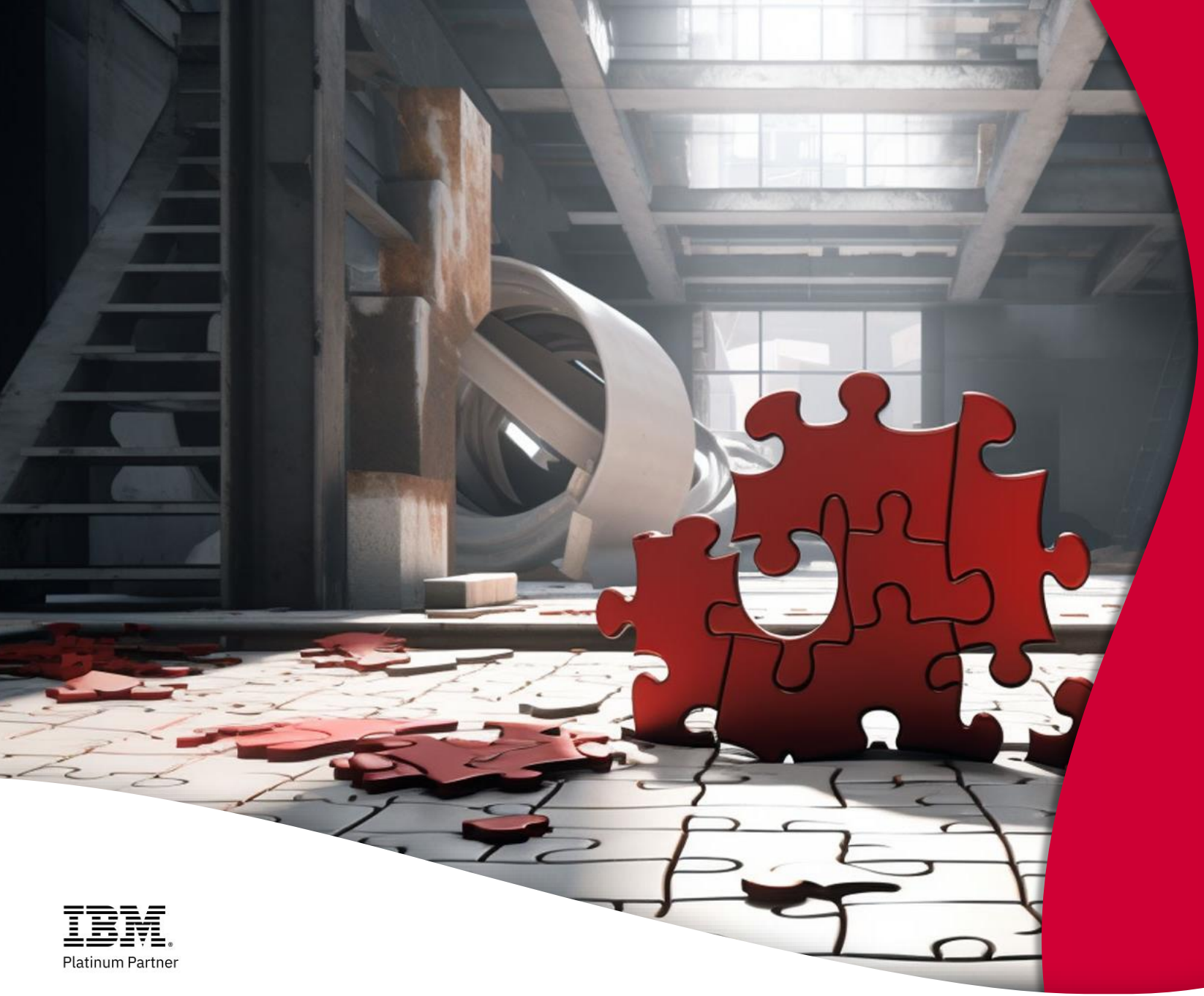


WEBINAR

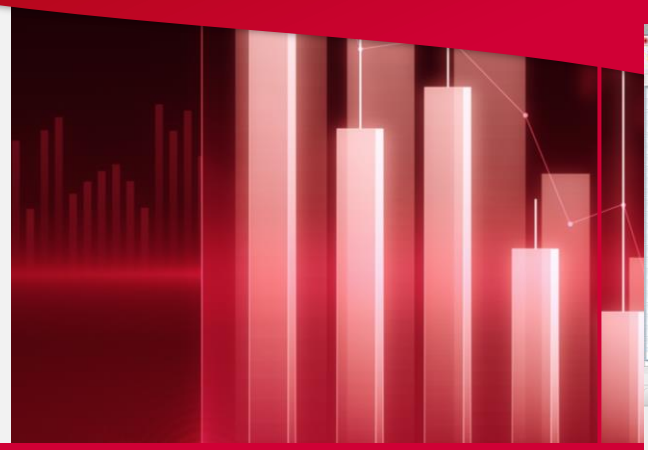
Missing Values with Statistics for Data Analysis

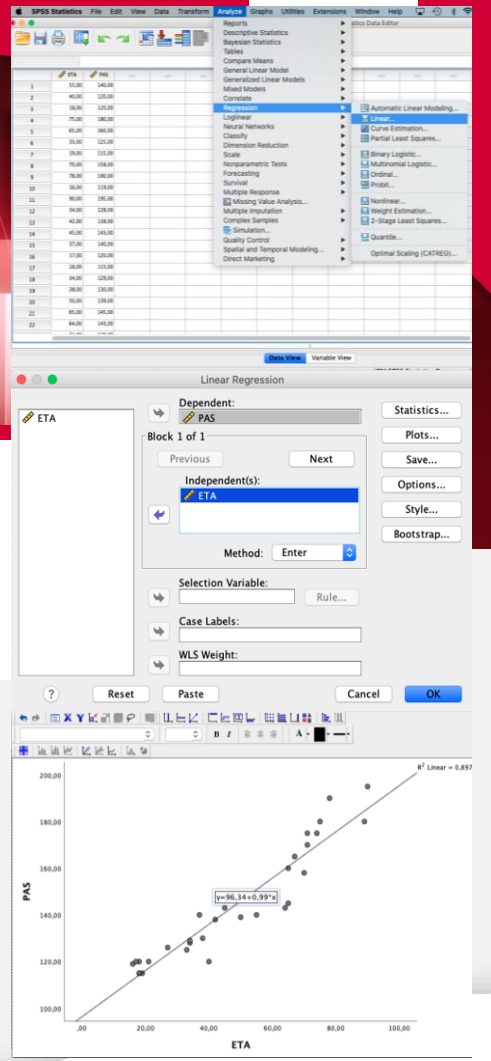
Statistics for Data Analysis

powerd by SPSS

IBM Platinum Partner

WEBINAR AGENDA





About Us

25+ years of history… in brief

Statistics for Data Analysis

Services included in the solution (LaunchBox, Add-On, APP, webinar, Support, …)

Missing Values with Statistics for Data Analysis

- What **Missing Values** are
- **Types** of missing information:
  - MCAR
  - MAR
  - MNAR
- **Replacement** methodologies
- Practicl **example**
- Q&A

Scarica la nostra APP sugli store

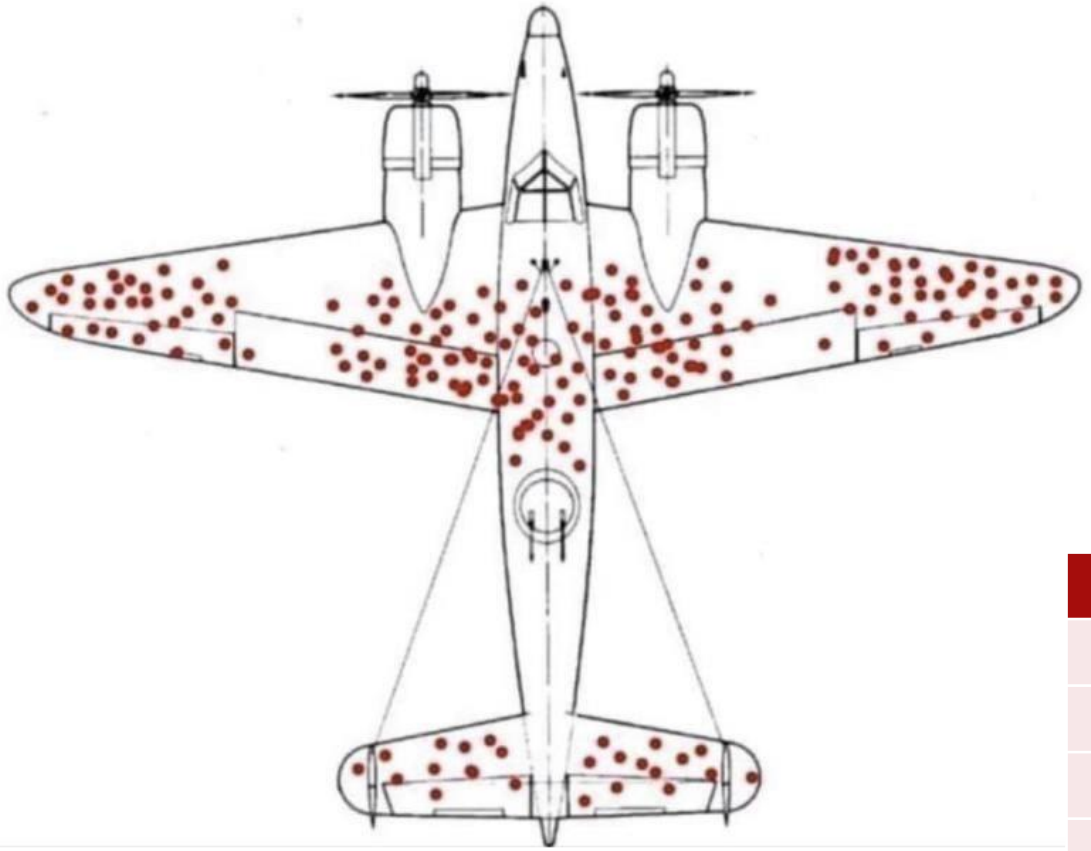
GET IT ON Google Play

Download on the App Store

www.statisticsfordataanalysis.com


Statistics for Data Analysis
SPSS
Missing Values
with
Statistics for Data Analysis
What are Missing Values?

What are missing Values?



Abraham Wald: (1902-1950), Hungarian mathematician, escaped from Europe to evade Nazi persecutions, and was a member of the Statistical Research Group (SRG). The SRG was one of the groups of scientists and university researchers created in the United States during World War II to support the war effort.

Airplane sections	Bullet holes for "square foot"
Engine	1,11
Fuselage	1,73
Tank	1,55
Rest of the plane	1,8

Missing Values
with
Statistics for Data Analysis

Types of missing information:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

Types of Missing Values

Missing Values

Missing Values analysis can make a difference

01

The reasons why some values are "missing" can be "ignorable" and "non-ignorable"

02

If the process by which missing values occur can be ignored when investigating a particular phenomenon, the presence of missing values can be considered neutral for the purposes of the study

03

If, however, the process is not "ignorable", the presence of missing values CANNOT be considered neutral for understanding the phenomenon of interest and these must be considered in its parameterization

Types of Missing Values

Rubin (1976) defined three different types of missing information

Missing Values

Missing Values analysis can make a difference

01

Missing Completely at Random (MCAR)

MCAR and MAR result from a generation mechanism that can be "ignorable" while MNAR belong to the "non-ignorable" type

02

Missing at Random (MAR)

MCAR assumes that there is no significant difference between complete and incomplete information.

03

Missing Not at Random (MNAR)

MAR supposes the variable Yi missing values can be linked to other measured variables, but they do not depend on Yi, so the mechanism for generating missing values can be assumed as MAR (e.g. male respondents who usually do not complete the survey on depression even if this has no bearing on their possible depressive state)

Example

- **Example:** Non-random NA

Data from Supplementary Survey of Older Americans

n_1=2343 complete data

n_2= 579 missing data

Years of schooling:

Complete case : $\overline{x_1}$=12.3

Incomplete case: $\overline{x_1}$=13.1

Intake of Vitamin C:

Complete case : $\overline{x_2}$=68

Incomplete case: $\overline{x_2}$=77

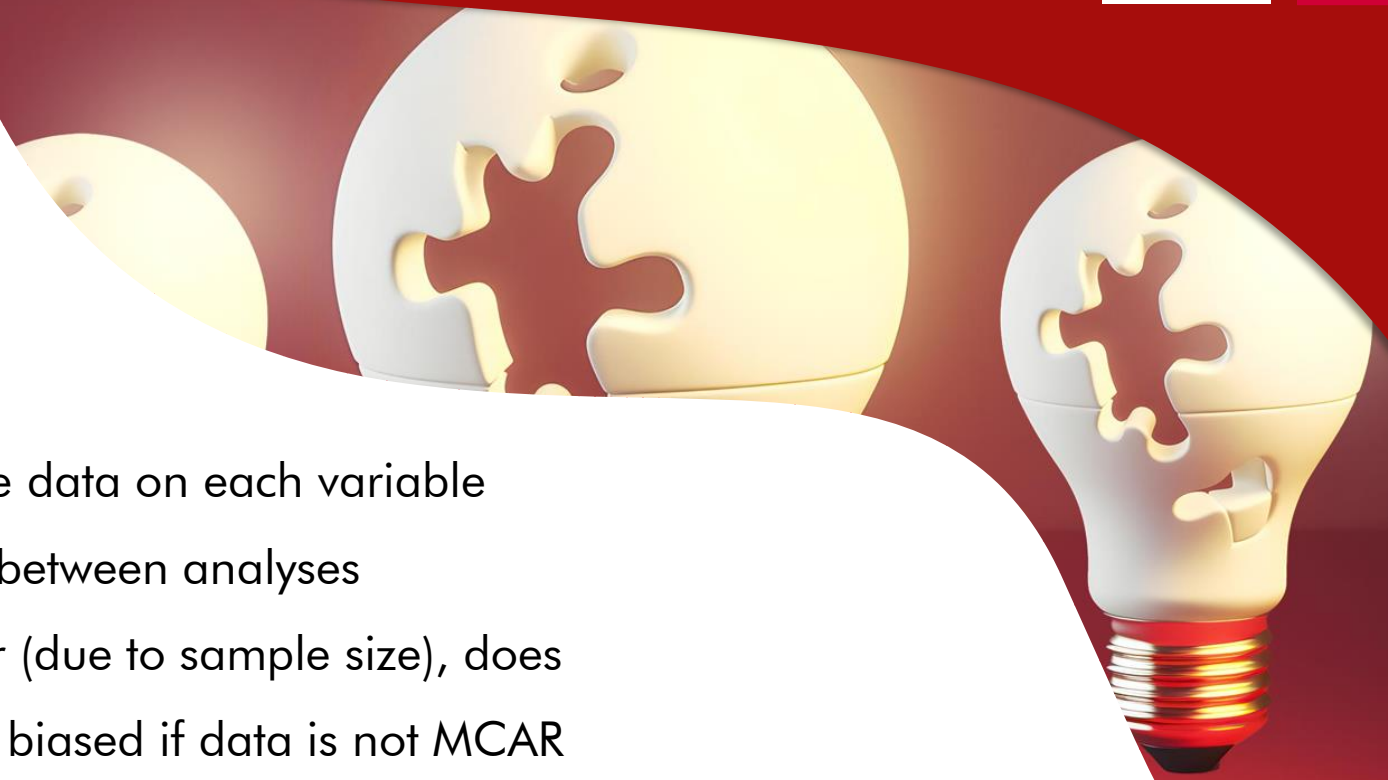
Both years of education and Vitamin C intake are higher in incomplete cases. Here, income variable is likely to be missing not at random


Statistics for Data Analysis
SPSS
Missing Values
with
Statistics for Data Analysis
Replacement methodologies

Missing values management methods

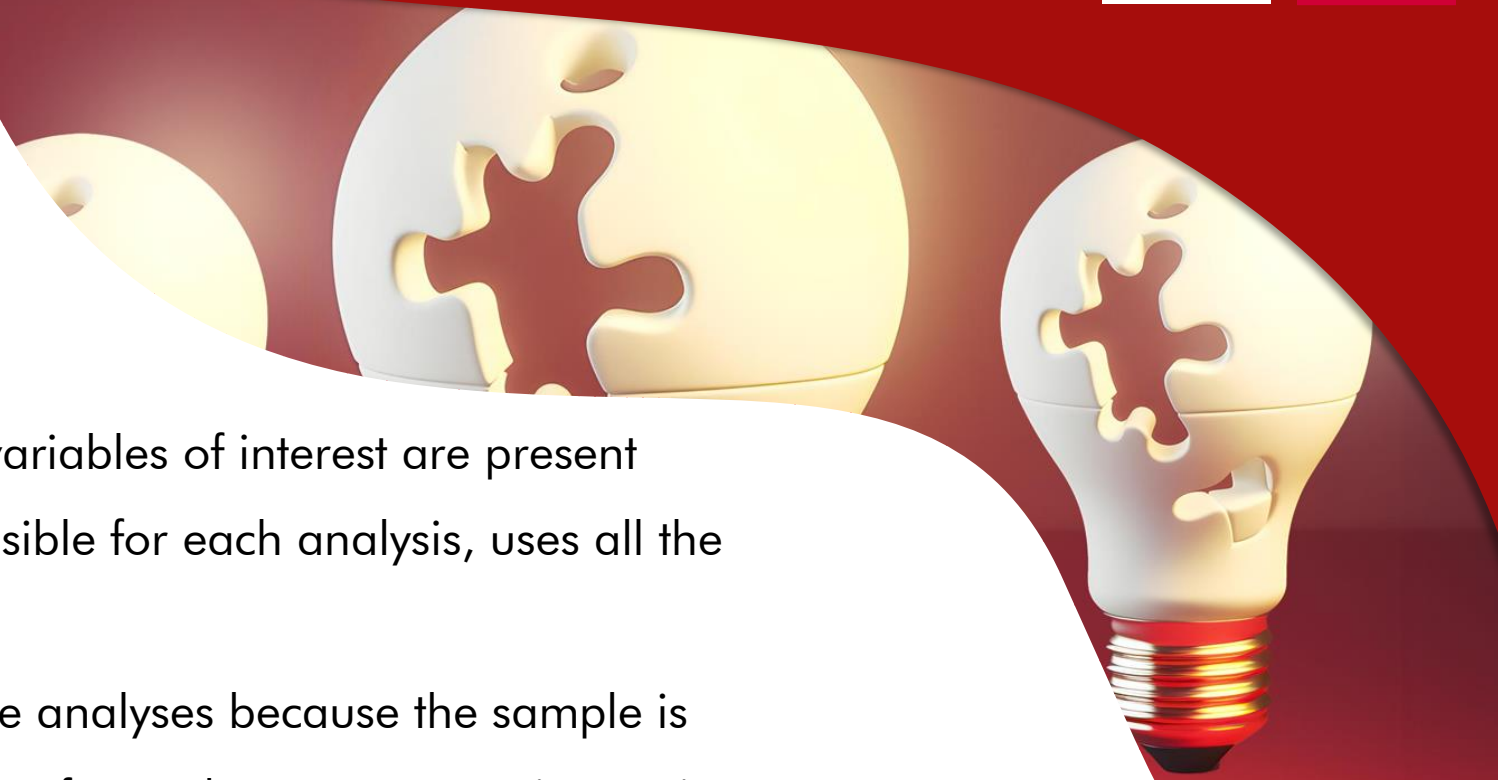
1. Deletion of the incomplete information (LISTWISE)
2. Partial deletion of the incomplete information (LPAIRWISE)
3. Missing Values Replacements

LISTWISE DELETION



- You can analyze only cases with available data on each variable
- Advantage: simplicity and comparability between analyses
- Disadvantage: it reduces statistical power (due to sample size), does not use all information, estimates can be biased if data is not MCAR
- Listwise deletion allows you to obtain regression slope correct estimates when the missing values are not a function of the response variable

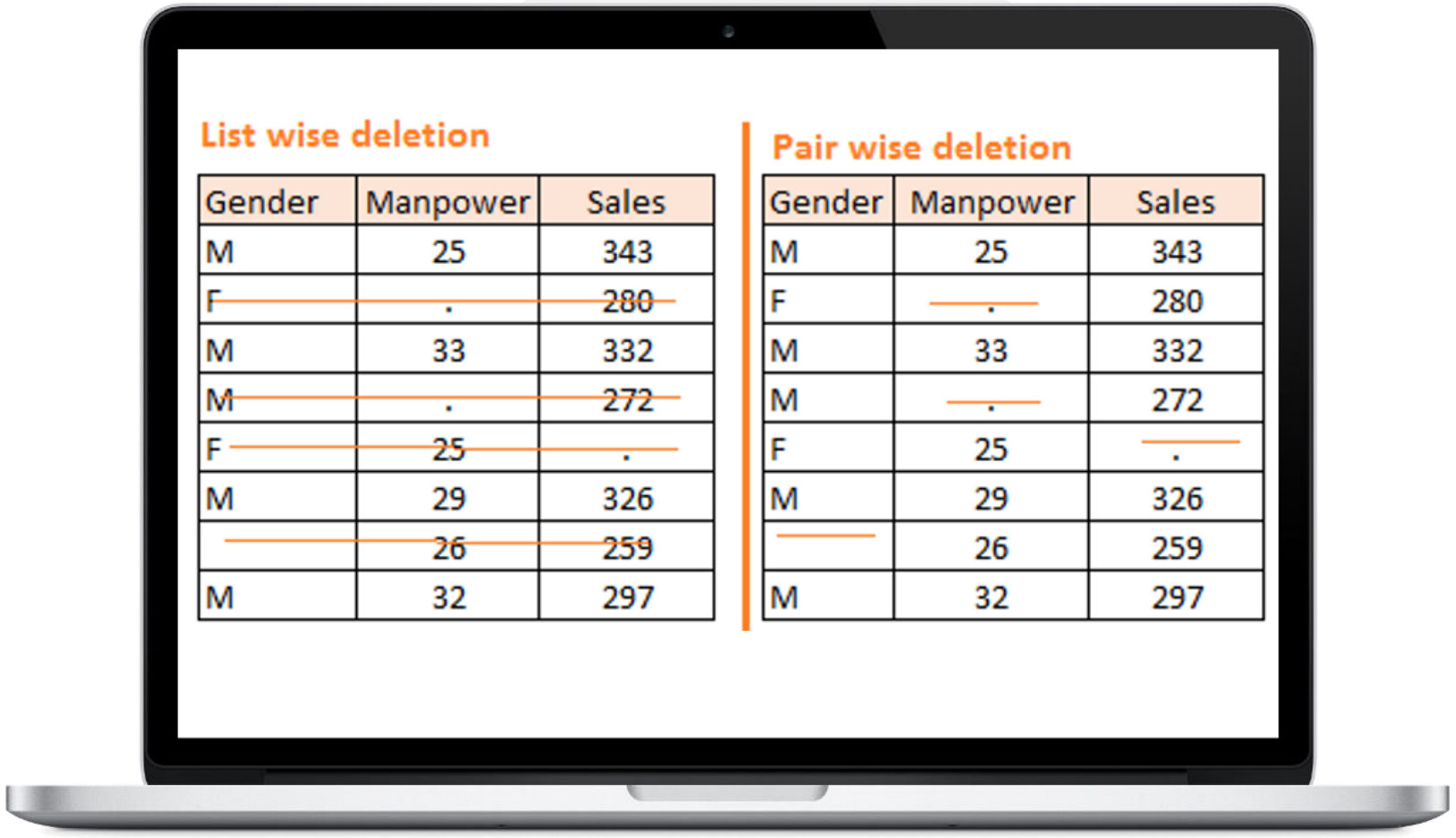
PAIRWISE DELETION



- You can analyse of all cases in which the variables of interest are present
- Advantage: it Keeps as many cases as possible for each analysis, uses all the information possible with each analysis
- Disadvantage: it is not possible to compare analyses because the sample is different each time, the sample size changes for each parameter estimate, it can lead to meaningless results
- Compute summary statistics using n_i observations not n
- Compute correlation type statistics using complete pairs for both variables


Statistics for Data Analysis
SPSS
Missing Values
with
Statistics for Data Analysis
Example

Example



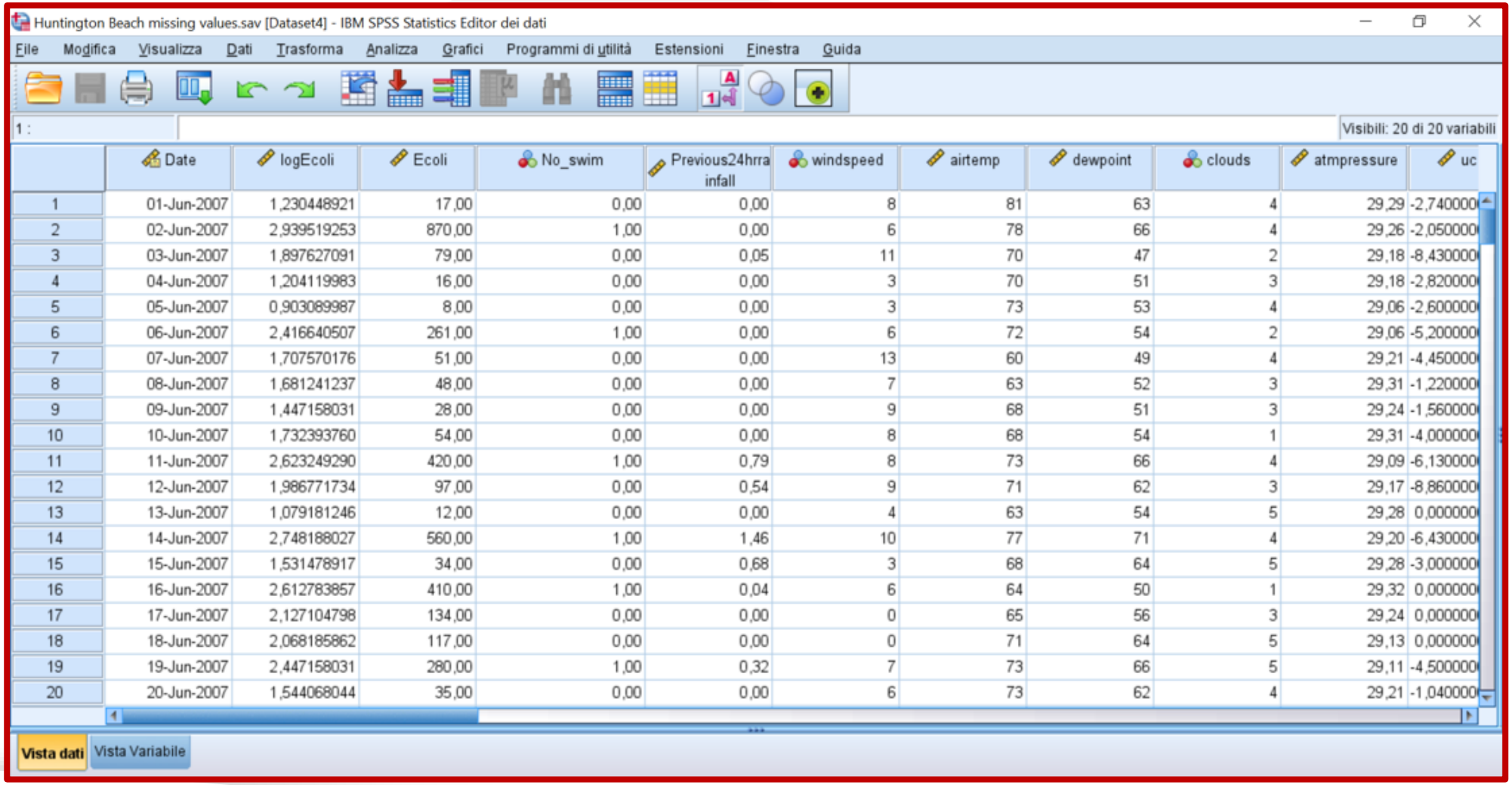
List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Example



	Date	logEcoli	Ecoli	No_swim	Previous24hrrainfall	windspeed	airtemp	dewpoint	clouds	atmpressure	uc
1	01-Jun-2007	1,230448921	17,00	0,00	0,00	8	81	63	4	29,29	-2,740000
2	02-Jun-2007	2,939519253	870,00	1,00	0,00	6	78	66	4	29,26	-2,050000
3	03-Jun-2007	1,897627091	79,00	0,00	0,05	11	70	47	2	29,18	-8,430000
4	04-Jun-2007	1,204119983	16,00	0,00	0,00	3	70	51	3	29,18	-2,820000
5	05-Jun-2007	0,903089987	8,00	0,00	0,00	3	73	53	4	29,06	-2,600000
6	06-Jun-2007	2,416640507	261,00	1,00	0,00	6	72	54	2	29,06	-5,200000
7	07-Jun-2007	1,707570176	51,00	0,00	0,00	13	60	49	4	29,21	-4,450000
8	08-Jun-2007	1,681241237	48,00	0,00	0,00	7	63	52	3	29,31	-1,220000
9	09-Jun-2007	1,447158031	28,00	0,00	0,00	9	68	51	3	29,24	-1,560000
10	10-Jun-2007	1,732393760	54,00	0,00	0,00	8	68	54	1	29,31	-4,000000
11	11-Jun-2007	2,623249290	420,00	1,00	0,79	8	73	66	4	29,09	-6,130000
12	12-Jun-2007	1,986771734	97,00	0,00	0,54	9	71	62	3	29,17	-8,860000
13	13-Jun-2007	1,079181246	12,00	0,00	0,00	4	63	54	5	29,28	0,000000
14	14-Jun-2007	2,748188027	560,00	1,00	1,46	10	77	71	4	29,20	-6,430000
15	15-Jun-2007	1,531478917	34,00	0,00	0,68	3	68	64	5	29,28	-3,000000
16	16-Jun-2007	2,612783857	410,00	1,00	0,04	6	64	50	1	29,32	0,000000
17	17-Jun-2007	2,127104798	134,00	0,00	0,00	0	65	56	3	29,24	0,000000
18	18-Jun-2007	2,068185862	117,00	0,00	0,00	0	71	64	5	29,13	0,000000
19	19-Jun-2007	2,447158031	280,00	1,00	0,32	7	73	66	5	29,11	-4,500000
20	20-Jun-2007	1,544068044	35,00	0,00	0,00	6	73	62	4	29,21	-1,040000

Example

Correlazioni

		Ecoli	Previous24hrrainfall	atmpressure	year	wind_sp	turbidity_miss	cuyahogariverflow_m	rockyriverflow_m
Ecoli	Correlazione di Pearson	1	0,377**	-0,208**	-0,185**	0,230**	0,743**	0,188**	0,116
	Sign. (a due code)		0,000	0,002	0,005	0,001	0,000	0,005	0,102
	N	225	225	225	225	225	206	217	199
Previous24hrrainfall	Correlazione di Pearson	0,377**	1	-0,239**	-0,005	0,072	0,190**	0,243**	0,315**
	Sign. (a due code)	0,000		0,000	0,945	0,283	0,006	0,000	0,000
	N	225	225	225	225	225	206	217	199
atmpressure	Correlazione di Pearson	-0,208**	-0,239**	1	-0,212**	-0,411**	-0,012	0,033	0,000
	Sign. (a due code)	0,002	0,000		0,001	0,000	0,866	0,634	0,999
	N	225	225	225	225	225	206	217	199
year	Correlazione di Pearson	-0,185**	-0,005	-0,212**	1	0,112	-0,100	-0,362**	-0,072
	Sign. (a due code)	0,005	0,945	0,001		0,093	0,154	0,000	0,312
	N	225	225	225	225	225	206	217	199
wind_sp	Correlazione di Pearson	0,230**	0,072	-0,411**	0,112	1	0,176*	-0,144*	-0,181*
	Sign. (a due code)	0,001	0,283	0,000	0,093		0,012	0,034	0,011
	N	225	225	225	225	225	206	217	199
turbidity_miss	Correlazione di Pearson	0,743**	0,190**	-0,012	-0,100	0,176*	1	0,043	0,084
	Sign. (a due code)	0,000	0,006	0,866	0,154	0,012		0,551	0,236
	N	206	206	206	206	206	206	199	199
cuyahogariverflow_m	Correlazione di Pearson	0,188**	0,243**	0,033	-0,362**	-0,144*	0,043	1	0,809**
	Sign. (a due code)	0,005	0,000	0,634	0,000	0,034	0,551		0,000
	N	217	217	217	217	217	199	217	199
rockyriverflow_m	Correlazione di Pearson	0,116	0,315**	0,000	-0,072	-0,181*	0,084	0,809**	1
	Sign. (a due code)	0,102	0,000	0,999	0,312	0,011	0,236	0,000	
	N	199	199	199	199	199	199	199	199

**. La correlazione è significativa a livello 0,01 (a due code).

*. La correlazione è significativa a livello 0,05 (a due code).

Example

Riepilogo elaborazione casi

	Casi					
	Valido		Mancante		Totale	
	N	Percentuale	N	Percentuale	N	Percentuale
Ecoli	199	88,4%	26	11,6%	225	100,0%
turbidity_miss	199	88,4%	26	11,6%	225	100,0%
rockyriverflow_m	199	88,4%	26	11,6%	225	100,0%

Riepilogo elaborazione casi

	Casi					
	Valido		Mancante		Totale	
	N	Percentuale	N	Percentuale	N	Percentuale
Ecoli	225	100,0%	0	0,0%	225	100,0%
turbidity_miss	206	91,6%	19	8,4%	225	100,0%
rockyriverflow_m	199	88,4%	26	11,6%	225	100,0%

Test t di varianza separati[a]

		Ecoli	Previous24hrrainfall	dewpoint	atmpressure	wind_sp	turbidity_miss	cuyahogariverflow_m	rockyriverflow_m
turbidity_miss	t	-0,1	0,8	0,7	-0,6	-0,8	.	2,2	.
	gl	26,7	30,4	20,4	20,0	22,8	.	31,5	.
	P (a 2 code)	0,924	0,445	0,474	0,555	0,448	.	0,039	.
	# presenti	206	206	206	206	206	206	199	199
	# mancanti	19	19	19	19	19	0	18	0
	Media (presenti)	156,1553	0,1249	58,25	29,1836	7,9230	17,4546	772,4422	235,7186
	Media (mancanti)	161,7895	0,0853	56,53	29,2089	8,5780	.	511,2222	.
rockyriverflow_m	t	-0,5	0,9	0,9	0,4	-0,8	0,3	2,2	.
	gl	47,1	59,5	31,2	28,0	37,6	8,0	31,5	.
	P (a 2 code)	0,621	0,354	0,396	0,658	0,425	0,782	0,039	.
	# presenti	199	199	199	199	199	199	199	199
	# mancanti	26	26	26	26	26	7	18	0
	Media (presenti)	153,8191	0,1261	58,29	29,1878	7,9152	17,5048	772,4422	235,7186
	Media (mancanti)	178,1538	0,0865	56,69	29,1700	8,4611	16,0271	511,2222	.

Per ogni variabile quantitativa, vengono formate coppie di gruppi di variabili indicatore (presente, mancante).

a. Le variabili indicatore con meno del 5% di valori mancanti non vengono visualizzate.

Coefficienti[a]

Modello		Coefficienti non standardizzati B	Coefficienti non standardizzati Errore standard	Coefficienti standardizzati Beta	t	Sign.
1	(Costante)	-24,728	20,706		-1,194	0,234
	turbidity_miss	10,200	0,652	0,745	15,654	0,000
2	(Costante)	-46,135	19,499		-2,366	0,019
	turbidity_miss	9,517	0,614	0,695	15,501	0,000
	Previous24hrrainfall	264,672	45,234	0,262	5,851	0,000
3	(Costante)	-545,444	113,993		-4,785	0,000
	turbidity_miss	10,137	0,603	0,740	16,811	0,000
	Previous24hrrainfall	203,420	45,366	0,202	4,484	0,000
	dewpoint	8,513	1,917	0,199	4,440	0,000
4	(Costante)	8150,451	3536,302		2,305	0,022
	turbidity_miss	10,102	0,596	0,737	16,964	0,000
	Previous24hrrainfall	182,735	45,572	0,181	4,010	0,000
	dewpoint	7,605	1,929	0,178	3,943	0,000
	atmpressure	-296,006	120,314	-0,107	-2,460	0,015

a. Variabile dipendente: Ecoli

Coefficienti[a]

Modello		Coefficienti non standardizzati B	Coefficienti non standardizzati Errore standard	Coefficienti standardizzati Beta	t	Sign.
1	(Costante)	97,169	17,337		5,605	0,000
	Previous24hrrainfall	189,243	45,330	0,355	4,175	0,000
2	(Costante)	31,317	21,498		1,457	0,148
	Previous24hrrainfall	197,234	41,998	0,370	4,696	0,000
	turbidity_miss60	4,523	0,982	0,363	4,604	0,000
3	(Costante)	-391,872	117,136		-3,345	0,001
	Previous24hrrainfall	156,649	41,476	0,294	3,777	0,000
	turbidity_miss60	6,009	1,019	0,482	5,897	0,000
	dewpoint	6,932	1,889	0,310	3,669	0,000
4	(Costante)	-352,861	117,318		-3,008	0,003
	Previous24hrrainfall	149,615	41,113	0,281	3,639	0,000
	turbidity_miss60	5,854	1,009	0,469	5,801	0,000
	dewpoint	5,762	1,956	0,258	2,946	0,004
	cuyahogariverflow_m	0,043	0,022	0,157	2,000	0,048
5	(Costante)	-316,966	114,680		-2,764	0,007
	Previous24hrrainfall	183,709	41,718	0,344	4,404	0,000
	turbidity_miss60	6,007	0,982	0,482	6,117	0,000
	dewpoint	5,052	1,916	0,226	2,636	0,010
	cuyahogariverflow_m	0,102	0,030	0,373	3,457	0,001
	rockyriverflow_m	-0,232	0,082	-0,304	-2,832	0,005

a. Variabile dipendente: Ecoli