better. In Design 1 any change could be

due to treatment or due to some difference in the farms.

The purpose of experimental design is to ensure

that observed differences are due to **treatment differences !**

Not due to other factors, dye, reagent, farm . . ..

These other factors are **confounding factors** which

are related to treatment and outcome.