Applied statistics with applications in R (and SAS)

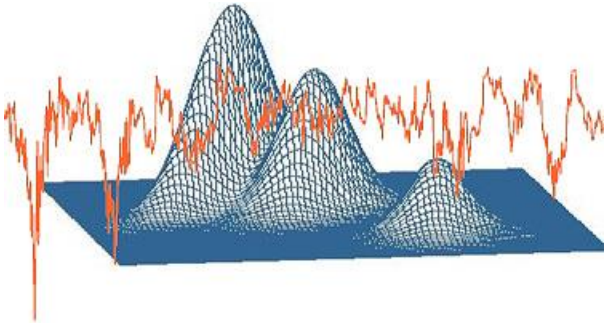
Alessio Cecchinato

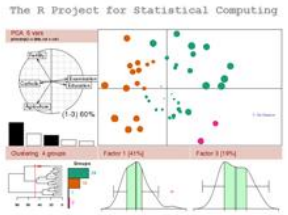
Hugo Toledo Alvarado

DAFNAE - Department of Agronomy, Food

Natural resources, Animals and Environment

University of Padova





Overview

- We will cover analysis of variance and analysis of covariance. The emphasis will be placed on selected practical tools using R software (and SAS, just some spots) rather than on the mathematical manipulations. Want to understand the theory so that we can apply it appropriately

Overview

1. Introduction to hypothesis testing
2. How to state a null hypothesis and alternative hypothesis
3. How to identify errors and interpret the level of significance
4. Analysis of variance (ANOVA):
5. One-way ANOVA
6. Two-way ANOVA
7. Two-way ANOVA and interactions
8. ANCOVA
9. Exercises and applications with R software

Overview

Team Project:

- Data will be provided to a team composed of two to three students and asked to employ techniques learned throughout this course to analyze the data set, interpret and report results

Final Exam:

- Written examinations. A written examination consists of exercises designed to test the basic knowledge acquired during the course

References

R users:

- Crawley, M. J. (2012). The R book. John Wiley & Sons.
- Kabacoff, R. I. (2010). R in Action. Manning.

SAS users:

- Text: Applied Linear Statistical Models, (4th ed.) by Neter, Kutner, Nachtsheim and Wasserman
- SAS System for Regression (2nd ed.) by Freund and Little
- SAS/STAT User's Guide Vol. 1 and 2

Notes of the course can be found here:

https://github.com/Hugo-Toledo/Applied-Statistics-R-UNIPD

A brief introduction: Scientific Method

a) Review and Research the problem

↓

b) Formulate Hypothesis

↓

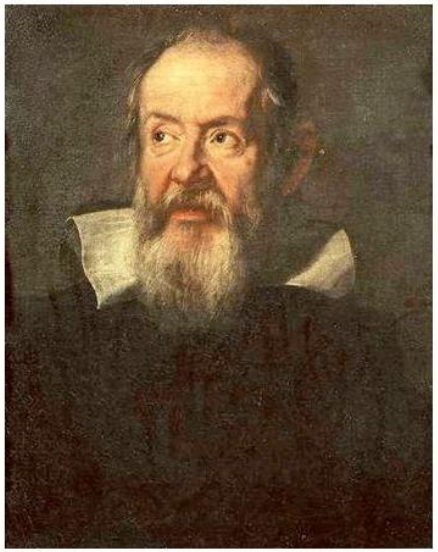
c) Design experiment that will allow test of hypothesis

↓

d) Evaluate the hypothesis

↓

e) Draw Conclusions



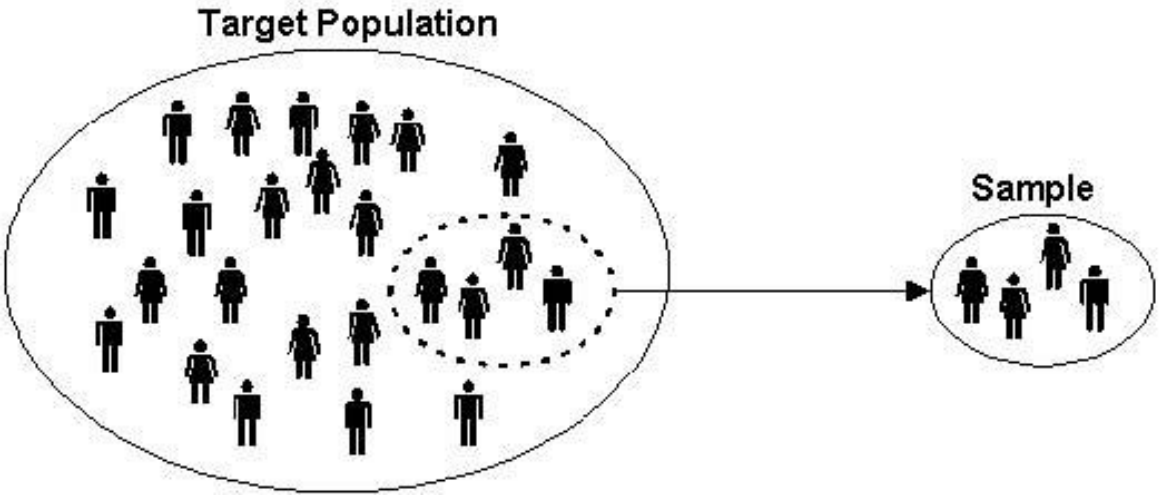
Galileo Galilei 1564-1642

How does statistical inference work?

- Infer upon the characteristics of a large population based on data from a finite random sample
- Mechanics (part of the scientific method):
  - Design and collect data from an experiment (e.g blood pressure)
  - Assess the probability of getting the experimental results assuming a true null hypothesis (status quo knowledge..e.g. no treatment difference)
  - Common investigator objective: disprove the status quo in favor of an alternative hypothesis (there is a treatment effect)
- Conclusions never made with absolute certainty…must establish proof beyond a reasonable doubt

Terminology

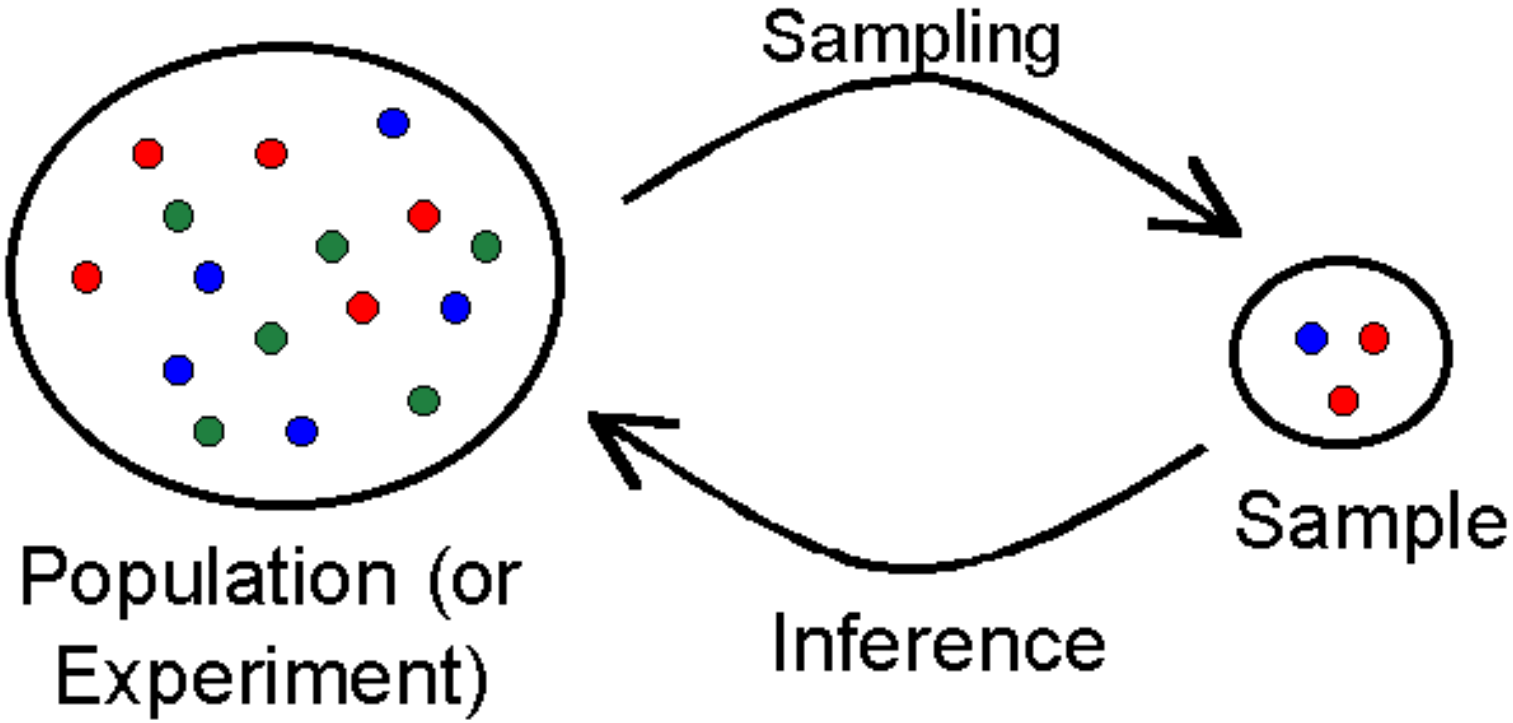
- **Population**: The collection of all responses, measurements, or counts that are of interest
- **Sample**: A portion or subset of the population



Population:
μ= mean
σ= standard deviation

Sample:
$\bar{x}$= mean
s = standard deviation

Populations and samples



Do it yourself …

- Identify the population and the sample in the following example: In a recent survey 2000 European Boxer dogs owners were asked if they submitted their dogs to x-ray screening for hip dysplasia. Eight hundred of the owners answered yes

- What is the population ?

- What is the sample ?

- What is the data set

Variables, Variables!!!

- Quantitative variables:
  - Due to a true numerical measurement
  - Ratio scale (e.g. weight) versus Interval Scale (e.g. temperature)
  - Discrete (countable) versus Continuous

- Qualitative variables:
  - Nominal scale (classification or group) GENOTYPE, SEX
  - Ordinal scale (ranked variables..small, medium, large)

Data types

Variables

- **Scale**
  - **Continuous**
    Measurements
    takes any value
  - **Discrete:**
    Counts/ integers
- **Categorical**
  - **Ordinal:**
    obvious order
  - **Nominal:**
    no meaningful order

Questionnaire

What data types relate to following questions?

- Q1: What is your favourite subject?

Maths	English	Science	Art	French

Nominal

- Q2: Gender:

Male	Female

Binary/ Nominal

- Q3: I consider myself to be good in statistics:

Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree

Ordinal

- Q4: Score in a recent exam:

Score between 0% and 100%

Scale

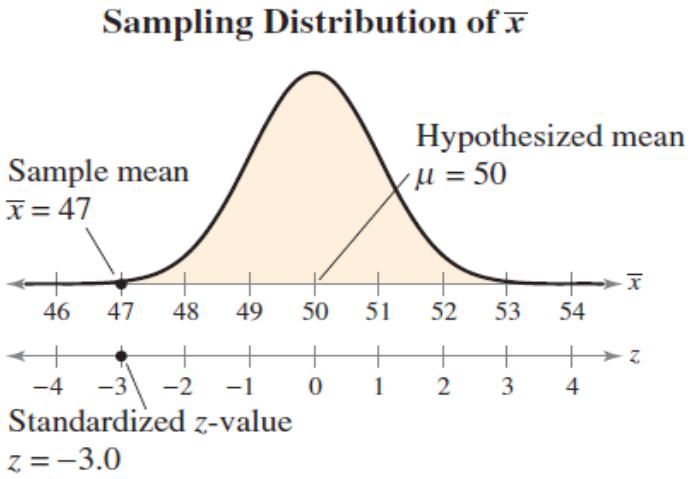
A (practical) introduction to hypothesis testing

Hypothesis tests

- A hypothesis test is a process that uses sample statistics to test a claim about the value of population parameter
- Example: suppose an automobile manufacturer advertises that its new hybrid car has a mean gas mileage of 50 miles per gallon.
- If you suspect that the mean mileage is not 50 miles per gallon, how could you show that the advertisement is false?
- Obviously, you cannot test *all* the vehicles, so we draw a sample from the population …
- There is a large probability that the value of the sample statistic differs from the that of the claim. Is it enough to conclude that the claim is wrong ?

Hypothesis tests

- Let's **assume the claim is correct**!
- you could take a random sample of 30 vehicles and measure the mileage of each ($\bar{x}$=47 miles/gallon and s = 5.5 miles/gallon)
- Recalling CLT, we know that the sampling distribution of the sample mean is normal with a mean of 47 and a standard error of $5.5/\sqrt{30} \approx 1$



- So either the sample is very unusual or the claim is false

Hypothesis tests

- This is an unusual event!
- Your assumption that the company's advertisement is correct has led you to an improbable result. So, either you had a very unusual sample, or the advertisement is probably false.
- **The logical conclusion is that the advertisement is probably false**

Stating a hypothesis

- A claim about a population parameter is called a statistical hypothesis
- To test a statistical hypothesis, we must carefully state a pair of hypotheses:
  - ***the claim***
  - ***its complement***
- When the first is true, the other is false
- Of the two hypotheses, the one that contains a statement of equality (=, ≤, ≥) is called the **null hypothesis (H_0)**, the other one (≠, <, >) is the **alternative hypothesis (H_a)**

Examples

Write the claim as a mathematical sentence. State the null and alternative hypotheses, and identify which represents the claim

- A school publicizes that the proportion of its students who are involved in at least one extracurricular activity is 61%
  - $H_0 : p = 0.61$ Claim: "the proportion is 61%"
  - $H_a : p \neq 0.61$
- A car leadership announces that the mean time for an oil change is less than 15 minutes
  - $H_0 : \mu \geq 15\ minutes$
  - $H_a : \mu < 15\ minutes$ Claim: "the mean is less than 15 minutes"
- A company advertises that the mean life of its furnaces is more than 18 years
  - $H_0 : \mu \leq 18\ years$
  - $H_a : \mu > 18\ years$ Claim: "the mean is more than 18 years"

Types of errors and level of significance

- We always test the null hypothesis H_0
- So, after the test, we make one of two possible decisions:
  - **reject the null hypothesis or**
  - **fail to reject the null hypothesis**
- Because the decision is based on incomplete information (the sample) there is always the chance of making the wrong decision
- There are two types of errors we can make:
  - **A type I error** occurs if the null hypothesis is rejected when it is true
  - **A type II error** occurs if the null hypothesis is not rejected when it is false

Types of errors and level of significance

	Truth of H_0	
Decision	**H_0 is true.**	**H_0 is false.**
Do not reject H_0.	Correct decision	Type II error
Reject H_0.	Type I error	Correct decision

Controlled via sample size (=1-Power of test)

Typically restrict to a 5% Risk = level of significance

Prob of this = Power of test

- In a hypothesis test, the **level of significance** α is your maximum allowable probability of making a type I error
- By setting the level of significance at a small value, we say that we want the probability of rejecting a true null hypothesis to be small
- Commonly used level of significance are $\alpha = 0.1$, $\alpha = 0.05$ and $\alpha = 0.01$

Summary for hypothesis testing

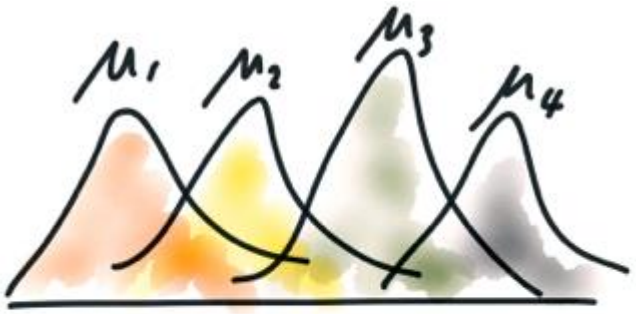
1. State the claim mathematically and verbally. Identify the null and alternative hypotheses (H_0 = ?, H_a = ?)
2. Choose the level of significance α = ?
3. Identify the type of test (left-, right-, or two-tailed) and identify the standardized sampling distribution (z, t, etc)
4. Determine critical values (z_0, t_0, etc)
5. Determine rejection regions
6. Calculate the test statistic and its standardized value
7. Decision rule:

   Is the standardized test statistic in the rejection region ?

   Yes → **Reject H_0**

   No → **Fail to reject H_0**
8. Interpret the decision in the context of the original claim

Analysis of variance (ANOVA)



Analysis of variance (ANOVA)

- Analysis of variance (ANOVA) is a hypothesis-testing technique that is used to compare means from three or more populations
- The response variable Y is continuous
- The explanatory variable is *categorical*
  - We call it a *factor*
  - The possible values are called *levels*
- This approach is a generalization of the *independent two-sample t-test*
- In other words, it can be used when there are more than two treatments

Analysis of variance (ANOVA)

- A **factor** refers to a categorical quantity under examination in an experiment as a possible cause of variation in the response variable

- **Levels** refer to the categories, measurements, or strata of a factor of interest in the experiment

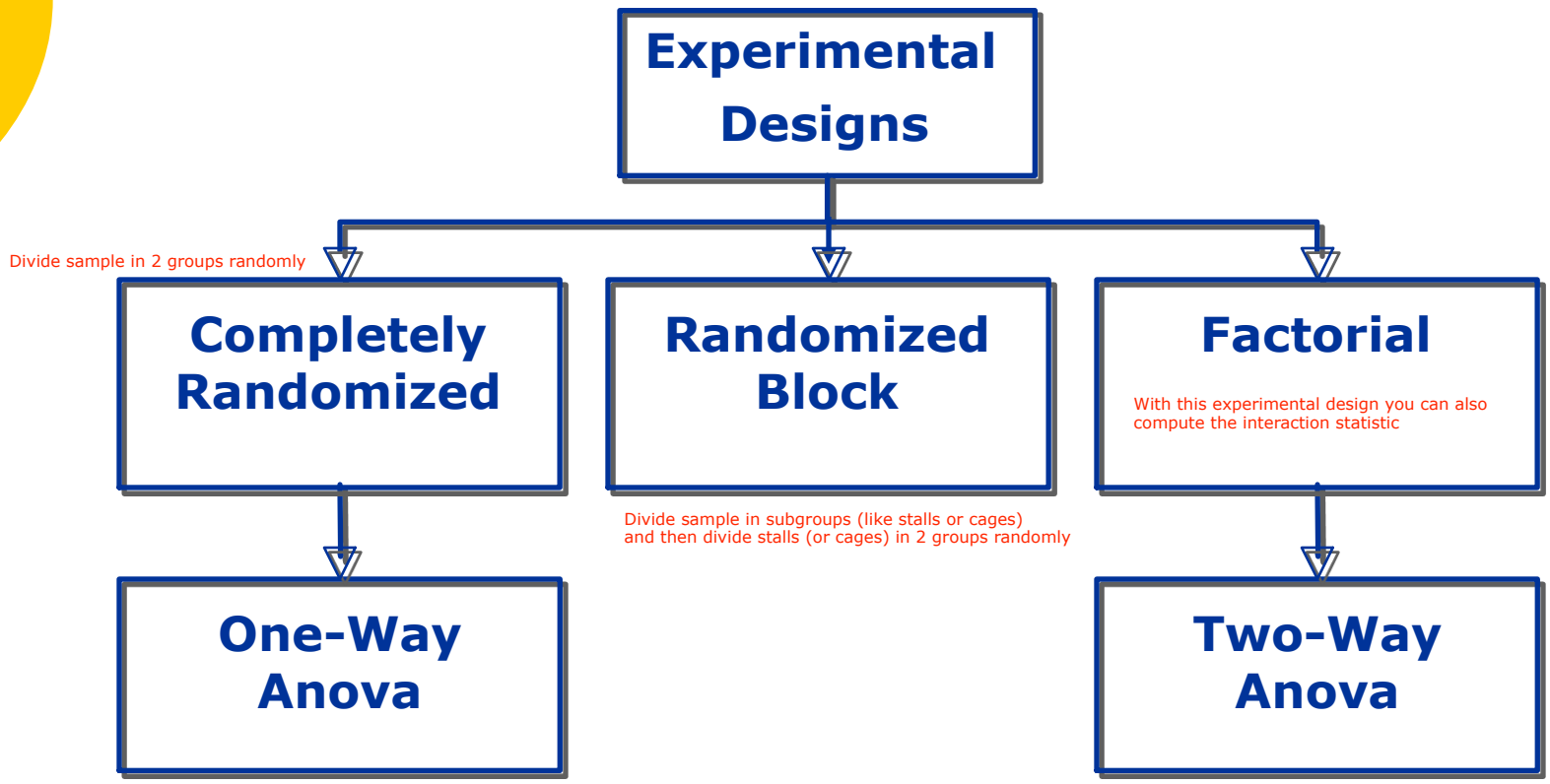
Data for One-Way ANOVA

- Y is the response variable
- X is the factor (it is qualitative/discrete)
  - r is the number of levels
  - often refer to these levels as groups or treatments
- Y_{ij} is the j^{th} observation in the i^{th} group

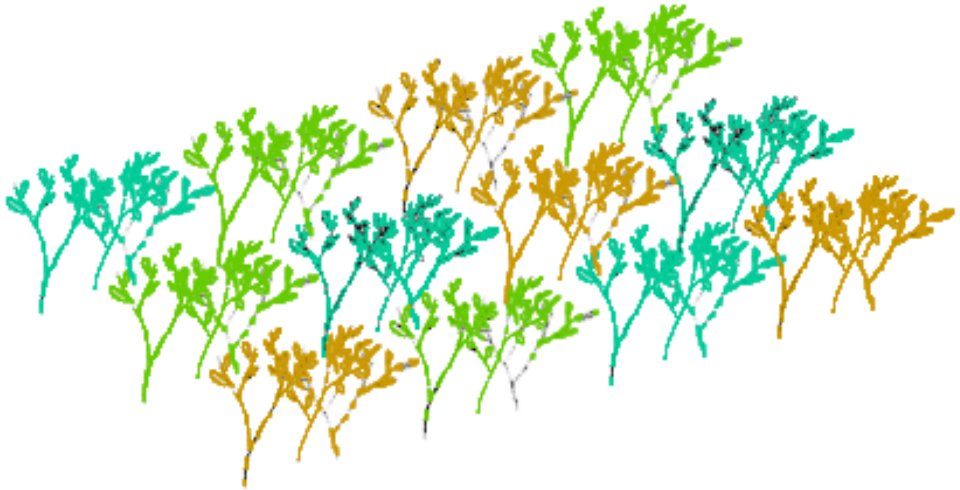
Notation

- For $Y_{i,j}$ we use
  - i to denote the level of the factor
  - j to denote the j^{th} observation at factor level i
- i = 1, . . . , r levels of factor X
- j = 1, . . . , n_i observations for level i of factor X

Types of Experimental Designs

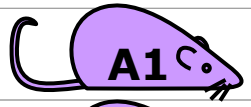
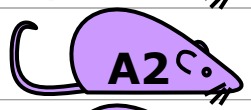


Completely Randomized Design



Blood cholestorol example

- Experimental study with randomization:
  - 6 rats assigned at random to one of 2 treatments (n=3 rats per treatment)
- Blood Cholesterol → data collected (mg/dl):

New Treatment (A)	Control Treatment (B)
y_{A1} = 144 	y_{B1} = 149 
y_{A2} = 148 	y_{B2} = 148 
y_{A3} = 146 	y_{B3} = 159 

- Would you conclude the treatments lead to different mean blood cholestorol levels ????

Ave $\bar{y}_A$ = 146 Ave $\bar{y}_B$ = 152

A mean treatment difference is found!

Statistical inference

But is it…….

Due to mere chance (biological noise) ???

Or…

The real thing (beyond reasonable doubt)?

Completely Randomized Design

- Experimental units (subjects) are assigned randomly to treatments
  - Subjects are assumed homogeneous
- One factor or independent variable
  - 2 or more treatment levels or groups
- Analyzed by one-way ANOVA

One-Way ANOVA F-Test

- Tests the equality of 2 or more (p) population means

- Variables
  - One nominal independent variable
  - One continuous dependent variable

One-Way ANOVA F-Test Assumptions

- Randomness & independence of errors
- Normality
  - Populations (for each condition) are normally distributed
- Homogeneity of variance
  - Populations (for each condition) have equal variances

Assumption

(1) $y_i = \alpha + \beta_1 x_{1i} + \ldots + \beta_n x_{ni} + \varepsilon_i$

y is a linear function of a,β,x_1, …, x_n and e (residual)

(2) $E(\varepsilon_i) = 0$

e does not depend on x_1, …, x_n (e and x_1, …, x_n are not correlated)

(3) $Var(\varepsilon_i) = \sigma_e^2$

Omoscedasticy assumption: the variance of e is equal for all the observations

(4) $Cov(\varepsilon_i, \varepsilon_j) = 0$ dove $i \neq j$

The residual of each observation is not related with the residual of other observations

(5) $\varepsilon_i \sim N(0, \sigma_e^2)$

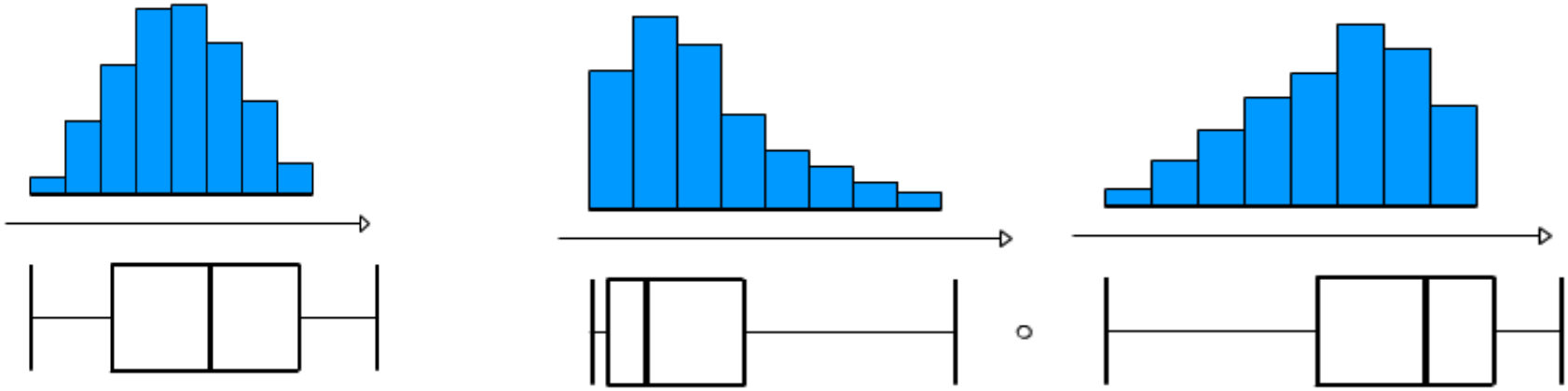
The residuals are normally distributed

Assessing Normality

Charts can be used to **informally** assess whether data is:

Normally distributed

Or….Skewed



The mean and median are very different for skewed data

Normality test

- Wide variety of tests we can perform to test if the data follows a normal distribution
- Mardia (1980) provides an extensive list for both the univariate and multivariate cases, categorizing them into two types
  - Properties of normal distribution, more specifically, the first four moments of the normal distribution
    - Shapiro-Wilk's W (compares the ratio of the standard deviation to the variance multiplied by a constant to one)
  - Goodness-of-fit tests,
    - Kolmogorov-Smirnov D
    - Cramer-von Mises W^2
    - Anderson-Darling A^2

Mardia, K.V. (1980): Tests of univariate and multivariate normality. In P.R. Krishnaiah (ed.), Handbook of statistics (vol. 1; pp 279-320). Amsterdam: North Holland.

Normality test

Descriptive Statistics and Normality Test

```
milk<-cows$milk      # Select the variable
summary(milk)        # Basic statistics
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.30   24.88   29.05   30.09   35.30   50.90
```

```
sd(milk)             # Standard Deviation function
```

```
## [1] 7.614805
```

```
range(milk)          # Range function
```

```
## [1] 15.3 50.9
```

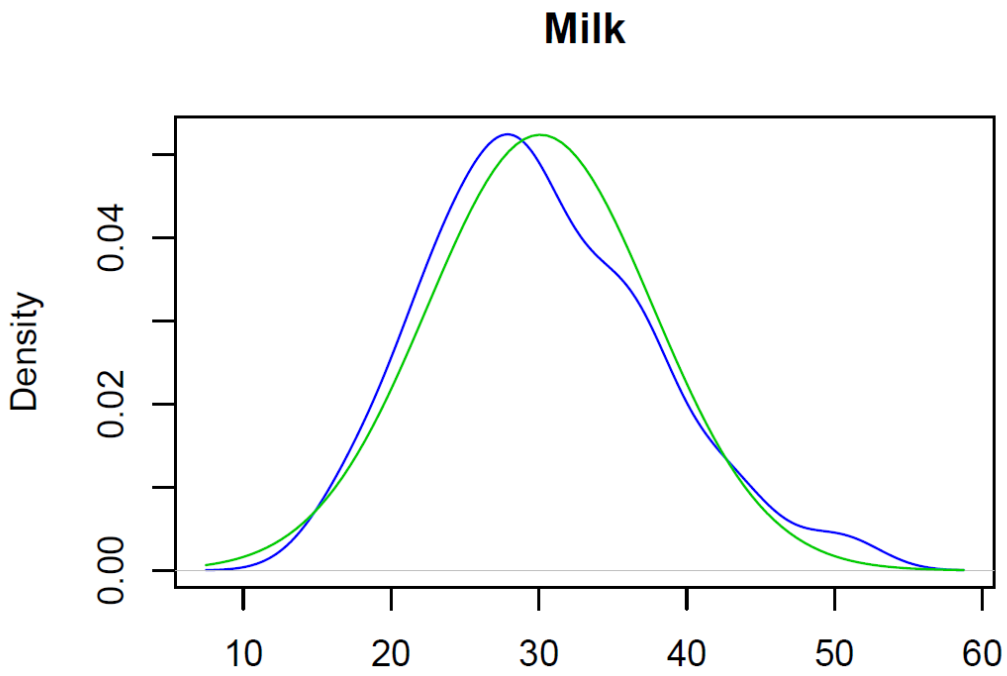
```
shapiro.test(milk)   # Shapiro - Wilk normality test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  milk
## W = 0.97876, p-value = 0.04779
```

Normality test

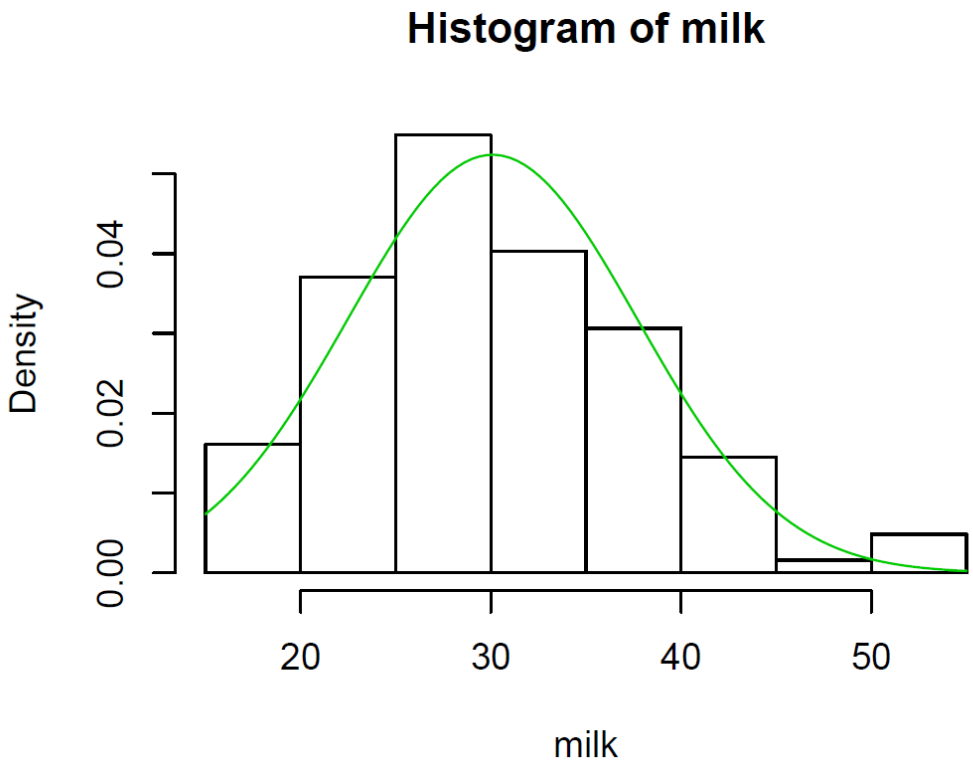
Density plot with normal distribution curve

```
plot(density(milk), col=4, main="Milk") # Create a density plot
curve(dnorm(x,mean=mean(milk),sd = sd(milk)),add = T,col=3) # Add normal dist.
```



Normality test

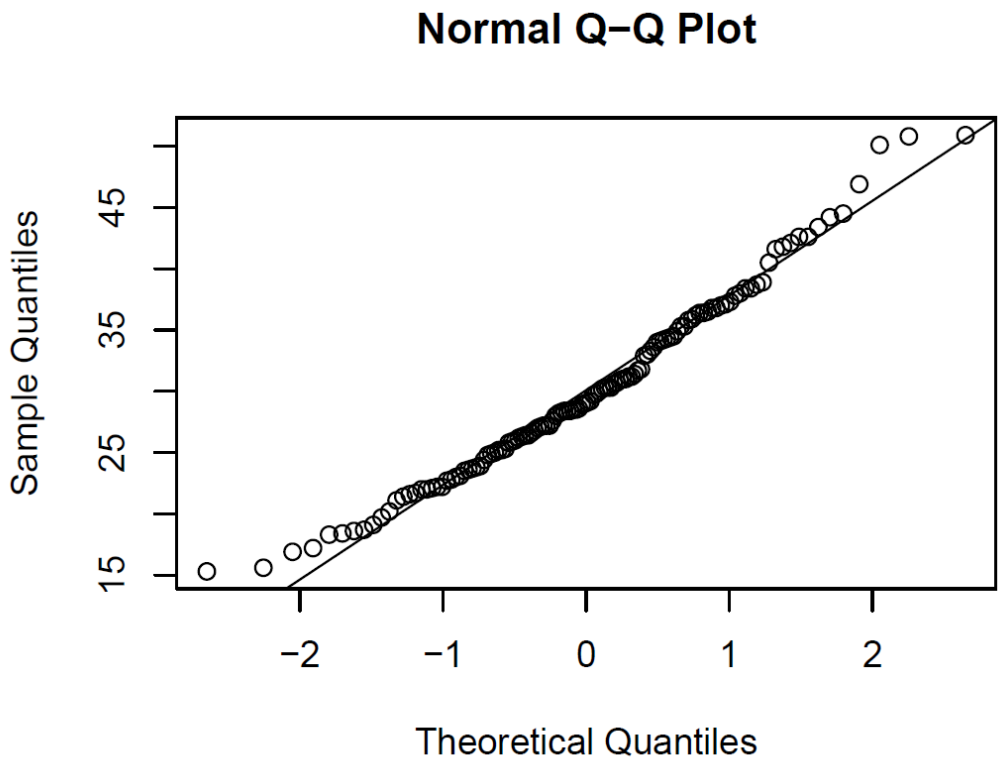
```
hist(milk, breaks = 10, freq = FALSE) # Create histogram
curve(dnorm(x,mean=mean(milk),sd = sd(milk)),add = T,col=3) # Add normal dist.
```



Normality test

Q-Q Plot

```
qqnorm(cows$milk) #  Create a qq-plot
qqline(cows$milk) # Add line
```



Normality test

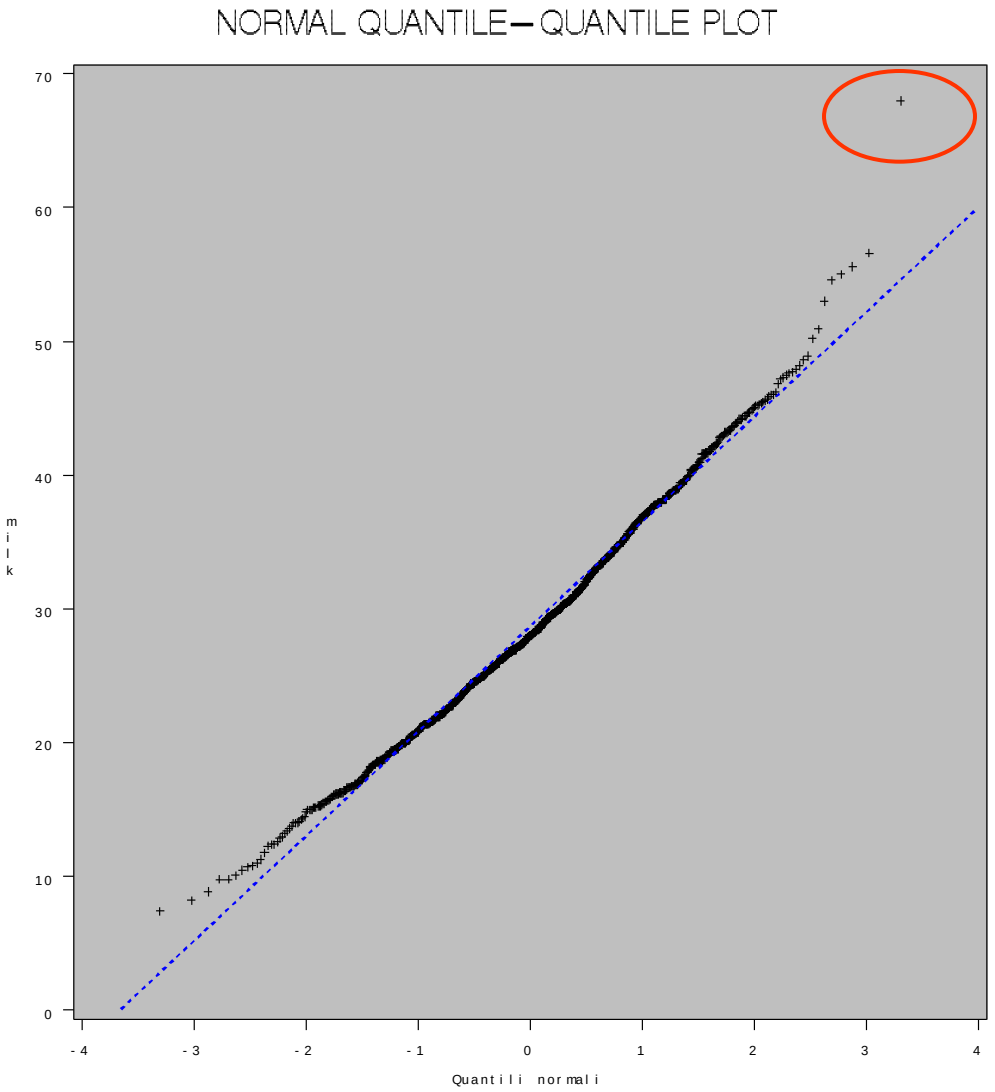
```
********verifica normalità*******************;

PROC UNIVARIATE DATA=brown NOPRINT;
VAR milk grs prt cellu scs casein indcas r k20 a30 sh ph pnc;
HISTOGRAM milk grs prt cellu scs casein indcas r k20 a30 sh ph pnc/NORMAL;
INSET NORMAL(AD ADPVAL CVM CVMPVAL KSD KSDPVAL);
RUN;
```

```
   SYMBOL1 v=PLUS;
   TITLE 'NORMAL QUANTILE-QUANTILE PLOT';
```

```
   PROC UNIVARIATE DATA=brown NOPRINT;
     QQPLOT milk  grs prt cellu scs casein indcas r indcas k20 a30 sh ph pnc/ normal(mu=est sigma=est color=blue
                              l=2 w=2 noprint)
                        square cframe = ligr;
   RUN;
```

```
   GOPTIONS RESET=ALL;
```

Normality test

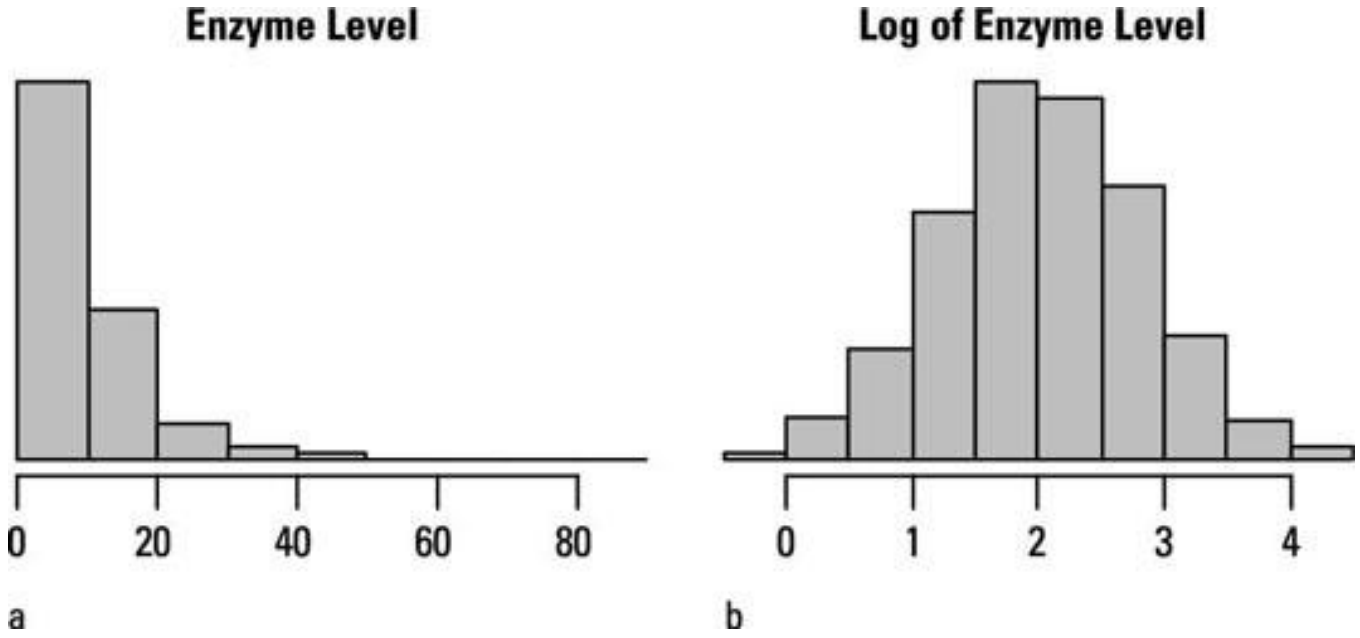


Consequences of Non-Normality

- F-test is very robust against non-normal data, especially in a fixed-effects model
- Large sample size will approximate normality by Central Limit Theorem (recommended sample size > 50)
- Simulations have shown unequal sample sizes between treatment groups magnify any departure from normality
- A large deviation from normality leads to hypothesis test conclusions that are too liberal and a decrease in power and efficiency

Remedial Measures for Non-Normality

- Data transformation (log, square root, etc…)
- Be aware - transformations may lead to a fundamental change in the relationship between the dependent and the independent variable and is not always recommended



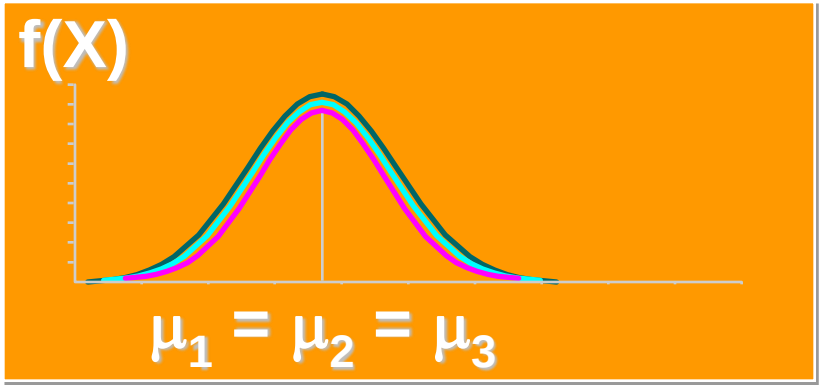
One-Way ANOVA F-Test Hypotheses

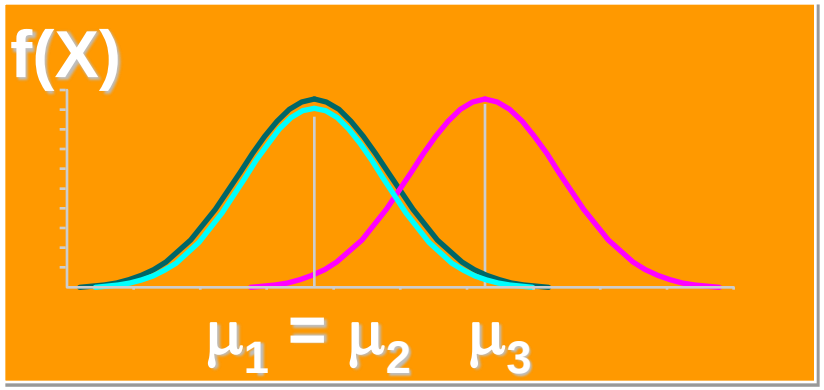
H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_p$

- All Population Means are Equal
- No Treatment Effect

H_a: Not All μ_j Are Equal

- At Least 1 Pop. Mean is Different
- Treatment Effect
- NOT $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_p$

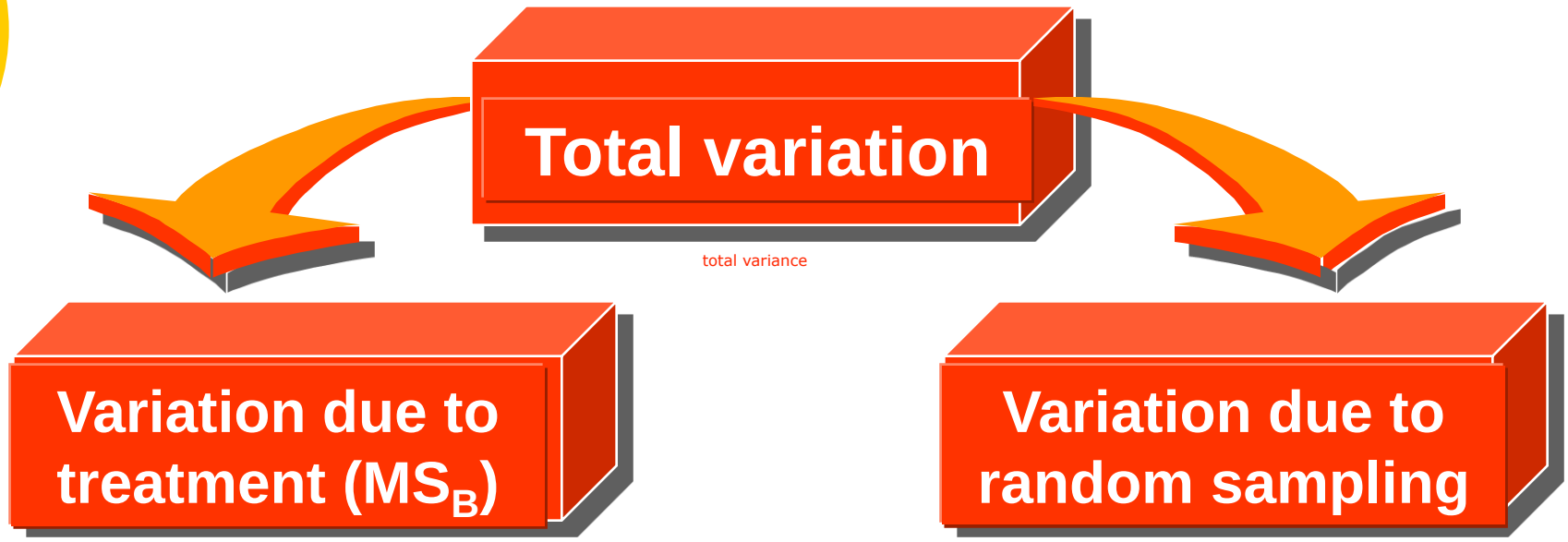




One-Way ANOVA Basic Idea

- Compares 2 types of variation to test equality of means
- If treatment variation is significantly greater than random variation then means are not equal
- Variation measures are obtained by ‘partitioning’ total variation

One-Way ANOVA Partitions Total Variation



Sum of squares due to treatment **SST** (i.e, between groups)

between-groups variation

Sum of squares error **SSE** (i.e, within groups)

One-Way ANOVA Partitions Total Variation

sum of squares due to the error (i.e. within-group variation)

$$SS_{ERROR} = \sum_j \sum_i (Y_{ij} - \hat{Y}_{ij})^2$$

→ "within"

sum of squares due to the treatment (i.e. between-groups variation)

$$+ \quad SS_{TREAT} = \sum_j \sum_i (\hat{Y}_{ij} - \bar{Y})^2$$

→ "between"

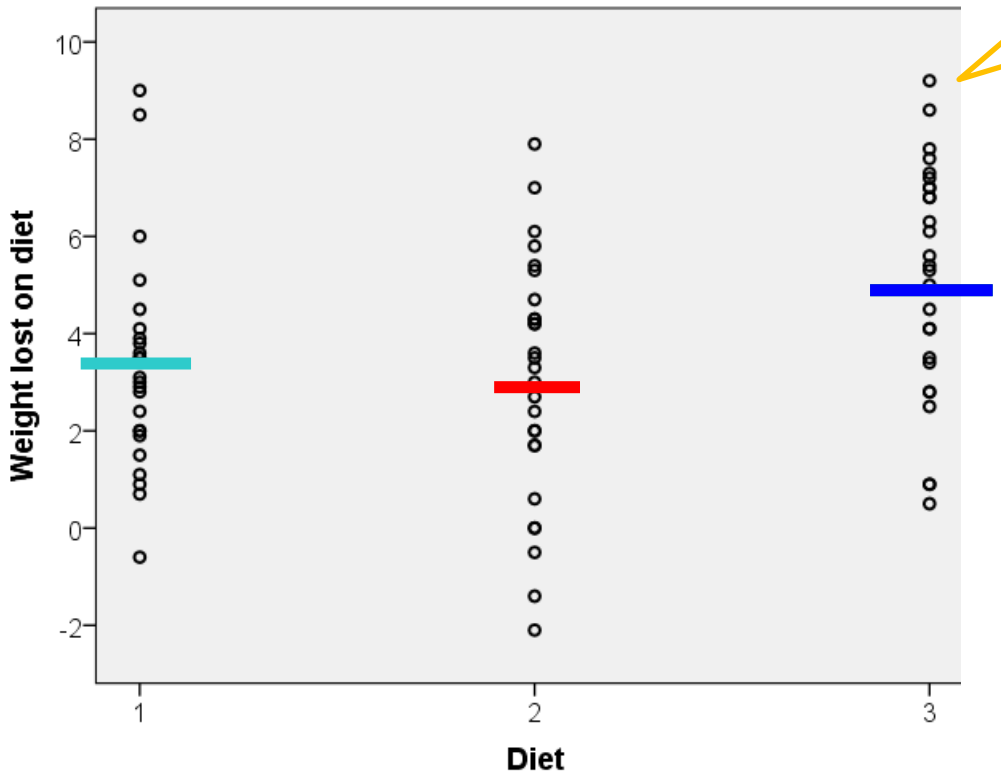
Y hat is the grand mean, i.e. the mean computed from the entire sample

$$SS_{TOTAL} = \sum_j \sum_i (Y_{ij} - \bar{Y})^2$$

Within group variation

Residual =difference between an individual and their group mean

$$SS_{ERROR} = \sum_j \sum_i (Y_{ij} - \hat{Y}_{ij})^2$$



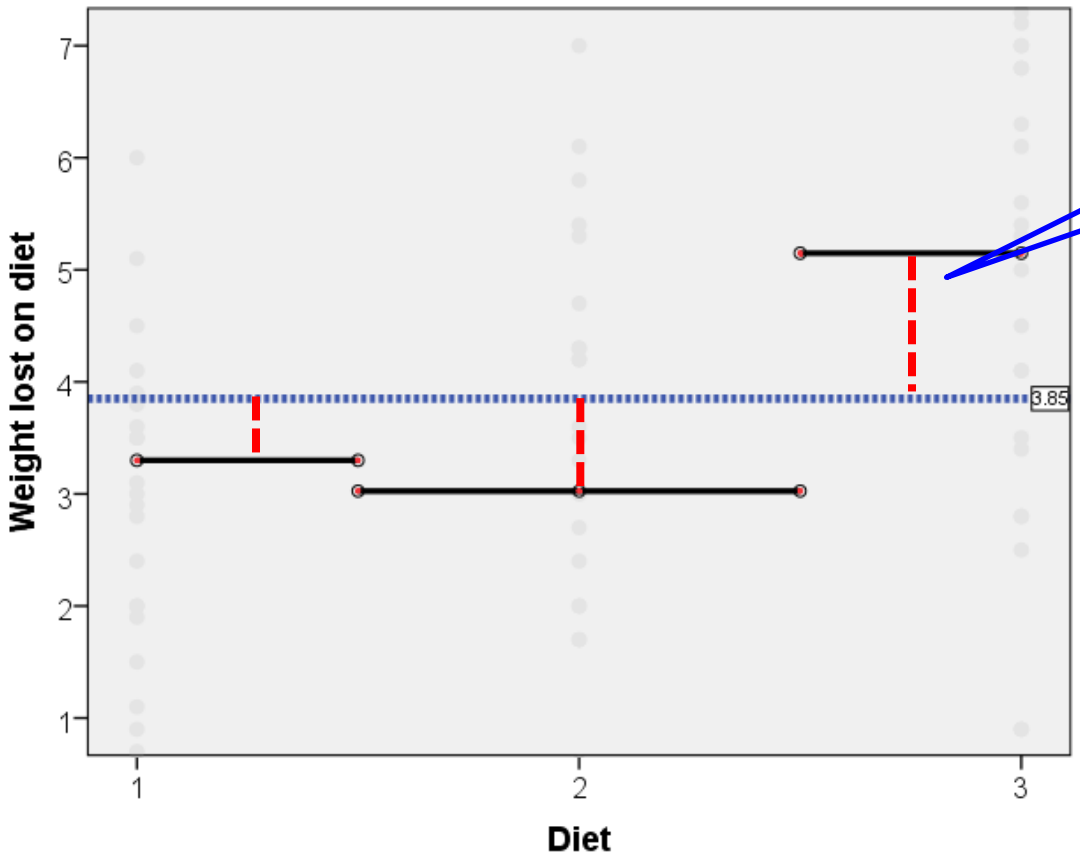
Person lost 9.2kg kg so residual =9.2 – 5.15 = 4.05

Mean weight lost on diet 3 = 5.15kg

Between group variation

Differences between each group mean and the overall mean

$$SS_{TREAT} = \sum_j \sum_i (\hat{Y}_{ij} - \bar{Y})^2$$



Diet 3 difference =5.15 – 3.85 = 1.3

Overall mean = 3.85

One-Way ANOVA F-Test

- 1. Test Statistic
  - $F = MST / MSE = \dfrac{SST/(p-1)}{SSE/(n-p)}$
    - MST Is Mean Square for Treatment
    - MSE Is Mean Square for Error
- 2. Degrees of Freedom
  - $df_1 = p - 1$
  - $df_2 = n - p$
    - p = # Populations, Groups
    - n = Total Sample Size

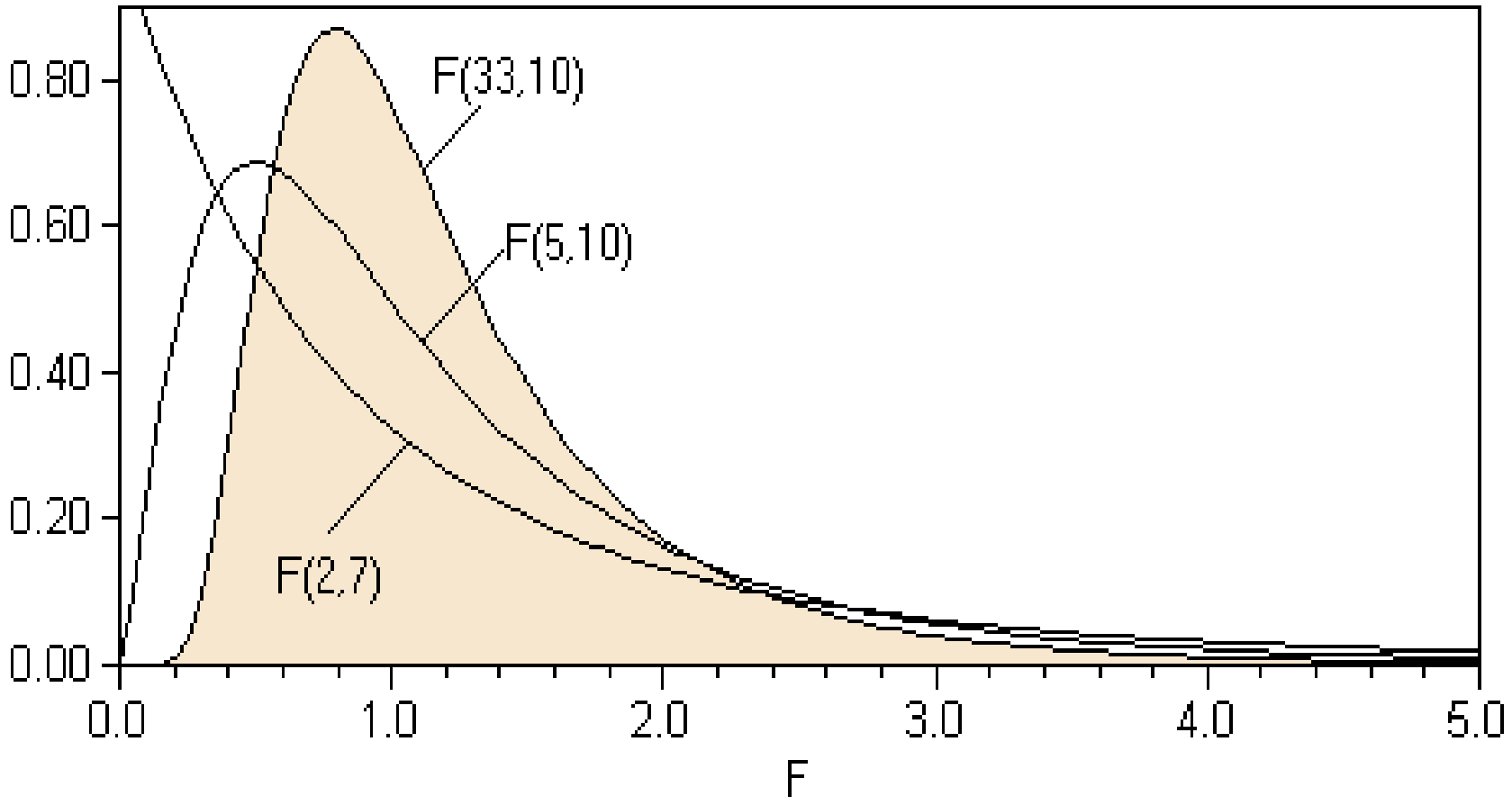
One-Way ANOVA - Summary Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Treatment	p - 1	SST	MST = SST/(p - 1)	$\frac{MST}{MSE}$
Error	n - p	SSE	MSE = SSE/(n - p)	
Total	n - 1	SS(Total) = SST+SSE		

ANOVA – the sampling distribution for F

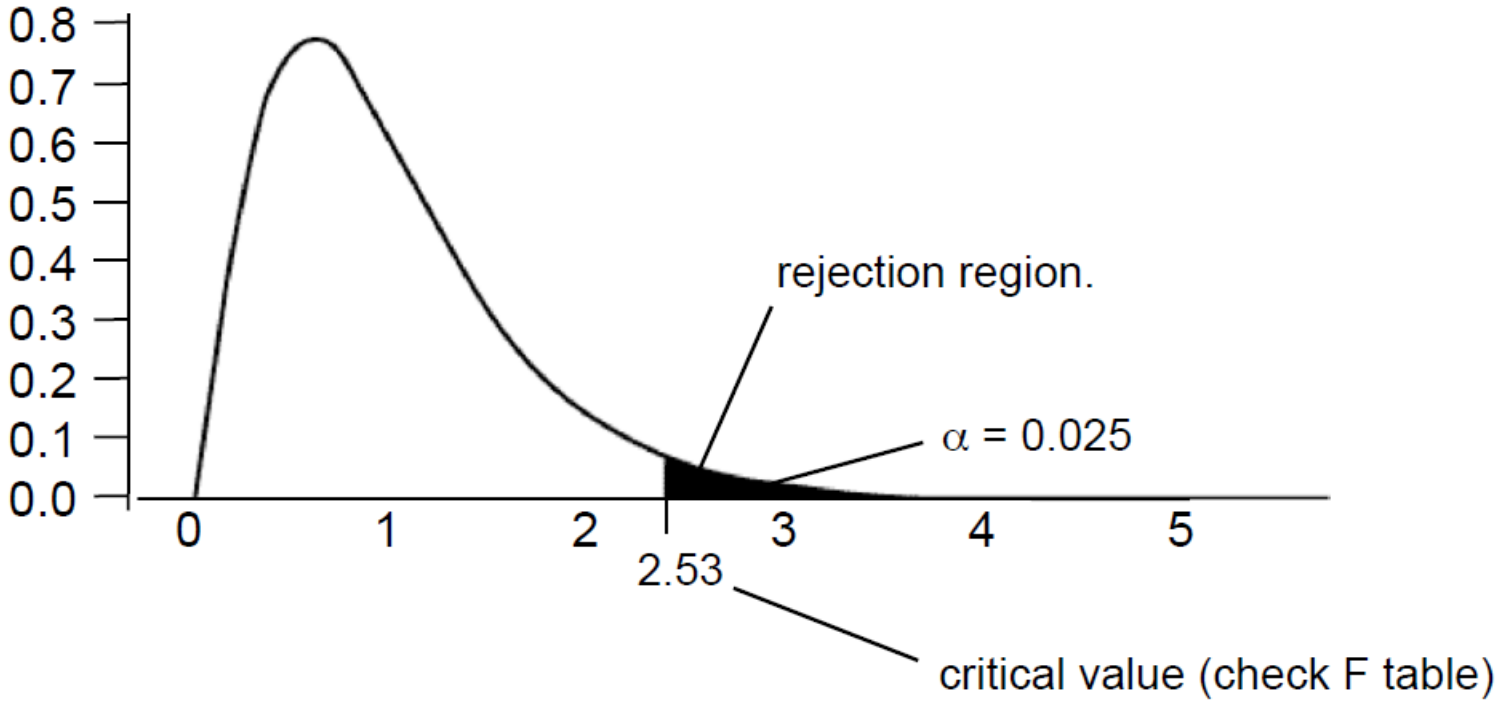
- The sampling distribution for the test statistic F is an F (Snedecor) distribution with k-1 degrees of freedom (df_N) for the numerator (df_N) and N-k degrees of freedom for the denominator (df_D)
- The F distribution :
  - is a family of curve each of which determined by the df of the numerator and by the df of the denominator
  - is positively skewed
  - has a total area under the curve equal to 1
  - F-values are greater than or equal to zero
  - Its mean value is approximately equal to 1

ANOVA – the sampling distribution for F



ANOVA – reject region

Example : an F distribution with $df_N=15$, $df_d=21$



One-Way ANOVA F-Test Example

- As a PhD student in animal and food science you want to see if 3 food supplements have different mean milk yields. You assign 15 cows, 5 per food supplement (CRD)
- Question: At the .05 level, is there a difference in mean yields?

Food1	Food2	Food3
25.4	23.4	20
26.31	21.8	22.2
24.1	23.5	19.75
23.74	22.75	20.6
25.1	21.6	20.4

One-Way ANOVA F-Test Solution

$\alpha = 0{,}05$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	236·8	238·9	240·5	241·9	243·9	245·9	248·0	249·1	250·1	251·1	252·2	253·3	254·3
2	18·51	19·00	19 16	19·25	19 30	19 33	19·35	19·37	19·38	19 40	19·41	19·43	19·45	19 45	19·46	19·47	19·48	19 49	19·50
3	10·13	9·55	9·28	9·12	9 01	8·94	8 89	8·85	8 81	8·79	8·74	8·70	8·66	8 64	8·62	8·59	8·57	8·55	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6 09	6·04	6 00	5·96	5 91	5·86	5·80	5·77	5 75	5·72	5·69	5·66	5 63
5	6·61	5·79	5 41	5·19	5·05	4·95	4·88	4·82	4·77	4·74	4·68	4·62	4 56	4·53	4·50	4·46	4·43	4·40	4·36
6	5·99	5·14	4 76	4 53	4·39	4·28	4·21	4·15	4·10	4·06	4 00	3 94	3·87	3·84	3·81	3·77	3 74	3·70	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·79	3·73	3·68	3·64	3·57	3·51	3 44	3 41	3·38	3·34	3·30	3 27	3·23
8	5·32	4·46	4·07	3 84	3 69	3 58	3 50	3·44	3·39	3 35	3 28	3·22	3·15	3·12	3 08	3·04	3·01	2 97	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·29	3·23	3·18	3·14	3 07	3 01	2·94	2·90	2 86	2 83	2 79	2 75	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·14	3 07	3 02	2 98	2 91	2·85	2·77	2·74	2·70	2·66	2·62	2·58	2·54
11	4 84	3·98	3 59	3 36	3 20	3·09	3 01	2 95	2·90	2 85	2 79	2·72	2·65	2 61	2·57	2·53	2 49	2·45	2·40
12	4 75	3·89	3 49	3·26	3 11	3 00	2 91	2·85	2 80	2 75	2 69	2 62	2 54	2 51	2·47	2 43	2 38	2·34	2·30
13	4 67	3·81	3·41	3 18	3·03	2·92	2·83	2·77	2·71	2·67	2·60	2·53	2·46	2·42	2 38	2·34	2·30	2·25	2·21
14	4 60	3 74	3·34	3·11	2·96	2·85	2·76	2·70	2·65	2·60	2 53	2·46	2 39	2 35	2·31	2·27	2·22	2·18	2 13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·71	2·64	2·59	2·54	2·48	2·40	2·33	2·29	2·25	2·20	2·16	2·11	2·07
16	4·49	3·63	3 24	3·01	2·85	2·74	2·66	2·59	2·54	2 49	2·42	2·35	2·28	2·24	2 19	2·15	2·11	2·06	2·01
17	4·45	3·59	3·20	2·96	2 81	2 70	2 61	2 55	2 49	2·45	2 38	2 31	2·23	2 19	2 15	2 10	2 06	2·01	1·96
18	4 41	3·55	3·16	2·93	2·77	2·66	2·58	2·51	2·46	2 41	2·34	2·27	2 19	2·15	2·11	2·06	2·02	1·97	1·92
19	4 38	3·52	3·13	2·90	2·74	2·63	2·54	2·48	2·42	2·38	2 31	2·23	2·16	2·11	2·07	2·03	1 98	1·93	1·88
20	4 35	3·49	3·10	2·87	2·71	2·60	2·51	2·45	2·39	2·35	2·28	2·20	2·12	2·08	2·04	1·99	1·95	1·90	1·84
21	4 32	3·47	3 07	2·84	2 68	2·57	2·49	2·42	2 37	2·32	2·25	2·18	2·10	2·05	2·01	1·96	1·92	1·87	1·81
22	4 30	3·44	3 05	2·82	2 66	2·55	2 46	2·40	2·34	2 30	2·23	2·15	2 07	2·03	1·98	1·94	1·89	1·84	1·78
23	4·28	3·42	3·03	2·80	2 64	2·53	2 44	2 37	2 32	2 27	2·20	2·13	2·05	2 01	1·96	1·91	1·86	1·81	1·76
24	4 26	3·40	3·01	2·78	2 62	2 51	2·42	2·36	2 30	2·25	2·18	2·11	2·03	1 98	1·94	1 89	1·84	1·79	1·73
25	4·24	3·39	2·99	2·76	2·60	2·49	2·40	2·34	2·28	2·24	2·16	2·09	2·01	1·96	1·92	1·87	1·82	1·77	1·71
26	4·23	3·37	2·98	2·74	2 59	2·47	2·39	2·32	2·27	2 22	2·15	2 07	1·99	1·95	1·90	1·85	1·80	1·75	1·69
27	4·21	3·35	2·96	2·73	2 57	2 46	2·37	2·31	2·25	2 20	2·13	2·06	1·97	1·93	1·88	1·84	1·79	1·73	1·67
28	4·20	3·34	2 95	2 71	2 56	2·45	2·36	2·29	2·24	2 19	2·12	2·04	1·96	1·91	1·87	1 82	1·77	1·71	1·65
29	4·18	3 33	2 93	2 70	2·55	2·43	2·35	2 28	2·22	2·18	2·10	2·03	1·94	1·90	1·85	1·81	1·75	1·70	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·33	2·27	2·21	2·16	2 09	2·01	1·93	1·89	1·84	1·79	1·74	1·68	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·25	2·18	2·12	2·08	2·00	1·92	1 84	1·79	1·74	·1·69	1 64	1·58	1·51
60	4·00	3 15	2·76	2·53	2 37	2·25	2·17	2·10	2·04	1 99	1·92	1 84	1·75	1·70	1·65	1·59	1·53	1 47	1·39
120	3·92	3·07	2 68	2·45	2 29	2·17	2·09	2 02	1·96	1·91	1·83	1·75	1·66	1·61	1·55	1·50	1·43	1·35	1·25
∞	3·84	3 00	2·60	2 37	2·21	2·10	2·01	1·94	1·88	1·83	1·75	1 67	1·57	1·52	1·46	1·39	1·32	1·22	1·00

ANOVA - calculations

$$MST = \frac{SST}{p-1} = \frac{\sum_{i=1}^{p} n_i \cdot (\hat{Y}_i - \bar{Y})^2}{p-1}$$

Where SST is the between groups sum of squares, p is the number of groups and n_i is the size group

$$MSE = \frac{SSE}{N-p} = \frac{\sum_{i=1}^{p} (n_i - 1) \cdot S^2{}_i}{p-1}$$

Where SSE is the within groups sum of squares (error), p is the number of groups, n_i is the size group, N is the sum of group sizes, $S^2{}_i$ is the within group variance

ANOVA - Example

Our data set: $\bar{Y}_{\text{(grand mean)}} = 22.71$

$\hat{Y}_{Food1} = 24.93$; $\hat{Y}_{Food2} = 22.61$; $\hat{Y}_{Food3} = 20.59$

$S^2_{\text{Food1}} = 1.0648$; $S^2_{\text{Food2}} = 0.778$; $S^2_{\text{Food3}} = 0.9205$

$$MST = \frac{SST}{p-1} = \frac{\sum_{i=1}^{p} n_i \cdot (\hat{Y}_i - \bar{Y})^2}{p-1} = \frac{5 \cdot (24.93-22.71)^2 + 5 \cdot (22.61-22.71)^2 + 5 \cdot (20.59-22.71)^2}{3-1} = \mathbf{23.582}$$

$$MSE = \frac{SSE}{N-p} = \frac{\sum_{i=1}^{p} (n_i - 1) \cdot S^2 i}{N-1} = \frac{(4 \cdot 1.0648) + (4 \cdot 0.778) + (4 \cdot 0.9295)}{15-3} = \mathbf{0.9211}$$

$df_N = p - 1 = 3 - 1$, $df_D = N - p = 15 - 3 = 12$

Summary Table Solution

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Food	3 - 1 = 2	47.1640	23.5820	25.60
Error	15 - 3 = 12	11.0532	0.9211	
Total	15 - 1 = 14	58.2172		

R CODES FOR ANOVA

Create Dataframe

```
foods<-data.frame(milk=c(25.40,26.31,24.10,23.74,25.10,
                         23.40,21.80,23.50,22.75,21.60,
                         20.00,22.20,19.75,20.60,20.40),
                  food=c("Food1","Food1","Food1","Food1","Food1",
                         "Food2","Food2","Food2","Food2","Food2",
                         "Food3","Food3","Food3","Food3","Food3"),
                  stringsAsFactors = TRUE)

print(foods) #See the Data
```

```
##     milk  food
## 1  25.40 Food1
## 2  26.31 Food1
## 3  24.10 Food1
## 4  23.74 Food1
## 5  25.10 Food1
## 6  23.40 Food2
## 7  21.80 Food2
## 8  23.50 Food2
## 9  22.75 Food2
## 10 21.60 Food2
## 11 20.00 Food3
## 12 22.20 Food3
## 13 19.75 Food3
## 14 20.60 Food3
## 15 20.40 Food3
```

R OUTPUT - ANOVA

```
tm<-lm(milk ~ food, data = foods)    # Fit the linear model
summary(tm)                          # Linear Model Summary
```

```
##
## Call:
## lm(formula = milk ~ food, data = foods)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.19  -0.82   0.01   0.63   1.61
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.9300     0.4292  58.084 4.48e-16 ***
## foodFood2    -2.3200     0.6070  -3.822  0.00243 **
## foodFood3    -4.3400     0.6070  -7.150 1.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9597 on 12 degrees of freedom
## Multiple R-squared:  0.8101, Adjusted R-squared:  0.7785
## F-statistic:  25.6 on 2 and 12 DF,  p-value: 4.684e-05
```

```
fm<-aov(milk ~ food, data = foods)   # Fit the ANOVA
summary(fm)                          # ANOVA table SS I
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## food         2  47.16  23.582    25.6 4.68e-05 ***
## Residuals   12  11.05   0.921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10% level of confidence is termed a "tendency" in the scientific community

5% level of confidence is considered much better for the scientific community

SAS CODES FOR ANOVA

```
Data Anova;
input group$ milk @@;
cards;
food1 25.40     food2 23.40     food3 20.00
food1 26.31     food2 21.80     food3 22.20
food1 24.10     food2 23.50     food3 19.75
food1 23.74     food2 22.75     food3 20.60
food1 25.10     food2 21.60     food3 20.40
;
run;

proc anova; /* or PROC GLM */
class group;
model milk=group;
run;
```

SAS OUTPUT - ANOVA

La procedura GLM

Variabile dipendente: milk

Origine	DF	Somma dei quadrati	Media quadratica	Valore F	Pr > F
Modello	2	47.16400000	23.58200000	25.60	<.0001
Errore	12	11.05320000	0.92110000		
Totale corretto	14	58.21720000			

R-quadro	Var coeff	Radice MSE	milk Media
0.810139	4.226066	0.959740	22.71000

Origine	DF	Tipo I SS	Media quadratica	Valore F	Pr > F
group	2	47.16400000	23.58200000	25.60	<.0001

Origine	DF	Tipo III SS	Media quadratica	Valore F	Pr > F
group	2	47.16400000	23.58200000	25.60	<.0001

Pair-wise comparison

- Needed when the overall F test is rejected
- Can be done without adjustment of type I error if other comparisons were planned in advance (least significant difference - LSD method)
- Type I error needs to be adjusted if other comparisons were not planned in advance (Bonferroni's and scheffe's methods)

Latin square experimental design…. what is it? This design does not need a post-hoc analysis (after the ANOVA).

Fisher's Least Significant Difference (LSD) Test

To compare level 1 and level 2

$$t = (\bar{y}_1 - \bar{y}_2) \Big/ \sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Compare this to $t_{\alpha/2}$ = Upper-tailed value or $-t_{\alpha/2}$ *lower*-tailed from Student's t-distribution for $\alpha/2$ and (n - p) degrees of freedom

MSE = Mean square within from ANOVA table

n = Number of subjects

p = Number of levels

Bonferroni's method

To compare level 1 and level 2

$$t = (\bar{y}_1 - \bar{y}_2) \Big/ \sqrt{MSE\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Adjust the significance level α by taking the new significance level α*

$$\alpha^* = \alpha / \binom{p}{2}$$

R CODES FOR multiple comparisons

```
TukeyHSD(fm)                    # Tukey test for multiple comparisons
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = milk ~ food, data = foods)
##
## $food
##              diff       lwr        upr     p adj
## Food2-Food1 -2.32 -3.939373 -0.7006265 0.0063517
## Food3-Food1 -4.34 -5.959373 -2.7206265 0.0000322
## Food3-Food2 -2.02 -3.639373 -0.4006265 0.0153900
```

Food2 and Food2 have significant difference (low p-value)

Food3 and Food1 have significant difference (low p-value)

Food3 and Food2 have significant difference, but their difference is weaker than the other differences

```
plot(TukeyHSD(fm))              # Plot for tukey test
aggregate(foods$milk, by=list(foods$food), FUN = mean) # Means by group
```

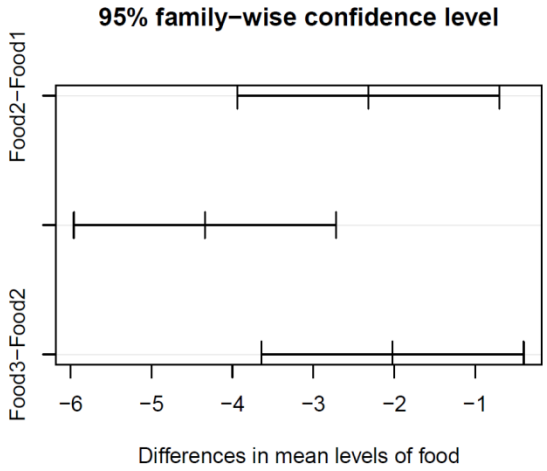
```
##   Group.1     x
## 1   Food1 24.93
## 2   Food2 22.61
## 3   Food3 20.59
```

SAS OUTPUT - Tukey

```
TukeyHSD(fm)                    # Tukey test for multiple comparisons
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = milk ~ food, data = foods)
##
## $food
##               diff       lwr        upr     p adj
## Food2-Food1 -2.32 -3.939373 -0.7006265 0.0063517
## Food3-Food1 -4.34 -5.959373 -2.7206265 0.0000322
## Food3-Food2 -2.02 -3.639373 -0.4006265 0.0153900
```

```
plot(TukeyHSD(fm))              # Plot for tukey test
```

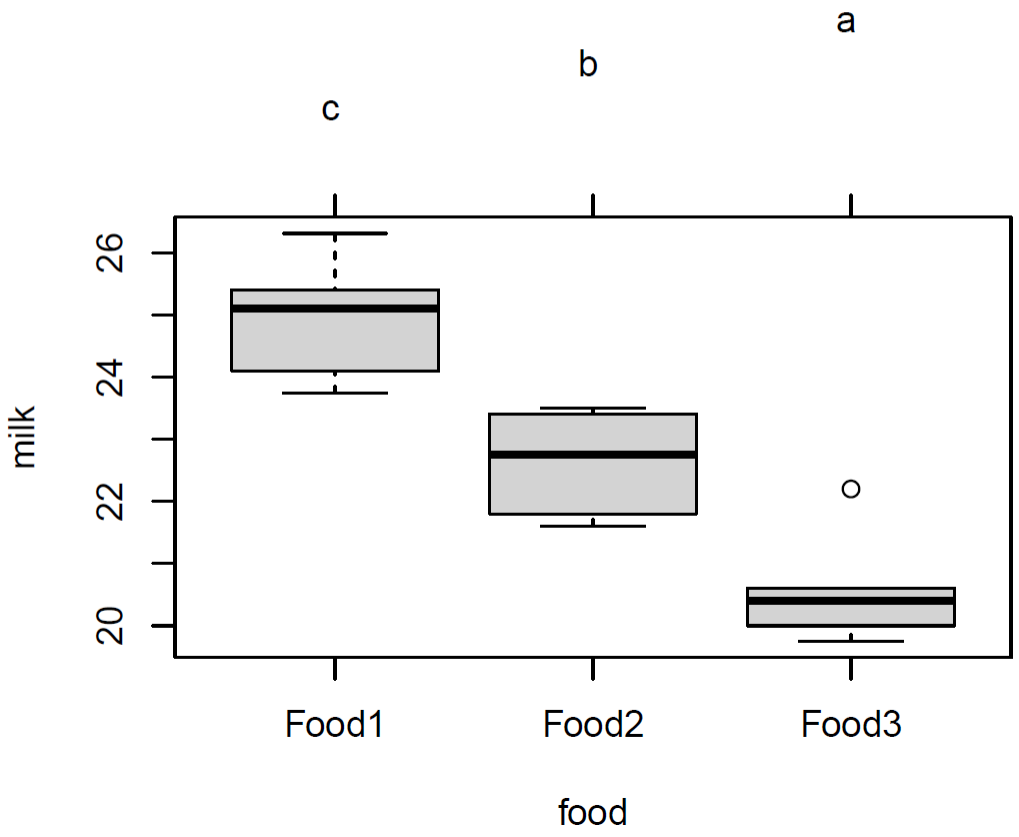



```
aggregate(foods$milk, by=list(foods$food), FUN = mean) # Means by group
```

```
##   Group.1     x
## 1   Food1 24.93
## 2   Food2 22.61
## 3   Food3 20.59
```

R Plot– Tukey differences

```r
#install.packages("multcomp")
library(multcomp)
par(mar=c(4,4,6,2))           # Change parameters for the plot margins
tuk <- glht(fm, linfct=mcp(food="Tukey")) # Fit the general Linear Hypotheses
plot(cld(tuk, level=0.05),col="lightgrey")   # Plot the mean differences
```



SAS CODES FOR multiple comparisons

```
proc glm;
class group;
model milk=group;
means group/ lsd bon;
run;
```

SAS OUTPUT - LSD

SAS System 16:10 Sunday, January 31, 2010 7

La procedura GLM

Test t (LSD) per milk

NOTA: Questo test controlla il tasso di errore di confronto di tipo I, non il tasso di errore sperimentale.

Alfa	0.05
Gradi di libertà per l'errore	12
Errore quadratico medio	0.9211
Valore critico di t	2.17881
Minore differenza significativa	1.3225

Le medie con la stessa lettera non sono significativamente diverse.

t Raggruppamento	Media	N	group
A	24.9300	5	food1
B	22.6100	5	food2
C	20.5900	5	food3

SAS OUTPUT - Bonferroni

SAS System 16:10 Sunday, January 31, 2010 8

La procedura GLM

Test t di Bonferroni (Dunn) per milk

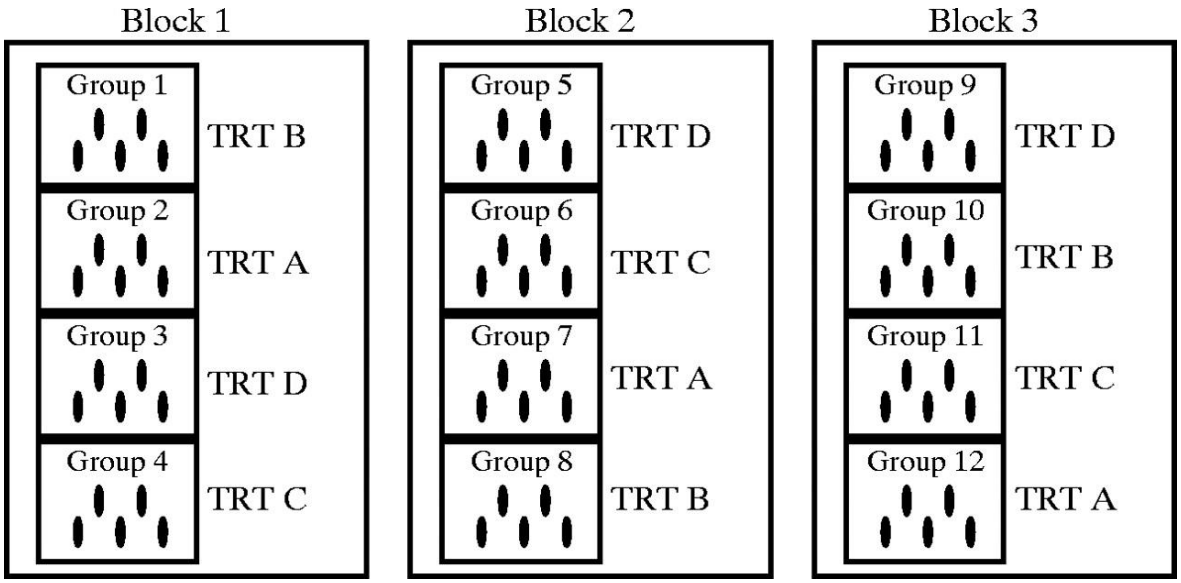
NOTA: Questo test controlla il tasso di errore sperimentale di tipo I, ma generalmente ha un tasso di errore di tipo II maggiore rispetto a REGWQ.

Alfa	0.05
Gradi di libertà per l'errore	12
Errore quadratico medio	0.9211
Valore critico di t	2.77947
Minima differenza significativa	1.6871

Le medie con la stessa lettera non sono significativamente diverse.

Bon Raggruppamento	Media	N	group
A	24.9300	5	food1
B	22.6100	5	food2
C	20.5900	5	food3

Randomized Block Design



Randomized Block Design

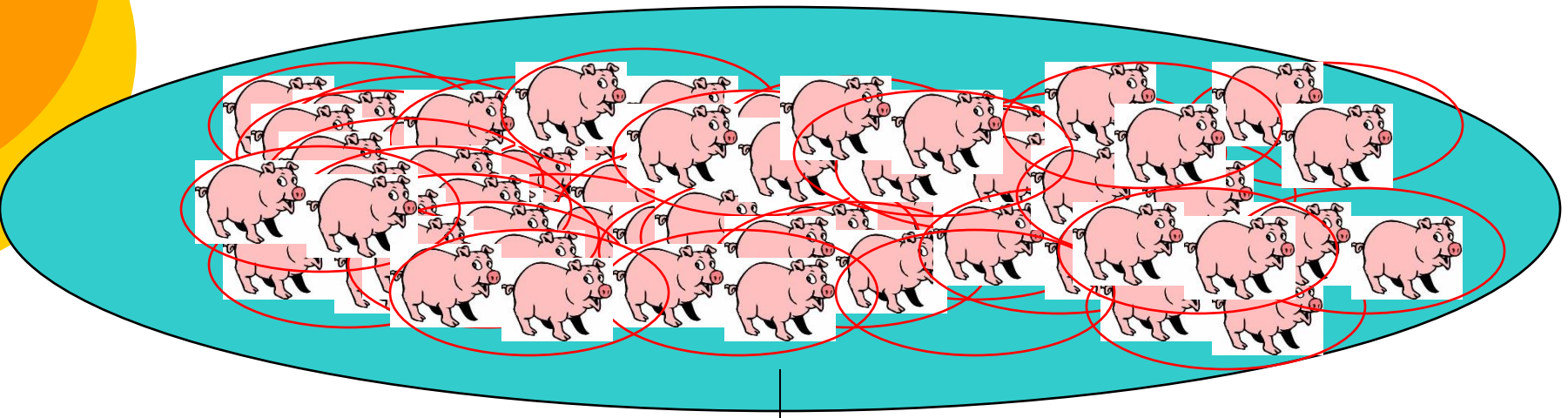
- Experimental Units (Subjects) Are Assigned Randomly within Blocks
  - Blocks are Assumed Homogeneous
- One Factor or Independent Variable of Interest
  - 2 or More Treatment Levels or Classifications
- One Blocking Factor

Randomized Block Design

Factor Levels: (Treatments)	A, B, C, D			
Experimental Units	Treatments are randomly assigned within blocks			
Block 1	A	C	D	B
Block 2	C	D	B	A
Block 3	B	A	D	C
. . .	. . .	. . .	. . .	. . .
Block b	D	C	A	B

The Randomized Complete Block Design (RCBD)

Population of litters of size 2



Draw random sample of litters

Litter 1 | Litter 2 | … | Litter *n*

Randomly assign treatments to piglets within litters

Randomized Block F-Test

- Tests the Equality of 2 or More (p) Population Means

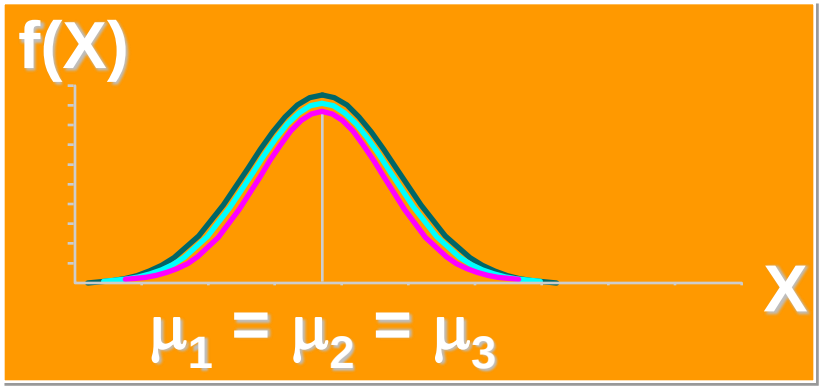
- Variables
  - One Nominal Independent Variable
  - One Nominal Blocking Variable
  - One Continuous Dependent Variable

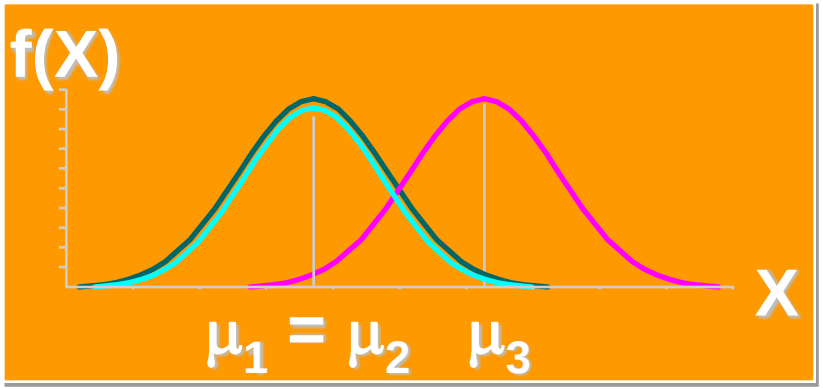
Randomized Block F-Test Assumptions

- Normality
  - Probability distribution of each block-treatment combination is normal
- Homogeneity of variance
  - Probability distributions of all block-treatment combinations have equal variances

Randomized Block F-Test Hypotheses

- H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_p$
  - All Population Means are Equal
  - No Treatment Effect
- H_a: Not All μ_j Are Equal
  - At Least 1 Pop. Mean is Different
  - Treatment Effect
  - $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_p$ Is wrong





Two-Way ANOVA - Summary Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Treatment (a)	a - 1	SS(A)	MS(A)	$\frac{MS(A)}{MSE}$
Blocks (b)	b - 1	SS(B)	MS(B)	$\frac{MS(B)}{MSE}$
Error	(a-1)(b-1)	SS(E)	MS(E)	
Total	N-1	SS(T)		

Randomized Block F-Test
Test Statistic

- Test Statistic
  - $F = MST / MSE$
    - MST Is Mean Square for Treatment
    - MSE Is Mean Square for Error
- Degrees of Freedom
  - $df_N = p - 1$
  - $df_D = n - b - p + 1$ or $(p-1)(b-1)$
    - p = # Treatments, b = # Blocks, N = Total Sample Size

Example: 15 piglets; weaned in 5 litters; treated with 3 treatments

14 degrees of freedom due to the piglets

4 degrees of freedom due to the litters

2 degrees of freedom due to the treatments

EXAMPLE and SAS CODE

Litter	T1	T2	T3
1	7.86	7.76	7.46
2	8.00	7.73	7.68
3	7.93	7.74	7.51
4	7.62	7.43	7.21
5	7.81	7.44	7.42

Alternatives: PROC ANOVA, GLM, MIXED, ANALYST, INTERACTIVE DATA ANALYSIS

```
data rcbd;
input trt $ litter y @@;
cards;
T1 1 7.86 T2 1 7.76 T3 1 7.46
T1 2 8.00 T2 2 7.73 T3 2 7.68
T1 3 7.93 T2 3 7.74 T3 3 7.51
T1 4 7.62 T2 4 7.43 T3 4 7.21
T1 5 7.81 T2 5 7.44 T3 5 7.42
;
run;

proc print data=rcbd;
run;

proc anova data=rcbd;
class litter trt;
model y = litter trt;
means trt;
run;

proc glm data=rcbd;
class litter trt;
model y = litter trt;
lsmeans trt / stderr;
run;

proc mixed data=rcbd;
class litter trt;
model y = trt;
random litter;
lsmeans trt;
run;
```

RESULTS (PROC ANOVA)

(Test for Fixed Effects)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	0.65649333	0.10941556	24.51	<.0001
Error	8	0.03570667	0.00446333		
Corrected Total	14	0.69220000			

at least one factor is significant…. but we don't know which one

0.10941556 / 0.00446333 = 24.51

R-Square	Coeff Var	Root MSE	y Mean
0.948416	0.874453	0.066808	7.640000

0.06928333 / 0.00446333 = 15.52

Source	DF	Anova SS	Mean Square	F Value	Pr > F
litter	4	0.27713333	0.06928333	15.52	0.0008
trt	2	0.37936000	0.18968000	42.50	<.0001

0.18968 / 0.00446333 = 42.50

RESULTS (PROC GLM)

(Test for Fixed Effects)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	0.65649333	0.10941556	24.51	<.0001
Error	8	0.03570667	0.00446333		
Corrected Total	14	0.69220000			

R-Square	Coeff Var	Root MSE	y Mean
0.948416	0.874453	0.066808	7.640000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
litter	4	0.27713333	0.06928333	15.52	0.0008
trt	2	0.37936000	0.18968000	42.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
litter	4	0.27713333	0.06928333	15.52	0.0008
trt	2	0.37936000	0.18968000	42.50	<.0001

EXAMPLE and R CODE

Litter	T1	T2	T3
1	7.86	7.76	7.46
2	8.00	7.73	7.68
3	7.93	7.74	7.51
4	7.62	7.43	7.21
5	7.81	7.44	7.42

```
pigs<-data.frame(trt=c("T1","T1","T1","T1","T1",
                       "T2","T2","T2","T2","T2",
                       "T3","T3","T3","T3","T3"),
                 litter=c("1","2","3","4","5",
                          "1","2","3","4","5",
                          "1","2","3","4","5"),
                 y=c(7.86,8.00,7.93,7.62,7.81,
                     7.76,7.73,7.74,7.43,7.44,
                     7.46,7.68,7.51,7.21,7.42),
                 stringsAsFactors = TRUE)
print(pigs) #See the Data
```

(Test for Fixed Effects)

```
tm<-lm(y ~  trt + litter, data = pigs) # Fit the linear model
summary(tm)                            # Linear Model Summary
```

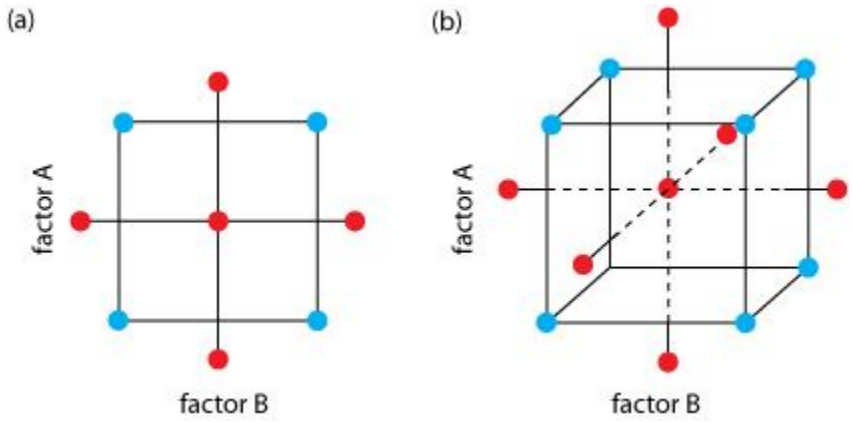
```
##
## Call:
## lm(formula = y ~ trt + litter, data = pigs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09667 -0.03500 -0.00400  0.04033  0.08667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.89733    0.04564 173.040 1.39e-15 ***
## trtT2       -0.22400    0.04225  -5.301 0.000727 ***
## trtT3       -0.38800    0.04225  -9.183 1.60e-05 ***
## litter2      0.11000    0.05455   2.017 0.078477 .
## litter3      0.03333    0.05455   0.611 0.558109
## litter4     -0.27333    0.05455  -5.011 0.001039 **
## litter5     -0.13667    0.05455  -2.505 0.036632 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06681 on 8 degrees of freedom
## Multiple R-squared:  0.9484, Adjusted R-squared:  0.9097
## F-statistic: 24.51 on 6 and 8 DF,  p-value: 9.763e-05
```

```
fm<-aov(y ~ trt + litter, data = pigs) # Fit the ANOVA
summary(fm)                            # ANOVA Table SS I
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## trt          2 0.3794 0.18968   42.50 5.48e-05 ***
## litter       4 0.2771 0.06928   15.52 0.000771 ***
## Residuals    8 0.0357 0.00446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

residuals degrees of freedom should be as large as possible to have robust inference

Factorial Design



Factorial Design

- Experimental units (subjects) are assigned randomly to treatments
  - Subjects are assumed homogeneous
- Two or more **factors** or independent variables
  - Each has 2 or more treatments (levels)
- Analyzed by two-way ANOVA

Advantages of Factorial Designs

- Saves time & effort
  - e.g., Could use separate completely randomized designs for each variable
- Controls confounding effects by putting other variables into model
- Can explore interaction between variables

Two-Way ANOVA

- Tests the equality of 2 or more population means when several independent variables are used

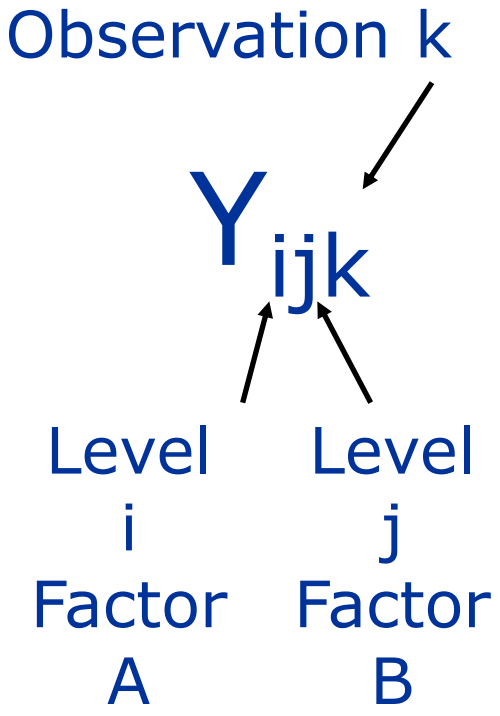
- Same results as separate one-way anova on each variable
  - But interaction can be tested

Two-Way ANOVA Assumptions

- Normality
  - Populations are Normally Distributed
- Homogeneity of Variance
  - Populations have Equal Variances
- Independence of Errors
  - Independent Random Samples are Drawn

Two-Way ANOVA Data Table

Factor A	Factor B 1	2	...	b
1	Y_{111}	Y_{121}	...	Y_{1b1}
	Y_{112}	Y_{122}	...	Y_{1b2}
2	Y_{211}	Y_{221}	...	Y_{2b1}
	Y_{212}	Y_{222}	...	Y_{2b2}
:	:	:	:	:
a	Y_{a11}	Y_{a21}	...	Y_{ab1}
	Y_{a12}	Y_{a22}	...	Y_{ab2}



rule of thumb: for every fixed i and j there should be at least 2 measurements of the response variable Y

The two factor factorial Experiment

The model for a factorial experiment with two factors A and B is:

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk} \qquad i = 1,\ldots,a;\ j = 1\ldots,b;\ k = 1,\ldots,n$$

Where:

y_{ijk} = observation k in level i of factor A and level j of factor B

μ = the overall mean

A_i = the effect of level i of factor A

B_j = the effect of level j of factor B

$(AB)_{ij}$ = the effect of the interaction of level i of factor A with level j of factor B

ε_{ijk} = random error with mean 0 and variance σ^2

a = number of levels of factor A; b =number of levels of factor B; n =number of observation for each A x B combination

Two-Way ANOVA - Null Hypotheses

No Difference in Means Due to Factor A

- $H_0: \mu_{1.} = \mu_{2.} = \ldots = \mu_{a.}$

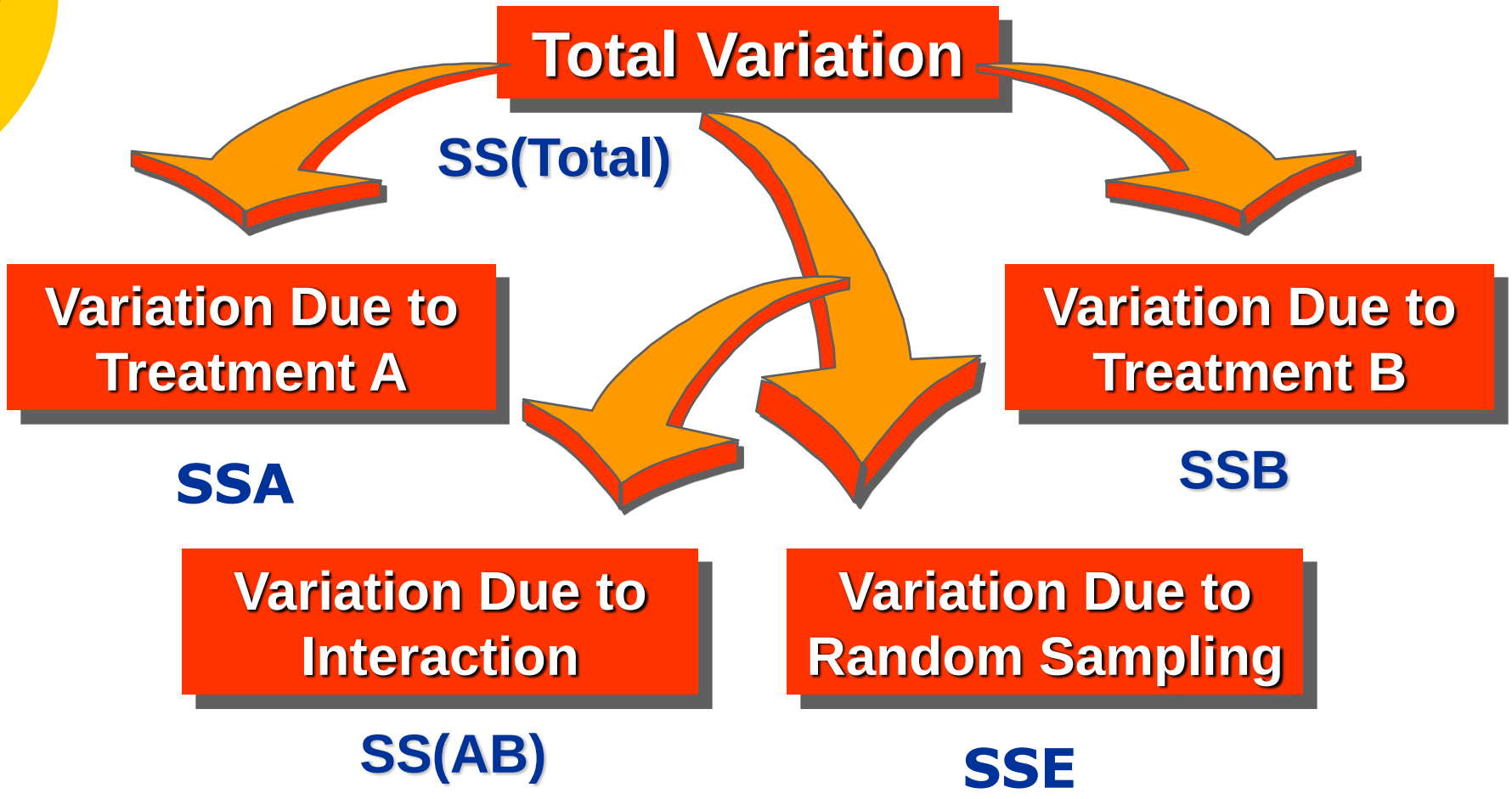
No Difference in Means Due to Factor B

- $H_0: \mu_{.1} = \mu_{.2} = \ldots = \mu_{.b}$

No Interaction of Factors A & B

- $H_0: AB_{ij} = 0$

Two-Way ANOVA - Total Variation Partitioning



Two-Way ANOVA - Summary Table

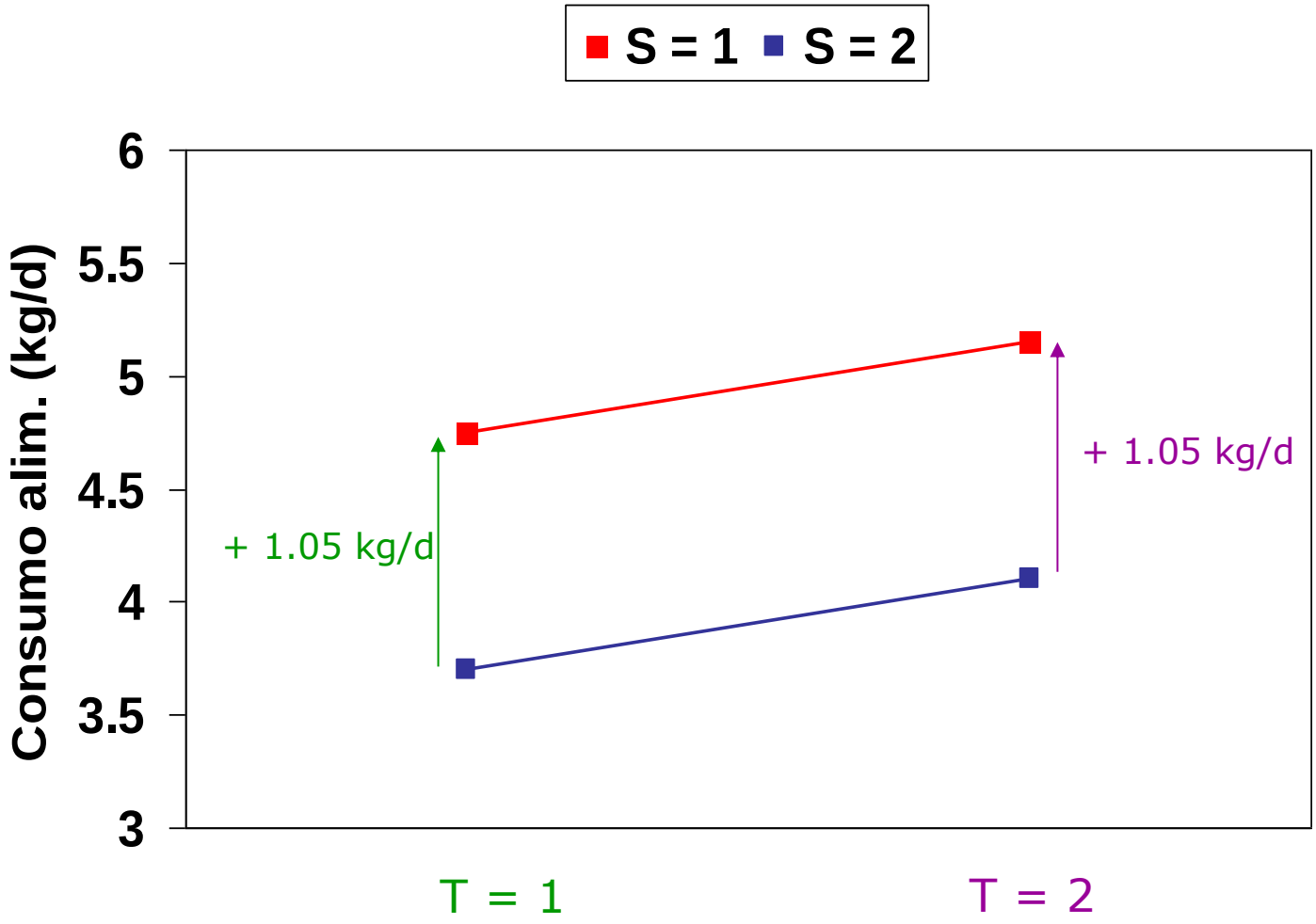
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
A (Row)	a - 1	SS(A)	MS(A)	$\frac{MS(A)}{MSE}$
B (Column)	b - 1	SS(B)	MS(B)	$\frac{MS(B)}{MSE}$
AB (Interaction)	(a-1)(b-1)	SS(AB)	MS(AB)	$\frac{MS(AB)}{MSE}$
Error	ab(n-1)	SSE	MSE	
Total	abn-1	SS(Total)		

← **Same as Other Designs**

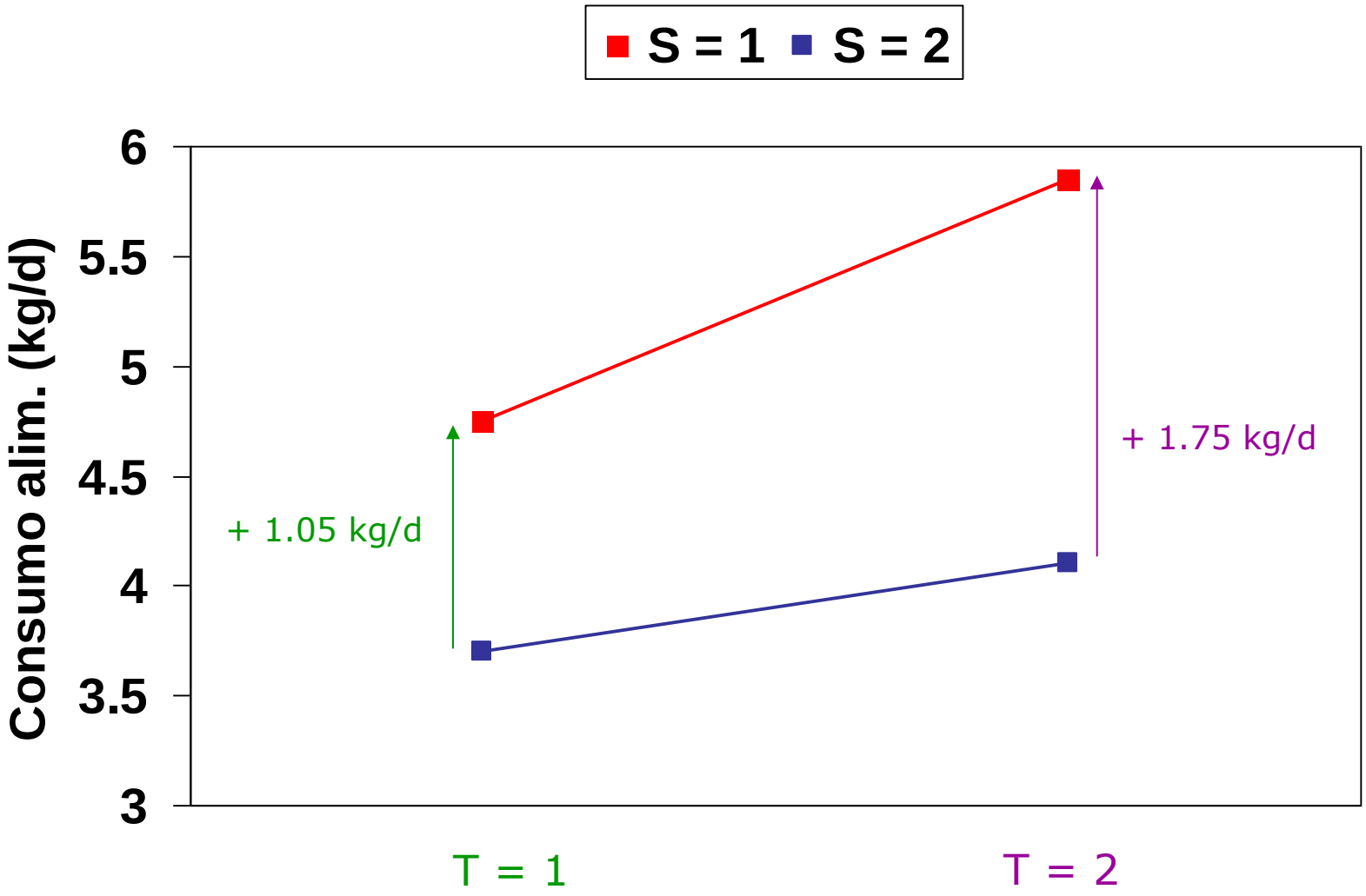
Interaction

- Occurs when effects of one factor vary according to levels of other factor
- When significant, interpretation of main effects (A & B) is complicated
- Can be detected
  - In data table, pattern of cell means in one row differs from another row
  - In graph of cell means, lines cross

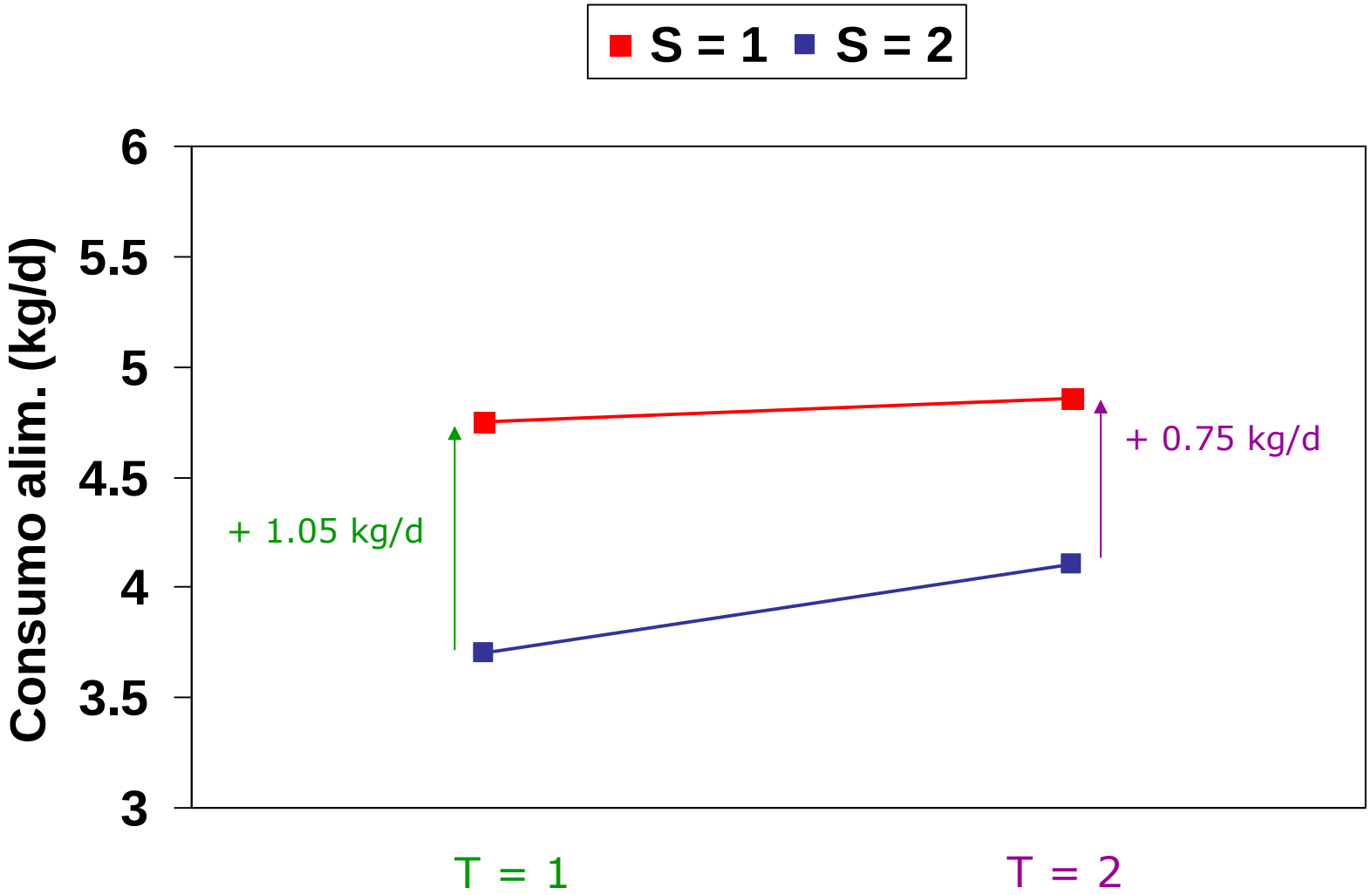
Example: graphs of interaction – **no** interaction



Example: graphs of Interaction – **positive** interaction



Example: graphs of Interaction – **negative** interaction



Two-Way ANOVA F - Test Example

Vitamin I	0 mg		4 mg	
Vitamin II	0 mg	5mg	0 mg	5 mg
	0.585	0.567	0.473	0.684
	0.536	0.545	0.450	0.702
	0.458	0.589	0.869	0.900
	0.486	0.536	0.473	0.698
	0.536	0.549	0.464	0.693

```
data gain;
input vitI vitII gain @@;
datalines;
1 1 0.585 2 1 0.473
1 1 0.536 2 1 0.450
1 1 0.458 2 1 0.869
1 1 0.486 2 1 0.473
1 1 0.536 2 1 0.464
1 2 0.567 2 2 0.684
1 2 0.545 2 2 0.702
1 2 0.589 2 2 0.900
1 2 0.536 2 2 0.698
1 2 0.549 2 2 0.693
;

proc glm;
class vitI vitII;
model gain=vitI vitII vitI*vitII;
lsmeans vitI*vitII/ tdiff pdiff stderr adjust=tukey;
run;
```

SAS OUTPUT - ANOVA

La procedura GLM

Variabile dipendente: gain

Origine	DF	Somma dei quadrati	Media quadratica	Valore F	Pr > F
Modello	3	0.14521095	0.04840365	4.39	0.0196
Errore	16	0.17648360	0.01103023		
Totale corretto	19	0.32169455			

R-quadro	Var coeff	Radice MSE	gain Media
0.451394	17.81139	0.105025	0.589650

Origine	DF	Tipo I SS	Media quadratica	Valore F	Pr > F
vitl	1	0.05191805	0.05191805	4.71	0.0454
vitll	1	0.06418445	0.06418445	5.82	0.0282
vitl*vitll	1	0.02910845	0.02910845	2.64	0.1238

Origine	DF	Tipo III SS	Media quadratica	Valore F	Pr > F
vitl	1	0.05191805	0.05191805	4.71	0.0454
vitll	1	0.06418445	0.06418445	5.82	0.0282
vitl*vitll	1	0.02910845	0.02910845	2.64	0.1238

SAS OUTPUT - ANOVA

La procedura GLM
Medie dei minimi quadrati
Correzione per i confronti multipli: Tukey

vit I	vit II	gain LSMEAN	Errore standard	Pr > |t|	Numero LSMEAN
1	1	0.52020000	0.04696855	<.0001	1
1	2	0.55720000	0.04696855	<.0001	2
2	1	0.54580000	0.04696855	<.0001	3
2	2	0.73540000	0.04696855	<.0001	4

Medie dei minimi quadrati per l'effetto vitI*vitII
t per H0: LSMean(i)=LSMean(j) / Pr > |t|

Variabile dipendente: gain

i/j	1	2	3	4
1		-0.55703 0.9433	-0.38541 0.9799	-3.23981 0.0238
2	0.557031 0.9433		0.171626 0.9981	-2.68278 0.0701
3	0.385405 0.9799	-0.17163 0.9981		-2.85441 0.0506
4	3.239814 0.0238	2.682783 0.0701	2.854409 0.0506	

Two-Way ANOVA F - Test Example

Create Dataframe

```
gain<-data.frame(vitI=c("1","1","1","1","1",
                        "1","1","1","1","1",
                        "2","2","2","2","2",
                        "2","2","2","2","2"),
                 vitII=c("1","1","1","1","1",
                         "2","2","2","2","2",
                         "1","1","1","1","1",
                         "2","2","2","2","2"),
                 gain=c(0.585,0.536,0.458,0.486,0.536,
                        0.567,0.545,0.589,0.536,0.549,
                        0.473,0.450,0.869,0.473,0.464,
                        0.684,0.702,0.900,0.698,0.693),
                 stringsAsFactors = TRUE)

print(gain) #See the Data
```

20 observations means 19 degrees of freedom

Vitamin I	0 mg		4 mg	
Vitamin II	0 mg	5mg	0 mg	5 mg
	0.585	0.567	0.473	0.684
	0.536	0.545	0.450	0.702
	0.458	0.589	0.869	0.900
	0.486	0.536	0.473	0.698
	0.536	0.549	0.464	0.693

0.585 means that the cow grows approximately 500 grams per day

R OUTPUT - ANOVA

```
tm<-lm(gain ~ vitI + vitII + vitI*vitII, data = gain) # Fit the linear model
summary(tm)                                  # Linear Model Summary
```

```
##
## Call:
## lm(formula = gain ~ vitI + vitII + vitI * vitII, data = gain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.0958 -0.0541 -0.0273  0.0158  0.3232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52020    0.04697  11.075 6.51e-09 ***
## vitI2        0.02560    0.06642   0.385    0.705
## vitII2       0.03700    0.06642   0.557    0.585
## vitI2:vitII2 0.15260    0.09394   1.624    0.124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.105 on 16 degrees of freedom
## Multiple R-squared:  0.4514, Adjusted R-squared:  0.3485
## F-statistic: 4.388 on 3 and 16 DF,  p-value: 0.01958
```

Reject null hypothesis: there is at least one significant factor

```
fm<-aov(gain ~ vitI + vitII + vitI*vitII, data = gain) # Fit the ANOVA
summary(fm)                                  # ANOVA Table
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## vitI         1 0.05192 0.05192   4.707 0.0454 *
## vitII        1 0.06418 0.06418   5.819 0.0282 *
## vitI:vitII   1 0.02911 0.02911   2.639 0.1238
## Residuals   16 0.17648 0.01103
## ---
```

The interaction is not significant because the p-value is higher than 5%

R OUTPUT - LSM

```
library(lsmeans)
lsmeans(tm,"vitI")   #LSM for VitI
```

```
##  vitI lsmean         SE df  lower.CL  upper.CL
##  1    0.5387 0.03321178 16 0.4682942 0.6091058
##  2    0.6406 0.03321178 16 0.5701942 0.7110058
##
## Results are averaged over the levels of: vitII
## Confidence level used: 0.95
```

```
lsmeans(tm,"vitII")  #LSM for VitII
```

```
##  vitII lsmean         SE df  lower.CL  upper.CL
##  1     0.5330 0.03321178 16 0.4625942 0.6034058
##  2     0.6463 0.03321178 16 0.5758942 0.7167058
##
## Results are averaged over the levels of: vitI
## Confidence level used: 0.95
```

```
summary(ref.grid(tm))#This is the reference grid of the model
```

```
##  vitI vitII prediction        SE df
##  1    1         0.5202 0.04696855 16
##  2    1         0.5458 0.04696855 16
##  1    2         0.5572 0.04696855 16
##  2    2         0.7354 0.04696855 16
```

These numbers are equal because for each combination of levels of the factors there is the same number of measurements of the response variable (balanced experimental design)

```
aggregate(gain$gain, by=list(gain$vitI,gain$vitII), FUN = mean) # Means by group
```

```
##   Group.1 Group.2      x
## 1       1       1 0.5202
## 2       2       1 0.5458
## 3       1       2 0.5572
## 4       2       2 0.7354
```

Multiple Comparisons

```
TukeyHSD(fm)          # Tukey test for multiple comparisons
```

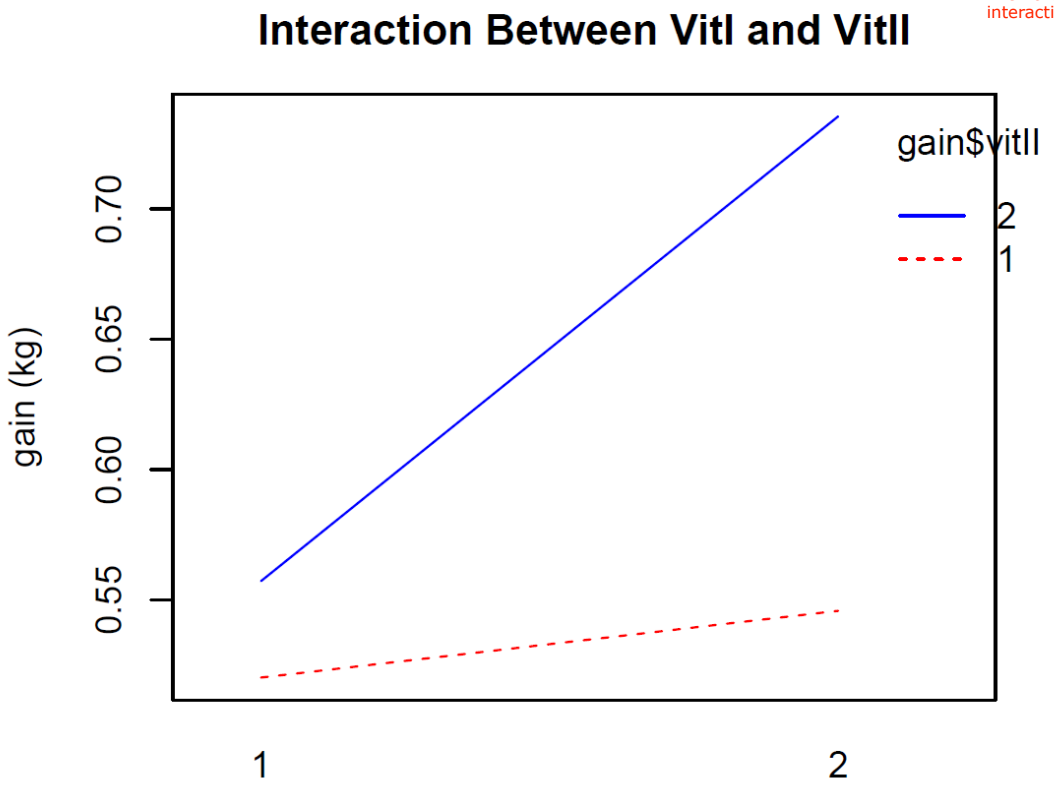
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = gain ~ vitI + vitII + vitI * vitII, data = gain)
##
## $vitI
##       diff         lwr       upr     p adj
## 2-1 0.1019 0.002331117 0.2014689 0.0454479
##
## $vitII
##       diff        lwr       upr     p adj
## 2-1 0.1133 0.01373112 0.2128689 0.0282222
##
## $`vitI:vitII`
##           diff           lwr       upr     p adj
## 2:1-1:1 0.0256 -0.1644391365 0.2156391 0.9798534
## 1:2-1:1 0.0370 -0.1530391365 0.2270391 0.9432622
## 2:2-1:1 0.2152  0.0251608635 0.4052391 0.0238280
## 1:2-2:1 0.0114 -0.1786391365 0.2014391 0.9981254
## 2:2-2:1 0.1896 -0.0004391365 0.3796391 0.0506375
## 2:2-1:2 0.1782 -0.0118391365 0.3682391 0.0700704
```

```
plot(TukeyHSD(fm))    # Plot for tukey test
```

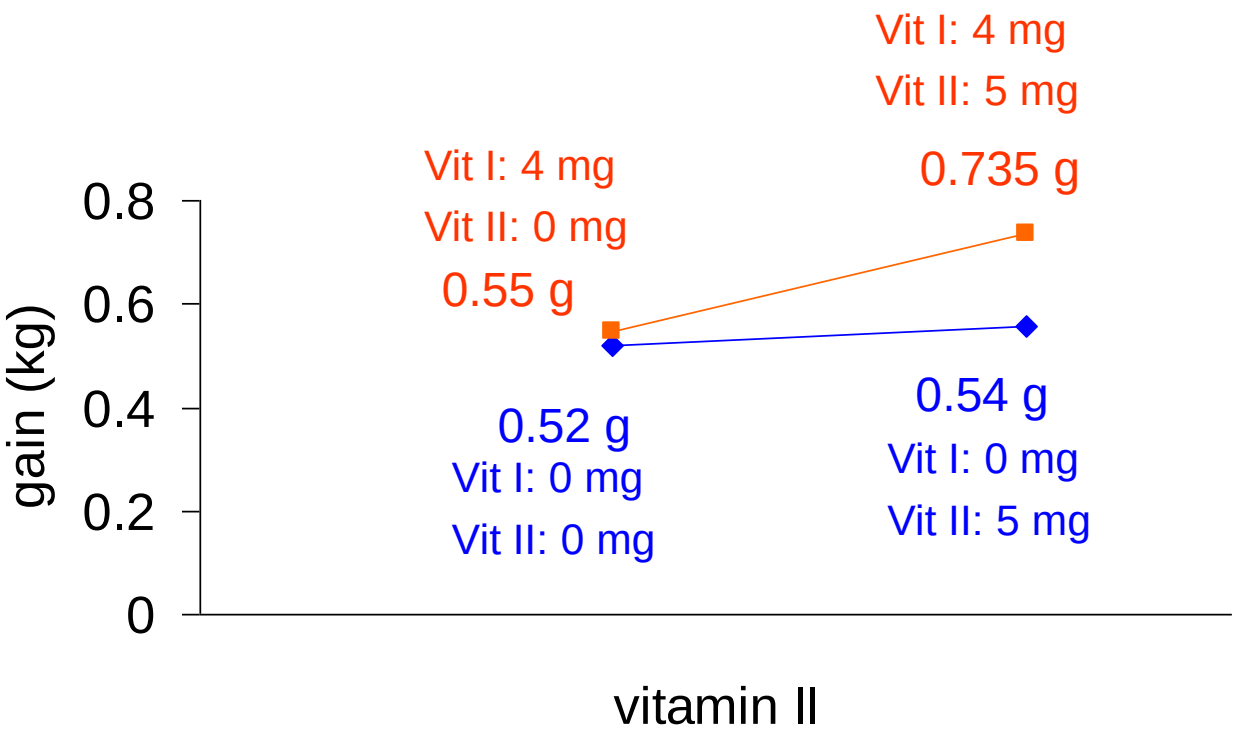
Interaction Plot

```
par(mar=c(3,4,3,2)) # Change parameters for the plot margins
interaction.plot(gain$vitI,gain$vitII,response = gain$gain,
                 col=c("red","blue"),pch = c(16,18),
                 main="Interaction Between VitI and VitII",
                 ylab = "gain (kg)")
```

Even the F p-values would not suggest existence of significant interactions, these plots indicate otherwise.



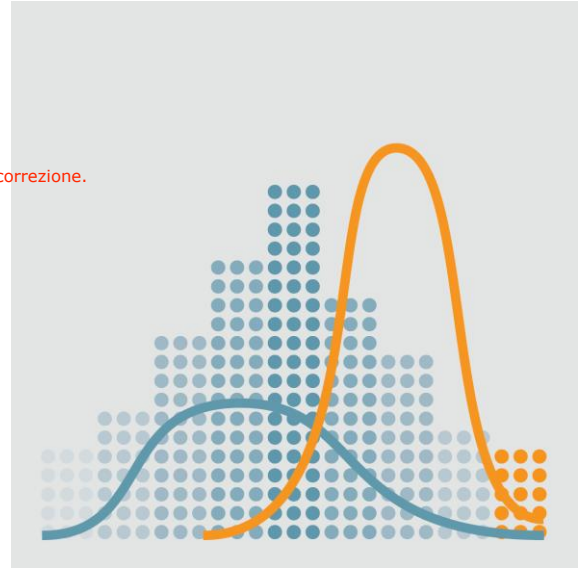
Interaction



According to the figure interaction possibly exists, but probably the power is not enough to detect it. Most likely more than 5 measurement per group are needed

Analysis of covariance

Caso possibile: si misura il avg mas gain di tori
studiando due diversi tipi di mangime ma lo si fa
con tori giovani che hanno una grande variabilità nel peso iniziale.
Allora la misurazione della crescita di peso e la previsione del peso finale
dopo un tempo prefissato utilizzando un particolare mangime
potrebbe essere viziato dalla variabilità insita nel peso
iniziale dei tori, che appunto da giovani hanno pesi tipicamente
diversi tra loro anche a parità di età. In questo caso quindi si applica una correzione.



Analysis of covariance

- Statistical procedure in which variability of a dependent variable is explained by both categorical and **continuous independent variables (covariate)**
- Common application of analysis of covariance is to adjust treatment means for known source of variability that can be explained by a continuous variable

Completely Randomized Design with a Covariate

In a CRD with a covariate the analysis of covariance is utilized for correcting treatment means, controlling the experimental error, and increase precision

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \tau_i + \varepsilon_{ij} \qquad i = 1,...,a;\ j = 1...,n$$

Where:

y_{ijk} = observation *j* in group *i* (treatment *i*)

β_0 = the intercept

x_{ij} = the regression coefficient

β_1 = a continuous independent variable with mean μ_x (covariate)

τ_i = the fixed effect of group or treatment *i*

ε_{ijk} = random error with mean 0 and variance σ^2

Assumptions

- Covariate is fixed and independent of treatments
- Error are independent each other
- Errors have a normal distribution with mean 0 and homogeneous variance σ^2

CRD with a Covariate - example

Diet A		Diet B		Diet C	
Initial weight (kg)	Gain (g/d)	Initial weight (kg)	Gain (g/d)	Initial weight (kg)	Gain (g/d)
350	970	390	990	400	990
400	1000	340	950	320	940
360	980	410	980	330	930
350	980	430	990	390	1000
340	970	390	980	420	1000

```
data gain;
input treatment $ initial gain @@;
datalines;
A 350 970 B 390 990 C 400 990
A 400 1000 B 340 950 C 320 940
A 360 980 B 410 980 C 330 930
A 350 980 B 430 990 C 390 1000
A 340 970 B 390 980 C 420 1000
;

proc print data=gain;
run;

proc glm;
class treatment;
model gain=initial treatment/solution;
lsmeans treatment/stderr pdiff tdiff adjust=tukey;
run;
```

SAS OUTPUT

SAS System 16:33 Monday, February 1, 2010 4

Oss	treatment	initial	gain
1	A	350	970
2	B	390	990
3	C	400	990
4	A	400	1000
5	B	340	950
6	C	320	940
7	A	360	980
8	B	410	980
9	C	330	930
10	A	350	980
11	B	430	990
12	C	390	1000
13	A	340	970
14	B	390	980
15	C	420	1000

La procedura GLM

Variabile dipendente: gain

Origine	DF	Somma dei quadrati	Media quadratica	Valore F	Pr > F
Modello	3	5492.014652	1830.671551	19.34	0.0001
Errore	11	1041.318681	94.665335		
Totale corretto	14	6533.333333			

R-quadro	Var coeff	Radice MSE	gain Media
0.840614	0.996206	9.729611	976.6667

SAS OUTPUT

Origine	DF	Tipo I SS	Media quadratica	Valore F	Pr > F
initial	1	4441.252588	4441.252588	46.92	<.0001
treatment	2	1050.762064	525.381032	5.55	0.0216

Origine	DF	Tipo III SS	Media quadratica	Valore F	Pr > F
initial	1	5318.681319	5318.681319	56.18	<.0001
treatment	2	1050.762064	525.381032	5.55	0.0216

Parametro	Stima	Errore standard	Valore t	Pr > |t|
Interc	747.1648352 B	30.30956710	24.65	<.0001
initial	0.6043956	0.08063337	7.50	<.0001
treatment A	15.2527473 B	6.22915600	2.45	0.0323
treatment B	-6.0879121 B	6.36135441	-0.96	0.3591
treatment C	0.0000000 B	.	.	.

NOTA: La matrice X'X è singolare ed è stata utilizzata una inversa generalizzata per risolvere le equazioni normali. I termini le cui stime sono seguite dalla lettera 'B' non sono stimati in modo univoco.

SAS OUTPUT

La procedura GLM
Medie dei minimi quadrati
Correzione per i confronti multipli: Tukey-Kramer

treatment	gain LSMEAN	Errore standard	Pr > |t|	Numero LSMEAN
A	988.864469	4.509065	<.0001	1
B	967.523810	4.570173	<.0001	2
C	973.611722	4.356524	<.0001	3

Medie dei minimi quadrati per l'effetto treatment
t per H0: LSMean(i)=LSMean(j) / Pr > |t|

Variabile dipendente: gain

i/j	1	2	3
1		3.198241 0.0213	2.448606 0.0765
2	-3.19824 0.0213		-0.95702 0.6175
3	-2.44861 0.0765	0.957015 0.6175	

CRD with a Covariate - example

```
gain<-data.frame(trt=c("A","A","A","A","A",
                       "B","B","B","B","B",
                       "C","C","C","C","C"),
                 in_weight=c(350,400,360,350,340,
                             390,340,410,430,390,
                             400,320,330,390,420),
                 gain=c(970,1000,980,980,970,
                        990, 950,980,990,980,
                        990,940,930,1000,1000),
                 stringsAsFactors = TRUE)


print(gain) #See the Data
```

Diet A		Diet B		Diet C	
Initial weight (kg)	Gain (g/d)	Initial weight (kg)	Gain (g/d)	Initial weight (kg)	Gain (g/d)
350	970	390	990	400	990
400	1000	340	950	320	940
360	980	410	980	330	930
350	980	430	990	390	1000
340	970	390	980	420	1000

R OUTPUT

```r
contrasts(gain$trt)<-contr.SAS # Set contrasts like SAS?

tm<-lm(gain ~ in_weight + trt, data=gain) # Fit the linear model
summary(tm)
```

```
##
## Call:
## lm(formula = gain ~ in_weight + trt, data = gain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.615  -4.066   0.000   3.319  17.121
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 747.16484   30.30957  24.651 5.61e-11 ***
## in_weight     0.60440    0.08063   7.496 1.21e-05 ***
## trt1         15.25275    6.22916   2.449   0.0323 *
## trt2         -6.08791    6.36135  -0.957   0.3591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.73 on 11 degrees of freedom
## Multiple R-squared:  0.8406, Adjusted R-squared:  0.7971
## F-statistic: 19.34 on 3 and 11 DF,  p-value: 0.0001078
```

these p-values are for a t-test whose null hypothesis is that the regression coefficient is not 0

It must be interpreted as 'standard' regression coefficient

ANOVA tables

```r
fm<-aov(gain ~ in_weight + trt, data=gain) # Fit the ANOVA
summary(fm) #ANOVA table SS I
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## in_weight    1   4441    4441   46.91 2.77e-05 ***
## trt          2   1051     525    5.55   0.0216 *
## Residuals   11   1041      95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#install.packages("car")
car::Anova(tm,type="III") #Anova table with SS III
```

```
## Anova Table (Type III tests)
##
## Response: gain
##             Sum Sq Df  F value    Pr(>F)
## (Intercept)  57526  1 607.6778 5.614e-11 ***
## in_weight     5319  1  56.1840 1.207e-05 ***
## trt           1051  2   5.5499   0.02155 *
## Residuals     1041 11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple comparisons

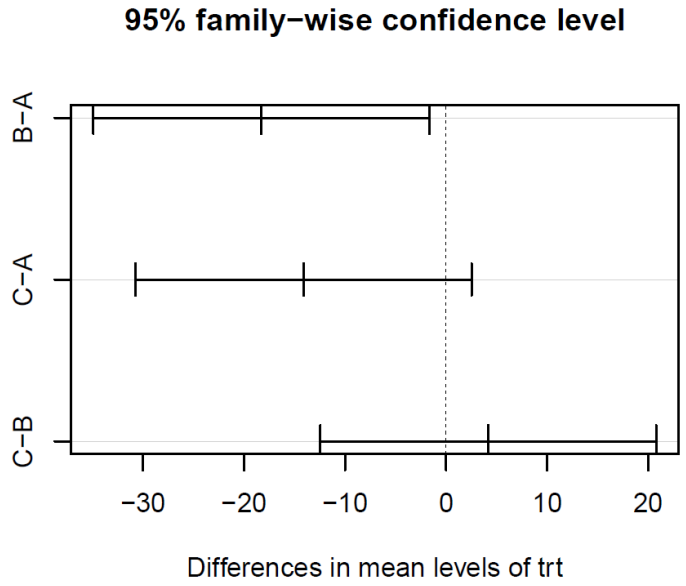
```
TukeyHSD(fm,"trt")          # Tukey test for multiple comparisons
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = gain ~ in_weight + trt, data = gain)
##
## $trt
##           diff       lwr       upr     p adj
## B-A -18.273292 -34.89314 -1.653444 0.0315687
## C-A -14.102484 -30.72233  2.517363 0.0991696
## C-B   4.170807 -12.44904 20.790655 0.7808618
```

this is the onl significant difference
because it is the only p-value < 0.05

0 is included in the interval [lwr, upr] so that the difference is not significant
because repeating the experiment many many times we could sometimes get 0

```
plot(TukeyHSD(fm,"trt"))    # Plot for tukey test
```




```
aggregate(gain$gain, by=list(gain$trt), FUN = mean) # Means by group
```

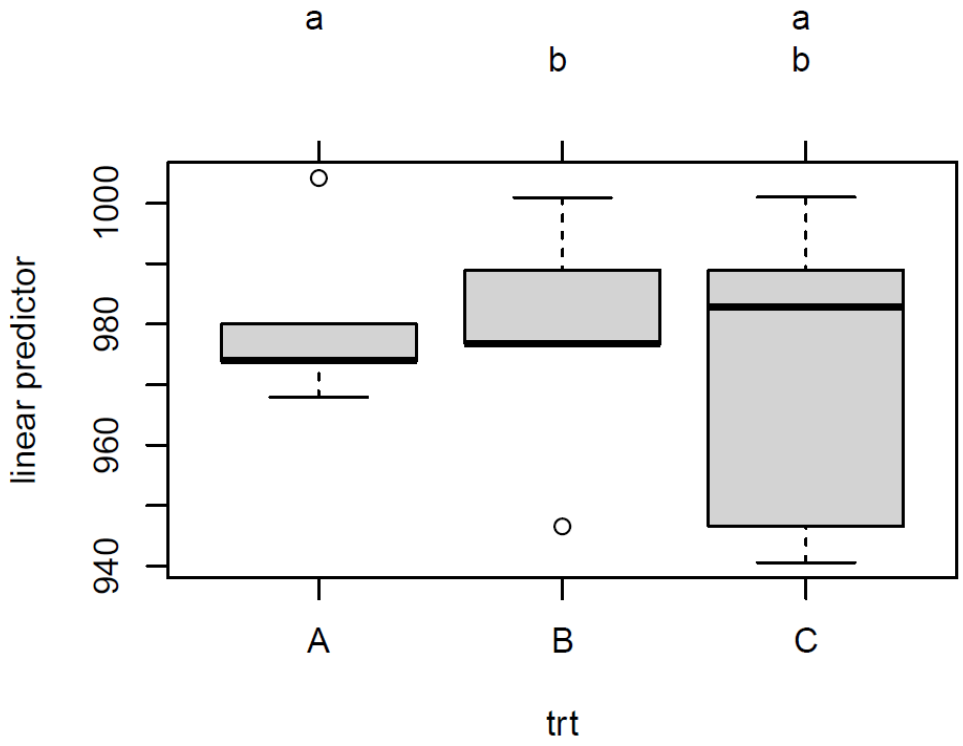
```
##   Group.1   x
## 1       A 980
## 2       B 978
## 3       C 972
```

```
lsmeans::lsmeans(tm,"trt")          #LSM for treatment
```

```
##  trt   lsmean       SE df lower.CL upper.CL
##  A   988.8645 4.509065 11 978.9401 998.7889
##  B   967.5238 4.570173 11 957.4649 977.5827
##  C   973.6117 4.356524 11 964.0231 983.2004
##
## Confidence level used: 0.95
```

Plot for differences

```
library(multcomp)
par(mar=c(4,4,6,2)) # Change parameters for the plot margins
tuk <- multcomp::glht(fm, linfct=multcomp::mcp(trt="Tukey")) # Fit the general Linear Hypotheses
plot(multcomp::cld(tuk, level=0.05),col="lightgrey") # Plot the mean differences
```



Questions?