Seminar

Multivariate Statistics for Biological Sciences

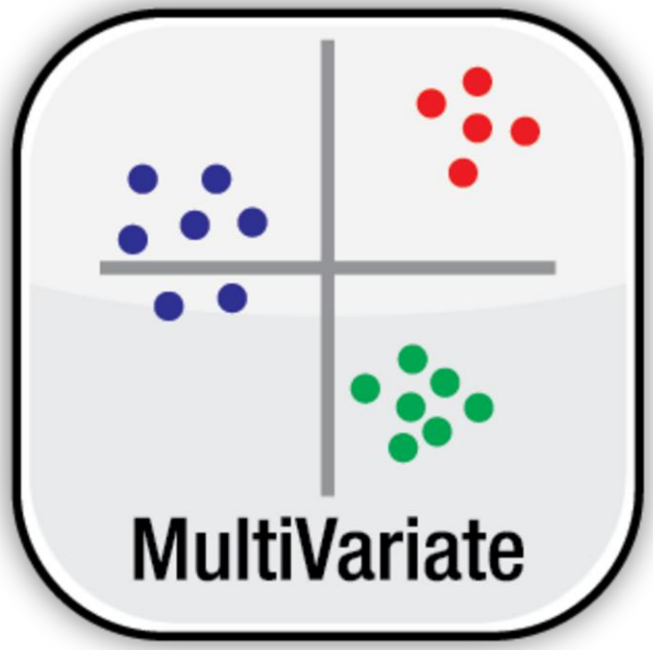
Anderson Freitas, MSc.

Piracicaba – SP – Brazil
May, 2025

Multivariate Statistics

Multivariate statistics (MVA) is a subdivision of statistics encompassing the **simultaneous observation and analysis of more than one outcome variable**.

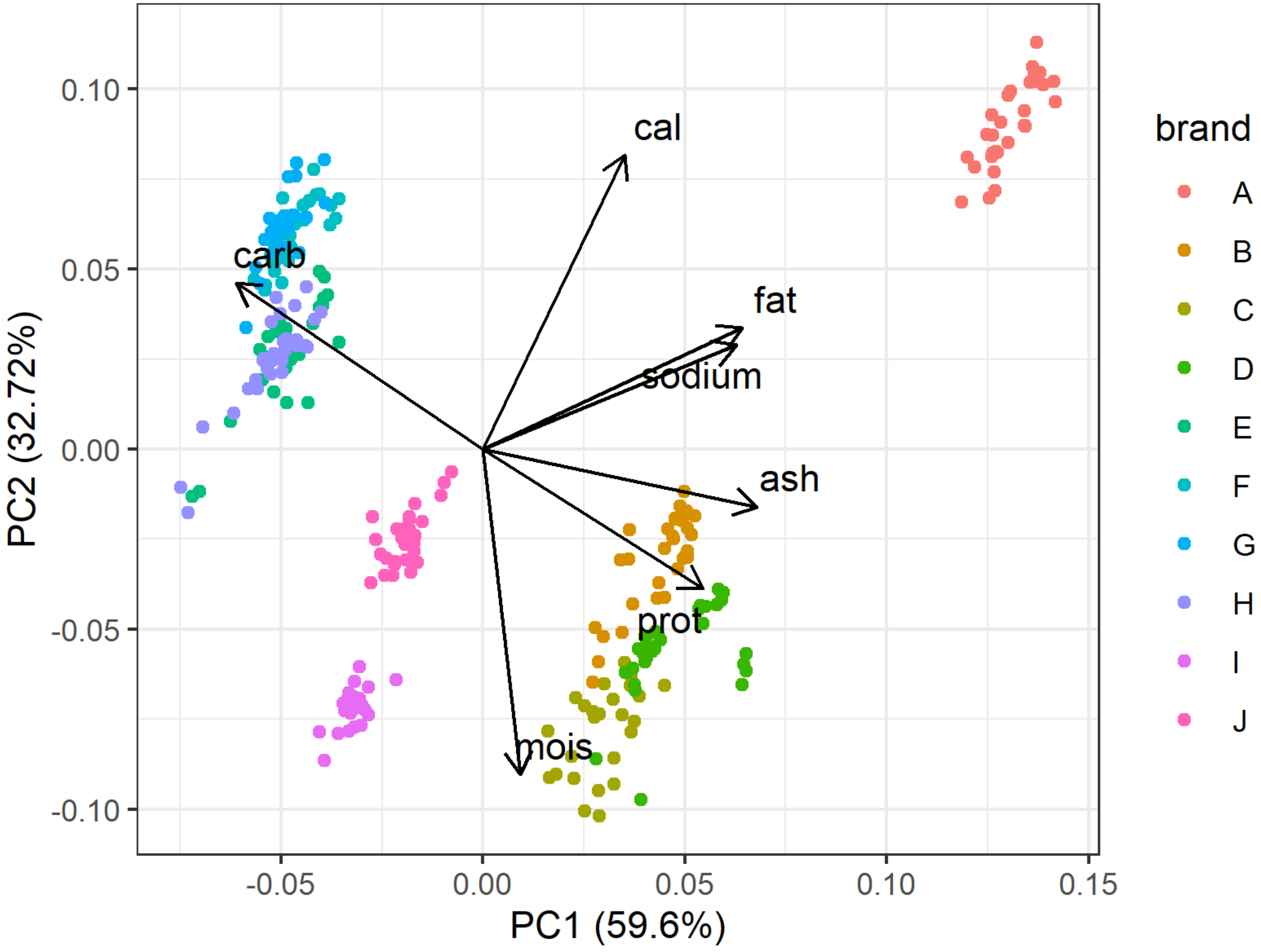
The main point of a multivariate analysis is to consider several related variables **simultaneously**, all of which are considered equally important, at least initially (MANLY, 2008).



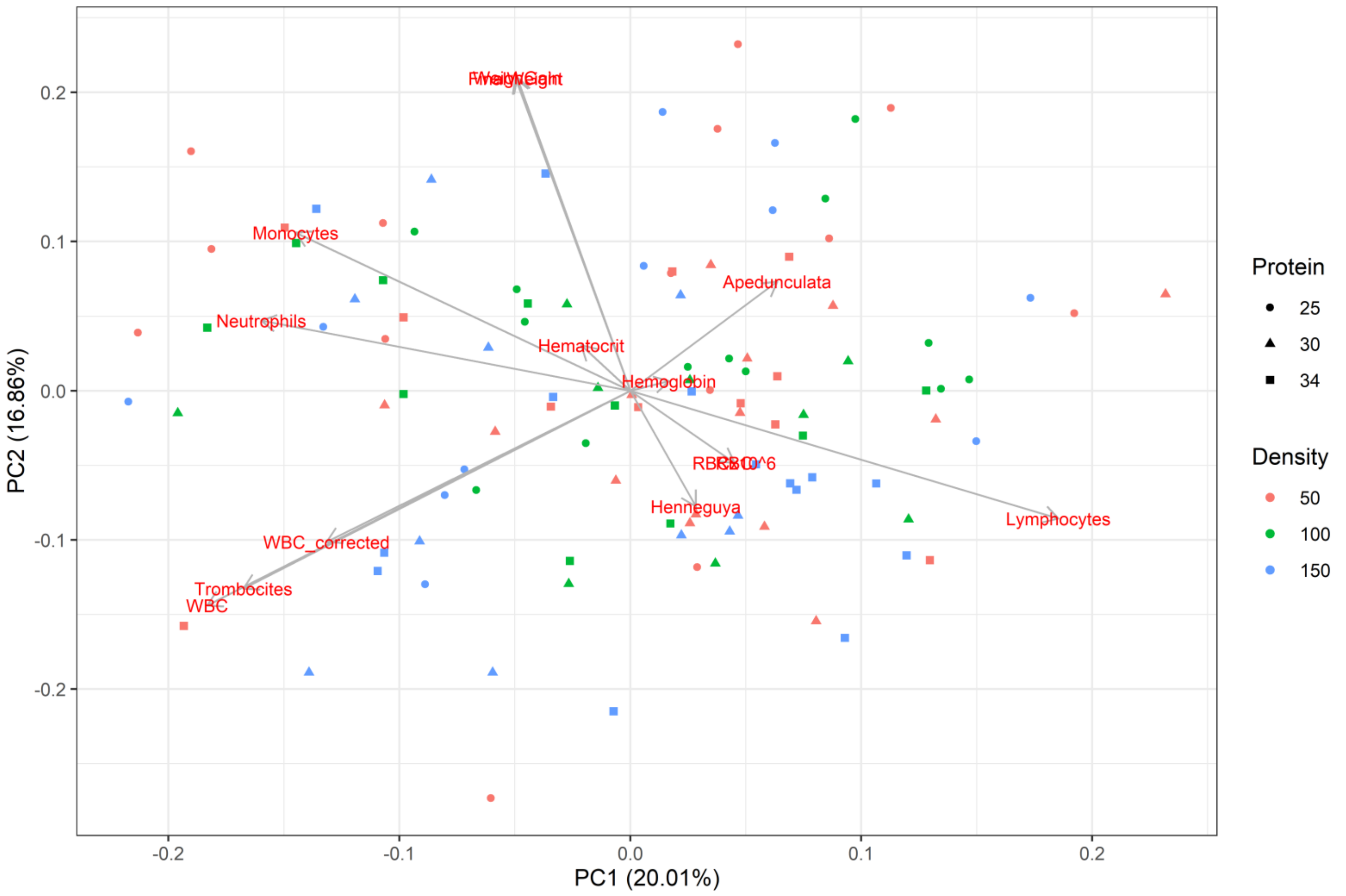
Examples of usages of MVA

Treatment	Temperature	Oxygen	pH	Conductivity	Alkalinity	Hardness	Ammonium	Nitrites
BioMos2%	27.32	7.77	6.78	578.3	26.67	13.33	0.2	0.61
BioMos2%	27.4	7.11	6.08	578.3	6.67	33.33	0	0.16
BioMos2%	27.46	6.95	5.19	131.3	18.33	3.33	0	0.21
BioMos2%	27.8	6.88	5.56	266.9	20	21.67	0	0.02
BioMos2%	27.76	6.75	6.35	207	16.67	11.67	0	0.05
BioMos2%	27.82	6.68	7.1	160.6	30	3.33	0	0.05
BioMos2%	27.67	7.66	7.33	320.4	19.7	14.44	0.03	0.18
BioMos2%	27.48	7.55	7.22	277.4	18.5	14.63	0.01	0.11
Biomos4%	27.1	7.76	6.98	615	23.33	13.33	0.01	0.6
Biomos4%	26.97	7.16	6.02	615	11.67	33.33	0	0.44
Biomos4%	27.06	6.88	5.01	117.8	3.33	3.33	0	0.01
Biomos4%	27.21	6.86	5.39	204.9	25	6.67	0	0.19
Biomos4%	27.31	6.79	6.22	204.7	16.67	10	0	0.27
Biomos4%	27.47	6.53	7.1	107.8	23.33	1.67	0	0.03
Biomos4%	27.56	7.41	7.42	310.87	17.22	11.39	0.00	0.26
Biomos4%	27.39	7.52	7.34	260.18	16.20	11.06	0.00	0.20
Biomos6%	26.63	7.66	6.82	630.3	26.67	16.67	0	1.83
Biomos6%	26.47	7.15	6.15	630.3	1.67	45	0	0.16
Biomos6%	26.51	7	5.71	116	6.67	6.67	0	2.11
Biomos6%	26.59	6.89	6.1	460.7	16.67	13.33	0	4.13
Biomos6%	26.72	6.82	6.39	215.7	40.33	13.33	0.1	1.27
Biomos6%	26.69	6.8	7.38	118	21.67	1.67	0	0.02
Biomos6%	26.68	7.67	7.67	361.83	18.95	16.11	0.02	1.59
Biomos6%	26.33	7.58	7.54	317.09	17.66	16.02	0.02	1.55

The Standard Figure - Expectation



The Standard Figure - Reality



How to calculate?

Considere dois indivíduos nos quais foi medida a variável quantitativa X. Naturalmente, uma medida de distância entre esses indivíduos seria:

$$d_{12} = x_1 - x_2$$

Ou ainda:

$$d_{12} = |x_1 - x_2|$$

$$d_{12} = \sqrt{(x_1 - x_2)^2}$$

How to calculate?

No caso em que p variáveis quantitativas foram medidas, pode-se generalizar para o caso multivariado, conhecida como **distância euclidiana**:

$$d_{ij} = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2}$$

How to calculate?

No caso em que p variáveis quantitativas foram medidas, pode-se generalizar para o caso multivariado, conhecida como **distância euclidiana**:

$$d_{ij} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$$

How to calculate?

- Em muitos métodos multivariados, a padronização de variáveis é um procedimento útil e, muitas vezes, necessário para eliminar influências das diferentes unidades de medida das variáveis nos resultados.

- A principal forma de padronização de variáveis é:

$$z = \frac{x - \mu}{\sigma}$$

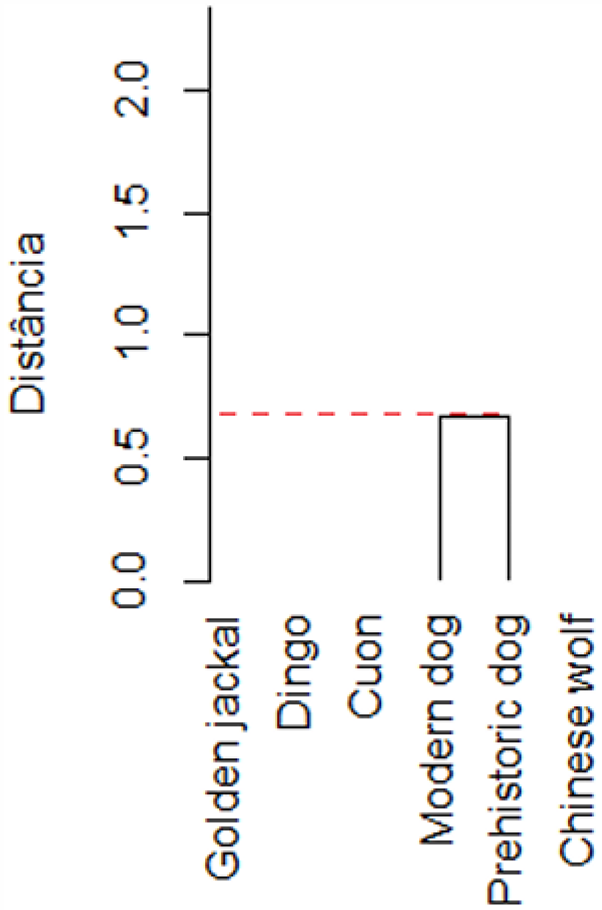
How to calculate?

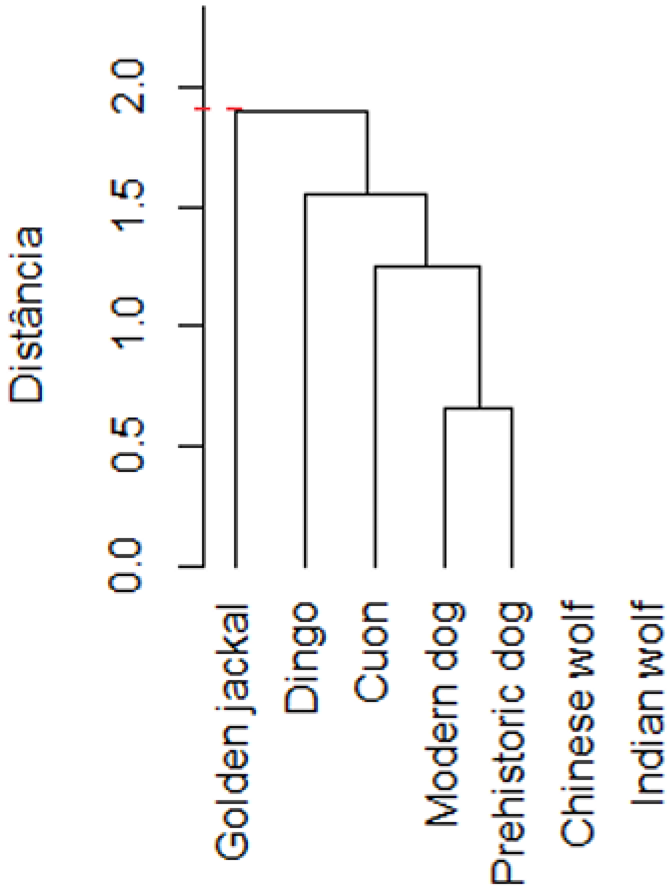
Tabela 1. Distâncias euclidianas entre sete grupos caninos.

	CM	CD	LC	LI	Cuon	Dingo	CPH
CM	-	-	-	-	-	-	-
CD	1,91	-	-	-	-	-	-
LC	5,38	7,12	-	-	-	-	-
LI	3,38	5,06	2,14	-	-	-	-
Cuon	1,51	3,19	4,57	2,91	-	-	-
Dingo	1,56	3,18	4,21	2,20	1,67	-	-
CPH	0,66	2,39	5,12	3,24	1,26	1,71	-

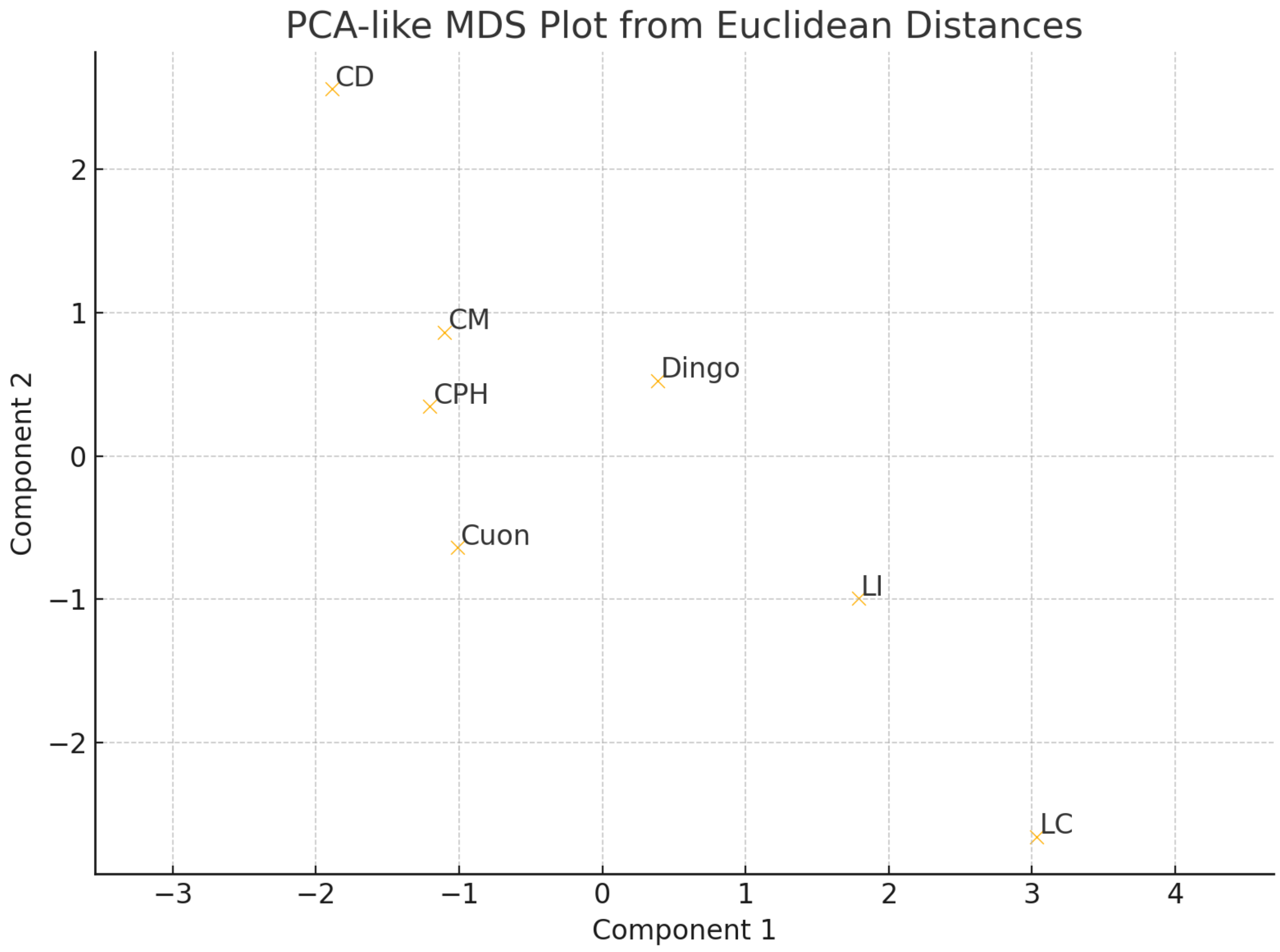
How to calculate?

	CM	CD	LC	LI	Cuon	Dingo	CPH
CM	-	-	-	-	-	-	-
CD	1,91	-	-	-	-	-	-
LC	5,38	7,12	-	-	-	-	-
LI	3,38	5,06	2,14	-	-	-	-
Cuon	1,51	3,19	4,57	2,91	-	-	-
Dingo	1,56	3,18	4,21	2,20	1,67	-	-
CPH	0,66	2,39	5,12	3,24	1,26	1,71	-





How to calculate?



How to calculate?

$$X_1 = +\mathbf{0,85}\ F_1 - 0{,}10\ F_2 - 0{,}27\ F_3 + 0{,}36\ F_4\ +\ e_1\ (0{,}93)$$

$$X_2 = +0{,}11\ F_1 - 0{,}30\ F_2 - \mathbf{0,86}\ F_3 + 0{,}10\ F_4\ +\ e_2\ (0{,}85)$$

$$X_3 = -0{,}03\ F_1 - 0{,}32\ F_2 + \mathbf{0,89}\ F_3 + 0{,}09\ F_4\ +\ e_3\ (0{,}91)$$

$$X_4 = -0{,}19\ F_1 + 0{,}04\ F_2 + \mathbf{0,64}\ F_3 - 0{,}14\ F_4\ +\ e_4\ (0{,}46)$$

$$X_5 = -0{,}02\ F_1 - 0{,}08\ F_2 + 0{,}04\ F_3 - \mathbf{0,95}\ F_4\ +\ e_5\ (0{,}92)$$

$$X_6 = -0{,}35\ F_1 + 0{,}48\ F_2 + 0{,}15\ F_3 - \mathbf{0,65}\ F_4\ +\ e_6\ (0{,}79)$$

$$X_7 = -0{,}08\ F_1 + \mathbf{0,93}\ F_2 + 0{,}00\ F_3 + 0{,}01\ F_4\ +\ e_7\ (0{,}87)$$

$$X_8 = -\mathbf{0,91}\ F_1 + 0{,}17\ F_2 + 0{,}12\ F_3 - 0{,}04\ F_4\ +\ e_8\ (0{,}88)$$

$$X_9 = -\mathbf{0,73}\ F_1 - \mathbf{0,57}\ F_2 + 0{,}30\ F_3 + 0{,}14\ F_4 +\ e_9\ (0{,}87)$$

Back to the basics: ANOVA

Tabela 1. Quadro da ANOVA de um fator.

Causas de variação	Graus de liberdade	Soma de Quadrados	Quadrados médios	F
Fatores	$k-1$	SQ_{Trat}	QM_{Trat}	$F=\frac{QM_{Trat}}{QM_{Res}}$
Resíduo	$k(r-1)$	SQ_{Res}	QM_{Res}	
Total	$\boldsymbol{kr-1}$	$\boldsymbol{SQ_{Total}}$		

p-value

- $SQ_{Total}=\sum y^2-C$
- $SQ_{Res}=SQ_{Total}-SQ_{Trat}$
- $SQ_{Trat}=\frac{\sum T^2}{r}-C$
- $C=\frac{(\sum y)^2}{n}$

ANOVA Output

Tests of Between-Subjects Effects

Dependent Variable: Political interest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5645.998[a]	5	1129.200	78.538	<.001
Intercept	132091.906	1	132091.906	9187.227	<.001
gender	8.420	1	8.420	.586	.448
education_level	5446.697	2	2723.348	189.414	<.001
gender * education_level	210.338	2	105.169	7.315	.002
Error	747.644	52	14.378		
Total	140265.750	58			
Corrected Total	6393.642	57			

a. R Squared = .883 (Adjusted R Squared = .872)

MANOVA

Or a PERMANOVA for non-parametric data!!!

The hypothesis to be verified is:

$$H0 = \boldsymbol{\mu}1 = \boldsymbol{\mu}2 = \cdots = \boldsymbol{\mu}I$$

In other words, the hypothesis is that there are no differences between the true treatment mean vectors.

The alternative hypothesis is that **at least one of the mean vectors** is **different** from the others.

Hands on

¡Muchas gracias!

Contato:

andersonfreitas@usp.br
anderson.defreitas@kaust.edu.sa

@FreitasSays

@senhorf

Follow me on
Social Media