Modeling and design of experiments

Dr Jean-Marie Fürbringer

École Polytechnique Fédérale de Lausanne

Fall 2025

1.1.1 Organization of the course Fall 2025

Wednesday, sept. 10 to dec. 17

- **Lecture**	9h15-11h	(BC 01)
- **Exercises**	11h15-12h	(BC 01)
- **Project**	11h15-12h	(PH H331)

Evaluation

- MA students : project presentation + questions of theory

1.1.2 Expected work

- Active participation during the lecture
- Conscious work on the exercises (statistical and computing aspects)
- Fill in the gaps

1.1.3 Resources

- Moodle page with the slides and the exercises
- Textbook
- Internet and library

1.1.4 References

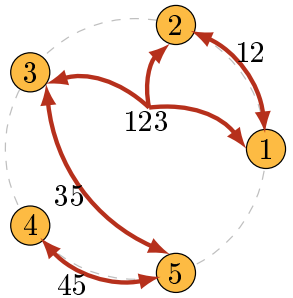
- Box, G., et al (2000), Statistics for experimenters, Wiley.
- Montgommery, D. (2009), Design and analysis of experiments, Wiley.
- Ryan,Th. (2007), Modern Experimental design, Wiley.
- Saltelli,A. (2000), Sensitivity analysis, Wiley.
- Lawson, J. (2014), Design and Analysis of Experiments with R, CRC press.

1.1.5 Softwares

- Matlab $\rightarrow$ the standard tool of the course
- Python $\rightarrow$ implies more personal work
- Excel $\rightarrow$ slower and a lot less powerful

1.2.1 Why learning DOE ?

- Nature answers questions in a very narrow way
- It is then key to question it with method
- DOE offer method to sharpen your experimental endeavor :
  - multifactorial approach
  - noise reduction strategy
  - taking interactions into account



Course chapters

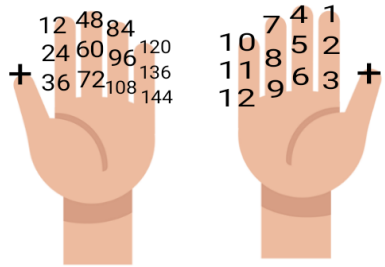
1. Introduction
2. Modeling
3. ANOVA
4. Factorial based designs
5. Response surface designs
6. Mixture designs
7. Qualitative factors

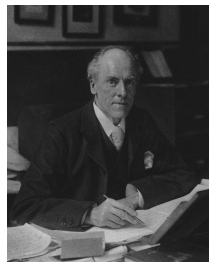
1.2.2 Pedagogical objectives of chapter 1

- Having a general orientation on the course
- Understand the origin of DOE
- Remember a few basic elements of statistics and data analysis
- Learning how to draw a mind map for a case
- Learning how to draw a causal model from data
- Training the basic operations of data analysis

1.2.3 A (very) brief time line of statistics

- Sumer (III millennium BC)
  - Livestock control (list, coding)
  - Prediction of sunrises and sunsets, tides and floods
- Aristotle (384-322 BC)
  - Things that change all the time can not be the objects of science
- René Descartes (1596-1650), Pierre de Fermat (1607-1665) et Blaise Pascal (1623-1662)
  - Theory of probability
- Thomas Bayes (1702-1761) et Pierre Simon de Laplace (1749-1827)
  - Conditional probability, prior vs posterior information
- Francis Galton (1822-1911), Karl Pearson (1857-1936) and Ronald Fisher (1890-1962)
  - Genetics, eugenics, correlations



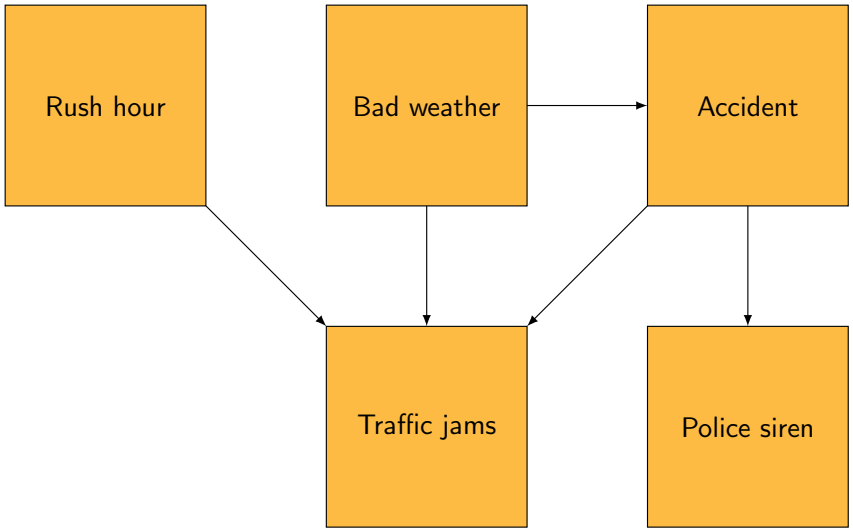


1.2.4 Sir Ronald Fisher (1890-1962)

- Agronomic station of Rothamsted, 1919
- Too frequent non-conclusive research
- Necessity of a collaboration between statisticians and experimenters
- Invention of ANOVA and DOE
- Statistical Methods for Research Workers (1925) ($\sim$ *Newton's Principia*)
- The Design of Experiments (1935)
- Founder of the neo-Darwinism and of the modern genetics

1.2.5 Causes, effects and contingencies

- Scientific experiment aims to determine relations between causes and effects (*Directed Acyclic Graph, Structural Causal Model*)
- Necessity to deal with confusion factors (see later)
- Necessity to deal with contingencies such as delays, costs, batches, security, ...

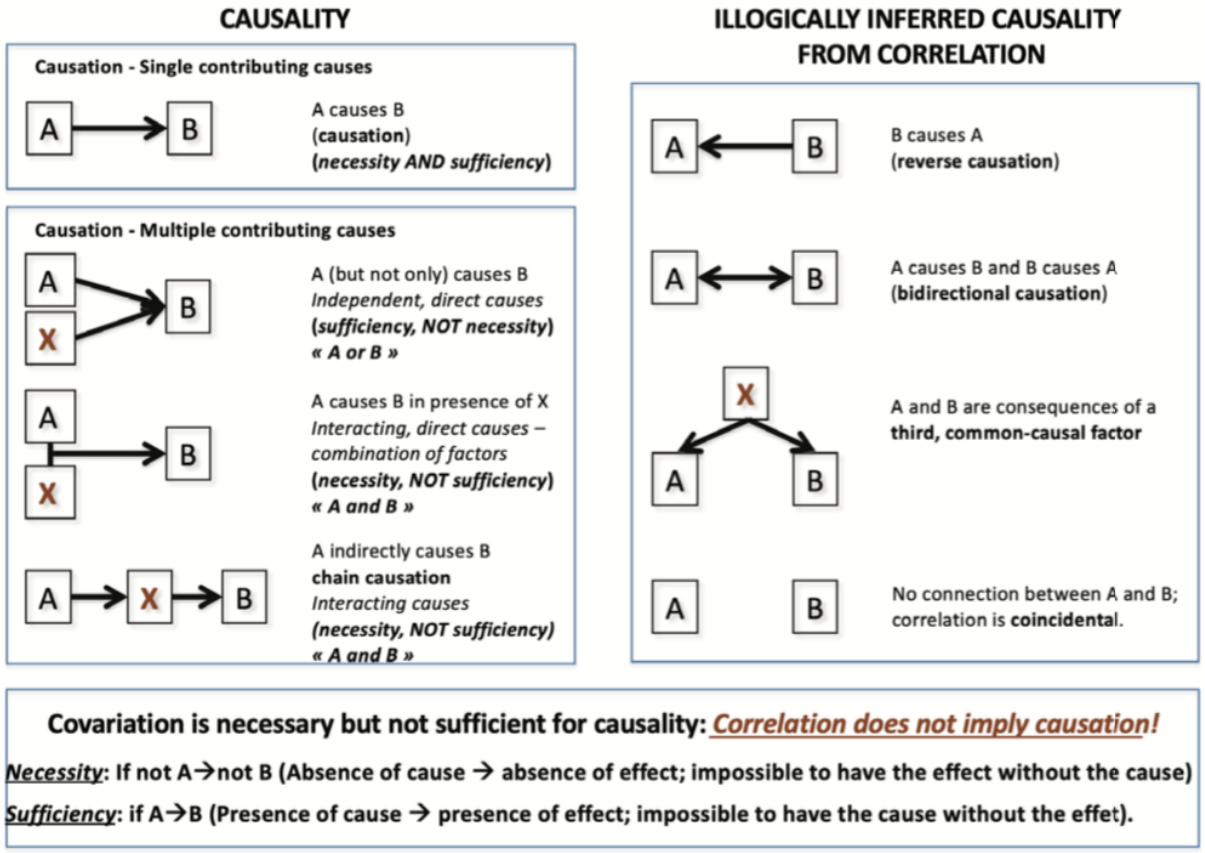


1.2.6 Untangling Cause and Effect

In nonlinear systems, the cause-effect relationship can be more complex due to several factors :

- **Nonlinearity in Response :**
  In linear systems, the response to an input is proportional. In nonlinear systems, the response can be disproportionate (e.g., doubling input might quadruple output).
- **Threshold Effects :**
  Nonlinear systems often have thresholds where behavior changes dramatically, making outcomes less predictable.
- **Feedback Loops :**
  Nonlinear systems may include feedback loops that amplify or dampen effects, complicating predictions.
- **Multiple Equilibria :**
  Multiple equilibrium points can lead to different outcomes for the same input depending on initial conditions.
- **Complex Interactions :**
  Nonlinear systems can exhibit chaos, where small changes in initial conditions lead to vastly different outcomes.

1.2.7 Causality



1.2.8 Correlation vs causality

Example	Correlation	Actual causation
Ice cream & sunburn	Positive correlation	Both caused by sunny weather (not each other)
Home work & grades	Correlation exists	More homework may improve grades (but confounds possible)
Smoking & lung cancer	Correlated	Smoking causally increases lung cancer risk
Solar panel & sunlight	Correlated	More sunlight $\rightarrow$ more energy (not vice versa)

1.3.1 What is the type of your data ?

- **Data** : Raw facts and figures collected through observations or measurements.
- **Types of Data** :
  - *Quantitative Data* : Numerical, can be discrete or continuous.
  - *Qualitative Data* : Categorical, can be nominal or ordinal.
- **Examples** : Test scores, survey responses, temperature readings.

1.3.1 Quantitative Input & Quantitative Response

Example : An energy engineer studies the relationship between outside temperature ($^\circ C$), insulation and household heating energy use (kWh/day).

Model :
Linear Regression

Model Formula

$$y = \beta_0 + \beta_T T + \beta_K K + \epsilon$$

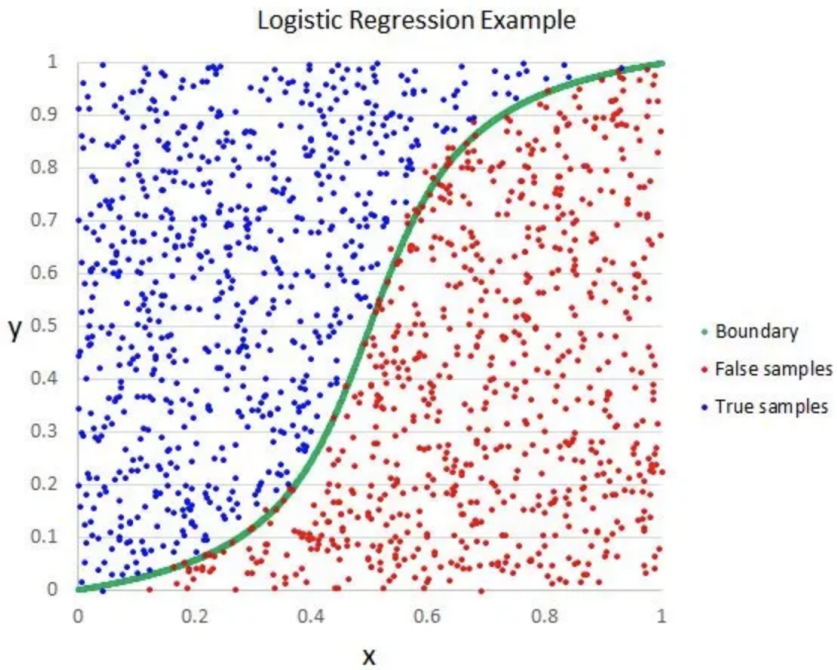
1.3.1 Quantitative Input & Qualitative Response

Example : Classifying patients as high-risk or low-risk based on blood pressure and cholesterol level.

Model :
Logistic Regression

Model Formula

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$



1.3.1 Qualitative Input & Qualitative Response

Example : A researcher wants to explore whether preferred commute mode (car, bike, public transport, walking) is associated with job type (office, factory, service, remote).

Model :
- Construct a contingency table
- Calculate row/column frequencies
- Apply a chi-squared test of independence to model the association.

Explanation

If significant, the model suggests that the type of job influences commuting preference.

1.3.1 Qualitative Input & Quantitive Response

Example : Modeling the quality of a product based on the categorical factors involved in it production

Model :
Constant coefficient model

Explanation

Find which categorical factors influence some characteristic of a product and which level of the factors produces the maximum or te minimum of the response.

$$Y = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

1.3.2 Modeling Based on Types of Data

Type of input	Type of response	Situation Description
Quantitative	Quantitative	**Regression**
Quantitative	Qualitative	Logistic regression
Qualitative	Quantitative	**Constant coefficient model**
Qualitative	Qualitative	Comparing frequencies between groups

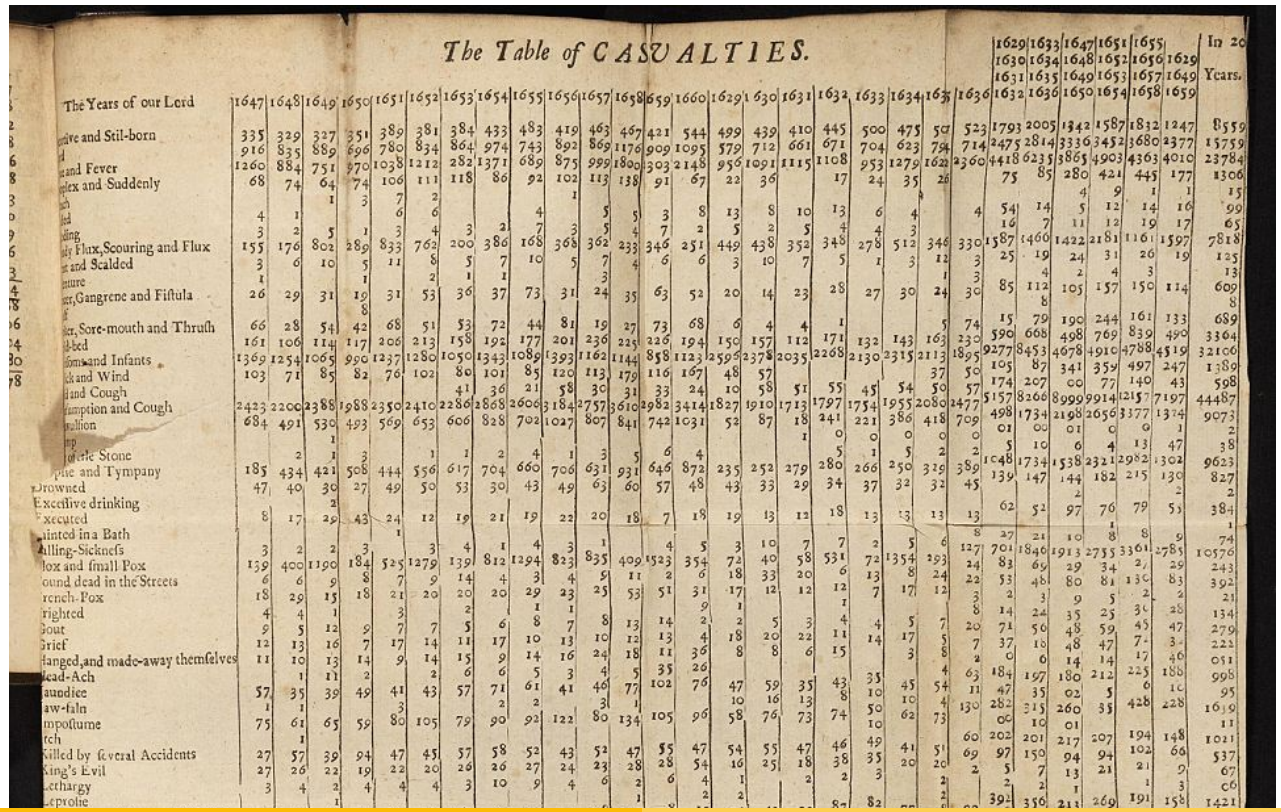
1.3.3 Importance of Meta-data

- **Meta-data** : Data about data ; provides context and additional information.
- **Purpose** :
  - Helps in understanding the context, quality, and limitations of the data.
  - Supports data management and interpretation.
- **Examples** : Date of data collection, methodology, units of measurement, source.

Graunt's book 1663 :

Natural and Political Observations Made upon the Bills of Mortality

- A historical milestone in applying data to understand society and health.
- An early example of quantitative reasoning.
- Evidence of the power of systematic observation to inform policy and science.



The Table of CASUALTIES.

The Years of our Lord	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1629	1630	1631	1632	1633	1634	1635	1636	1629 1630 1631 1632	1633 1634 1635 1636	1647 1648 1649 1650	1651 1652 1653 1654	1655 1656 1657 1658	1629 1649	In 20 Years.
...ive and Stil-born	335	329	327	351	389	381	384	433	483	419	463	467	421	544	499	439	410	445	500	475	587	523	1793	2005	1342	1587	1832	1247	8559
...d	916	835	889	696	780	834	864	974	743	892	869	1176	909	1095	579	712	661	671	704	623	794	714	2475	2814	3336	3452	3680	2377	15759
...and Fever	1260	884	751	970	1038	1212	282	1371	689	875	999	1800	303	2148	956	1091	1115	1108	953	1279	1622	2360	4418	6235	3865	4903	4363	4010	23784
...plex and Suddenly	68	74	64	74	106	111	118	86	92	102	113	138	91	67	22	36		17	24	35	26		75	85	280	421	445	177	1306
...ich			1	3	7	2			1																4	9	1	1	15
...ild	4	1			6	6			4		5	5	3	8	13	8	10	13	6	4		4	54	14	5	12	14	16	99
...ding	3	2	5	1	3	4	3	2	7	3	5	4	7	2	5	2	5	4	4	3			16	7	11	12	19	17	65
...ly Flux, Scouring and Flux	155	176	802	289	833	762	200	386	168	368	362	233	346	251	449	438	352	348	278	512	346	330	1587	1460	1422	2181	1161	1597	7818
...t and Scalded	3	6	10	5	11	8	5	7	10	5	7	4	6	6	3	10	7	5	1	3	12	3	25	19	24	31	26	19	125
...nture	1			1		2	1	1			3										1	3		4	2	4	3		13
...er, Gangrene and Fistula	26	29	31	19	31	53	36	37	73	31	24	35	63	52	20	14	23	28	27	30	24	30	85	112	105	157	150	114	609
...f				8																				8					8
...er, Sore-mouth and Thrush	66	28	54	42	68	51	53	72	44	81	19	27	73	68	6	4	4	1			5	74	15	79	190	244	161	133	689
...d-bed	161	106	114	117	206	213	158	192	177	201	236	225	226	194	150	157	112	171	132	143	163	230	590	668	498	769	839	490	3364
...ions and Infants	1369	1254	1065	990	1237	1280	1050	1343	1089	1393	1162	1144	858	1123	2595	2378	2035	2268	2130	2315	2113	1895	9277	8453	4678	4910	4788	4519	32106
...k and Wind	103	71	85	82	76	102	80	101	85	120	113	179	116	167	48	57					37	50	105	87	341	359	497	247	1389
...d and Cough							41	36	21	58	30	31	33	24	10	58	51	55	45	54	50	57	174	207	60	77	140	43	598
...umption and Cough	2423	2200	2388	1988	2350	2410	2286	2868	2600	3184	2757	3610	2982	3414	1827	1910	1713	1797	1754	1955	2080	2477	5157	8266	8999	9914	1215	7197	44487
...sullion	684	491	530	493	569	653	606	828	702	1027	807	841	742	1031	52	87	18	241	221	386	418	709	498	1734	2198	2656	3377	1374	9073
...p			1														1	0	0	0	0	0	01	00	01	0	0	1	2
...of the Stone		2	1	3		1	1	2	4	1	3	5	6	4				5	1	5	2	2	5	10	6	4	13	47	38
...pine and Tympany	185	434	421	508	444	556	617	704	660	706	631	931	646	872	235	252	279	280	266	250	329	389	1648	1734	1538	2321	2982	1302	9623
...Drowned	47	40	30	27	49	50	53	30	43	49	63	60	57	48	43	33	29	34	37	32	32	45	139	147	144	182	215	130	827
...Excessive drinking			2																									2	2
...Executed	8	17	29	43	24	12	19	21	19	22	20	18	7	18	19	13	12	18	13	13	13	13	62	52	97	76	79	53	384
...ainted in a Bath					1																	8	27	21	10	8	8	9	1
...alling-Sickneſs	3	2	2	3		3	4	1	4	3	1		4	5	3	10	7	7	2	5	6	127	701	1846	1913	2735	3361	2785	10576
...Pox and ſmall Pox	139	400	1190	184	525	1279	139	812	1294	823	835	409	1523	354	72	40	58	531	72	1354	293	24	83	69	29	34	2	29	243
...found dead in the Streets	6	6	9	8	7	9	14	4	3	4	5	11	2	6	18	33	20	6	13	8	24	22	53	46	80	81	130	83	392
...French-Pox	18	29	15	18	21	20	20	20	29	23	25	53	51	31	17	12	12	12	7	17	12	3	2	3	9	5	2	2	21
...Frighted	4	4	1		3		2		1	1				9	1			1				8	14	24	35	25	30	28	134
...Gout	9	5	12	9	7	7	5	6	8	7	8	13	14	2	2	5	3	4	4	5	7	20	71	56	48	59	45	47	279
...Grief	12	13	16	7	17	14	11	17	10	13	10	12	13	4	18	20	22	11	14	17	5	7	37	18	48	47	74	3	222
...Hanged, and made-away themſelves	11	10	13	14	9	14	15	9	14	16	24	18	11	36	8	8	6	15		3	8	2	0	6	14	14	17	4	011
...Head-Ach		1	11	2		2	6	6	5	3	4	5	35	26							4	63	184	197	180	212	225	188	998
...Jaundice	57	35	39	49	41	43	57	71	61	41	46	77	102	76	47	59	35	43	35	45	54	11	47	35	92	5	428	228	95
...Jaw-faln	1	1		3		2		2		3	1				10	16	13	8	10	10	4	130	282	315	260	35			1619
...Impoſtume	75	61	65	59	80	105	79	90	92	122	80	134	105	96	58	76	73	74	50	62	73	60	00	10	01				11
...Itch		1																	10				202	201	217	207	194	148	1021
...Killed by ſeveral Accidents	27	57	39	94	47	45	57	58	52	43	52	47	55	47	54	55	47	46	49	41	51	69	97	150	94	94	102	66	537
...King's Evil	27	26	22	19	22	20	26	26	27	24	23	28	28	54	16	25	18	38	35	20	20	2	5	7	13	21	21	9	67
...Lethargy	3	4	2	4	4	4	3	10	9	4	6	2	6	4	1		2	2	3		2		2	2	1		1	3	06
...Leproſie			1									1		2							2		392	350	213	269	191	158	1421

1.3.4 Categorization : Titanic passengers

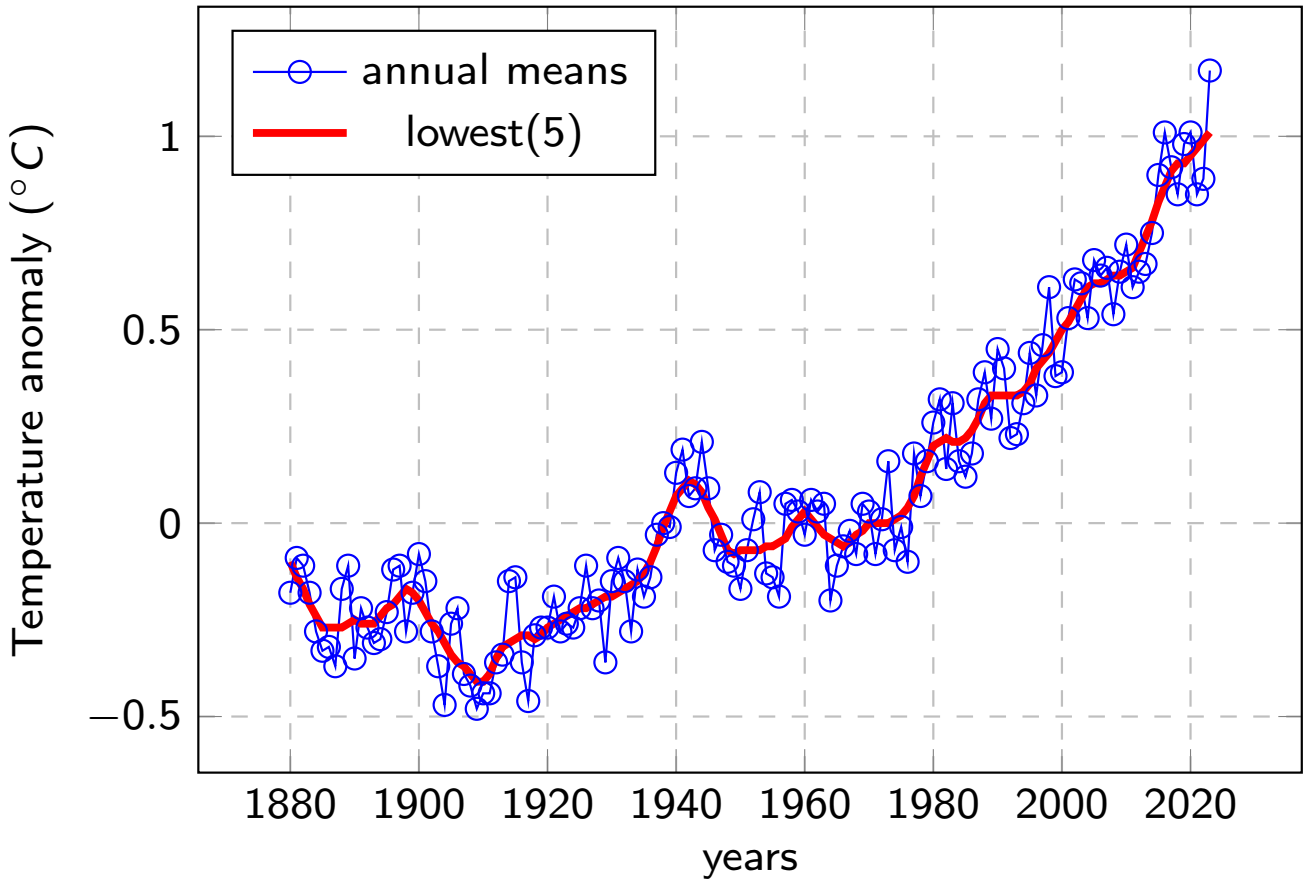
		Dead		Survivors	
Age	Class	M	F	M	F
Children	1st	0	0	5	1
	2nd	0	0	11	13
	3rd	35	17	13	14
	Crew	0	0	0	0
Adults	1st	118	4	57	140
	2nd	154	13	14	80
	3rd	387	89	75	76
	Crew	670	3	192	20

"[...] Those methods deal comprehensively with entire species, and with entire groups of influences, just as if they were single entities, and express the relations between them in an equally compendious manner. They commence by marshalling the values in order of magnitude from the smallest up to the largest, thereby converting a mob into an orderly array, which like a regiment thenceforth becomes a tactical unit. ", F. Galton, Biometrika, Volume 1, Issue 1, October 1901

1.3.5 Data : Earth Surface Temperature (NASA)

year	crude	lowest(5)	year	crude	lowest(5)	year	crude	lowest(5)
1901	-0.15	-0.23	1951	-0.07	-0.07	2001	0.53	0.52
1902	-0.28	-0.26	1952	0.01	-0.07	2002	0.63	0.55
1903	-0.37	-0.28	1953	0.08	-0.07	2003	0.62	0.58
1904	-0.47	-0.31	1954	-0.13	-0.06	2004	0.53	0.61
1905	-0.26	-0.34	1955	-0.14	-0.06	2005	0.68	0.62
1906	-0.22	-0.36	1956	-0.19	-0.05	2006	0.64	0.62
1907	-0.39	-0.37	1957	0.05	-0.04	2007	0.66	0.63
1908	-0.42	-0.39	1958	0.06	-0.01	2008	0.54	0.64
1909	-0.48	-0.41	1959	0.03	0.01	2009	0.65	0.64
1910	-0.44	-0.41	1960	-0.03	0.03	2010	0.72	0.65
1911	-0.44	-0.39	1961	0.06	0.01	2011	0.61	0.66
1912	-0.36	-0.35	1962	0.03	-0.01	2012	0.65	0.7
1913	-0.34	-0.32	1963	0.05	-0.03	2013	0.67	0.74
1914	-0.15	-0.31	1964	-0.2	-0.04	2014	0.75	0.78
1915	-0.14	-0.3	1965	-0.11	-0.05	2015	0.9	0.83
1916	-0.36	-0.29	1966	-0.06	-0.06	2016	1.01	0.87
1917	-0.46	-0.29	1967	-0.02	-0.05	2017	0.92	0.91
1918	-0.29	-0.3	1968	-0.08	-0.03	2018	0.85	0.93
1919	-0.27	-0.29	1969	0.05	-0.02	2019	0.98	0.93
1920	-0.27	-0.27	1970	0.03	0	2020	1.01	0.95
1921	-0.19	-0.26	1971	-0.08	0	2021	0.85	0.97
1922	-0.28	-0.25	1972	0.01	0	2022	0.89	0.99
1923	-0.26	-0.24	1973	0.16	0	2023	1.17	1.01
1924	-0.27	-0.23	1974	-0.07	0.01			
1925	-0.22	-0.22	1975	-0.01	0.02			
...	...	...	...	...	...			

1.3.6 A picture is worth a thousand words



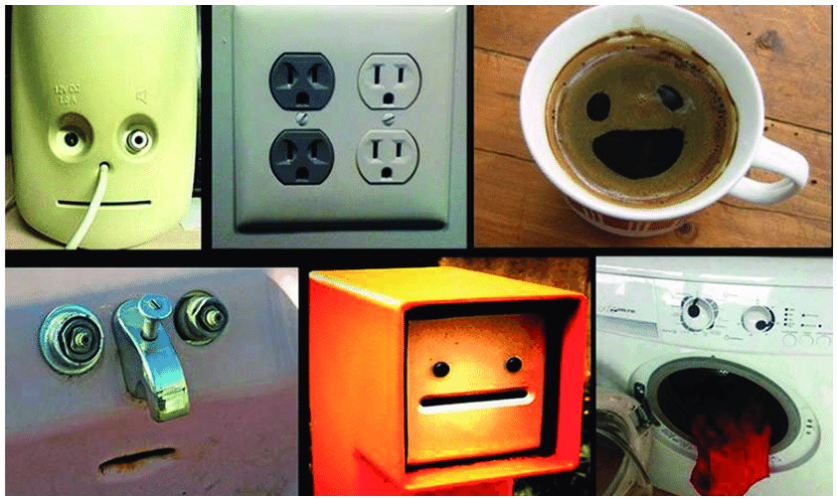
1.3.7 Importance and Risks

- **Importance :**
  - Categorizing data or creating visual representations helps in **identifying patterns**, trends, and relationships.
  - Graphs make complex data more accessible and **easier to interpret**.
  - Visual **summaries** provide a quick overview of data distribution and key features.

- **Risks of Bias :**
  - Categorization can **oversimplify** important variations within data.
  - Choice of graph type, scale, or binning can **mislead** or distort the true nature of data.
  - Visual representations might **exaggerate or minimize effects**, leading to misinterpretations.

- **Key Takeaway :**
  - While categorization and graphing are powerful tools, analysts must remain mindful of **biases** that can affect data interpretation.

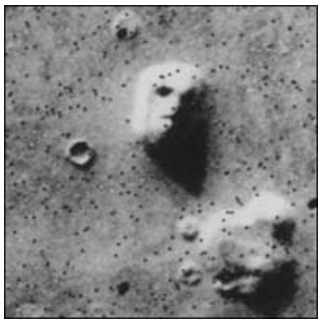
1.3.8 Pareidolia



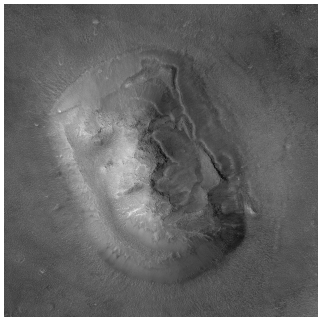
Our brain has the tendency to see motives in random sets as well as a tendency to make a script, which means to tell a story, and give a meaning to an image. Statistics offers tools, like statistical tests, to pass the perception.

1.3.9 Cydonia Mensae

- The « Mars' face » put in evidence the influence of the resolution of a numerical image over the interpretations than is given of it.
- The principle of parsimony of the hypotheses : it is less costly in hypotheses to consider that our brain looks for and finds a meaning to this picture, recognizing a human face in a picture, than to consider the existence of an extraterrestrial civilization who would have built a huge construction representing a human face.



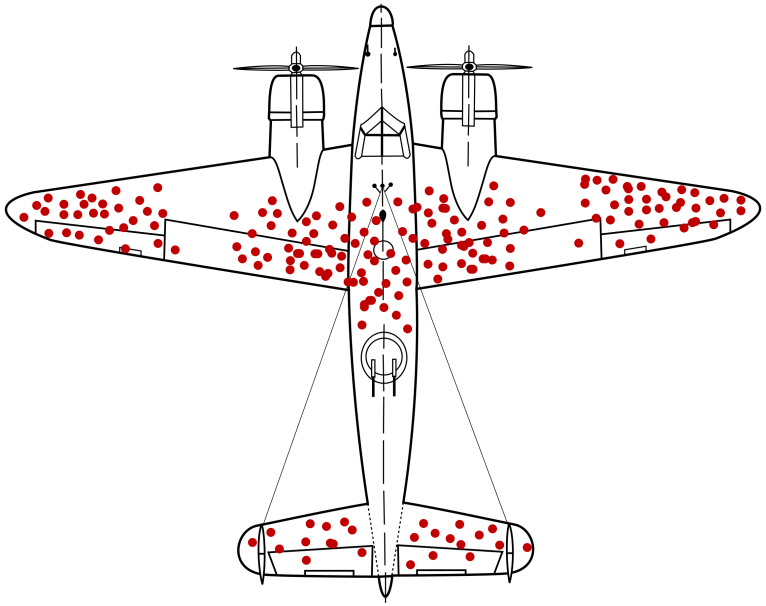
Viking (1971)



Mars global surveyor (2001)

https://fr.wikipedia.org/wiki/Cydonia_Mensae

1.3.10 Survivorship bias



The damaged portions of returning planes show locations where they can sustain damage and still return home ; those hit in other places presumably do not survive. (Image shows hypothetical data.) The error was to consider that the planes are *more probably* hit in those places and then reinforce them.

1.3.11 Pitfalls in data analysis

Cognitive biases

- of attention
- of memory
- of judgment
- of reasoning

Examples

- Pareidolia
- Confusion correlation/causation
- Cherry picking
- Error of attribution
- Barnum effect (experience of Forer)

`https://en.wikipedia.org/wiki/List_of_cognitive_biases`

1.3.12 Falsifiability (Popper)

- A statement is said refutable (falsifiable) if and only if it can be logically contradicts by an empirical test. ... more precisely if and only if it exists a possible observation statement (true or false) that would logically contradict the theory.
- To be accepted as scientific a statement, a theory must be falsifiable.

1.3.13 Simpson paradox

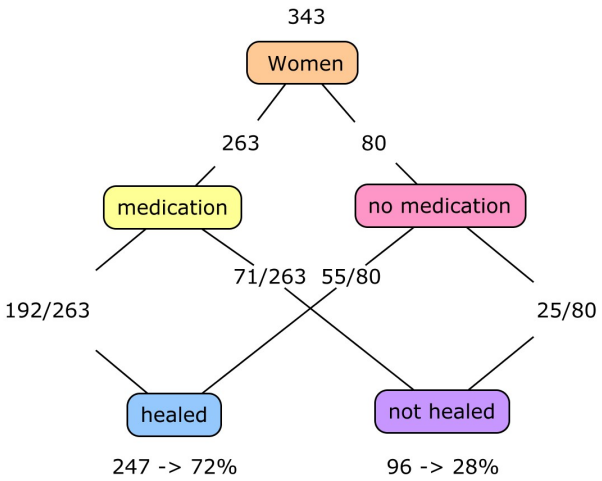
The **fraction of healing** of 700 patients with and without the use of a given drug. 350 individuals with the medication (group "with"), 350 without the medication (group "without").

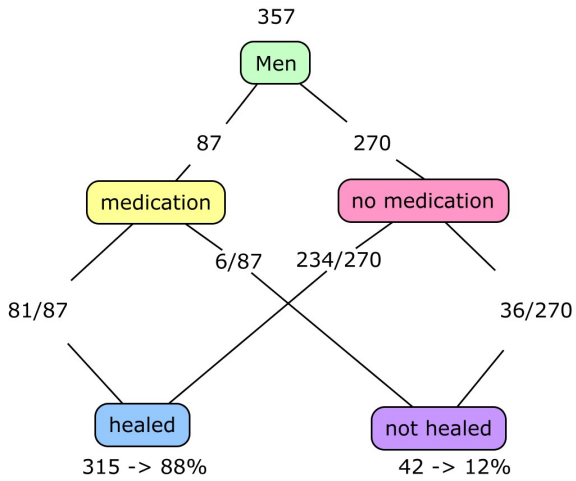
	With	**Without**
Men	81/87	234/270
Women	192/263	55/80
All	273/350	289/350

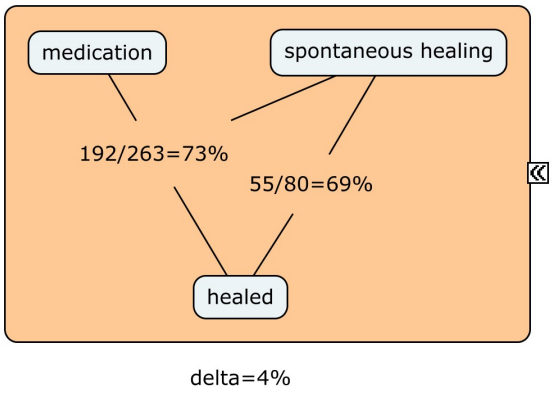
	With	**Without**
Men	93%	87%
Women	72%	69%
All	78%	83%

Is the drug to be proposed or not ?

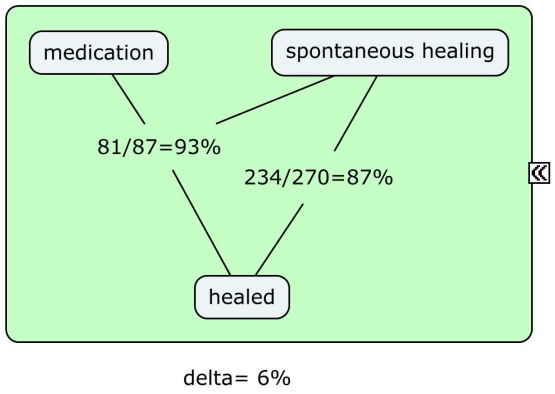
1.3.14 Cause analysis







delta=4%



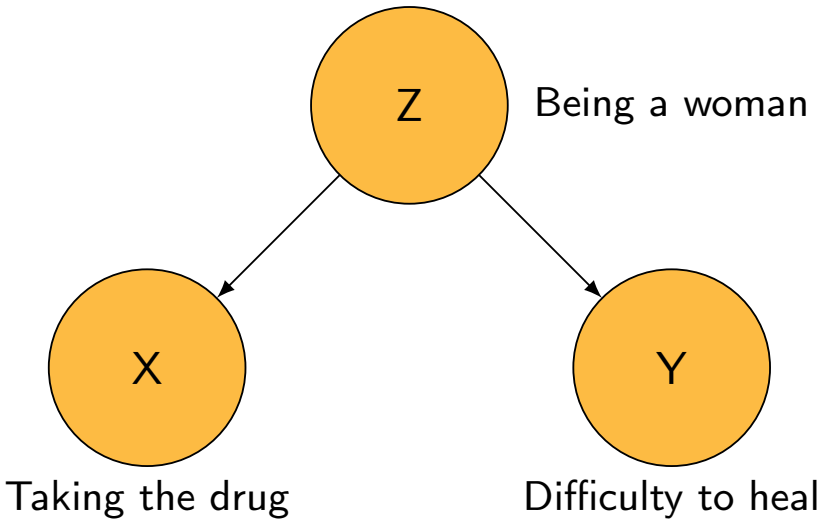
delta= 6%

1.3.15 Analysis

- If **taking the drug** appears less effective than **doing nothing**, it is because if we randomly choose a person who has taken the drug, the probability is high that this person is a woman and therefore that his recovery is less easy than a person chosen at random from those who did not use the drug, in which case the probability is greater that it is a man and that he recovered more easily.
- In the example, **being a woman** is a common cause **to take the medicine** and **to heal less easily**.
- In this case, the aggregated data does not allow to really test the effectiveness of the drug.
- To test the efficacy of the drug **a homogeneous group** is needed so that the difference in cure rate is attributable to the drug alone and not to the effect of estrogen.

1.3.16 Confounding factor

A confounding factor, (or extraneous determinant or lurking variable) is a variable that influences both the dependent variable and independent variable, causing a **spurious association**. Confounding is a causal concept, and as such, cannot be described in terms of correlations or associations



1.3.17 Data about scientific publications

	Country	Code	Region	Income	Level	2 000	2 018	pop2000	pop2018
2	Afghanistan	AFG	South Asia	Low income	1	4	112	20 779 957	37 171 922
3	Angola	AGO	Sub-Saharan Africa	Lower middle income	2	7	30	16 395 477	30 809 787
4	Albania	ALB	Europe & Central Asia	Upper middle income	3	22	180	3 089 027	2 866 376
5	Andorra	AND	Europe & Central Asia	High income	4	-	4	65 390	77 008
6	United Arab Emirates	ARE	Middle East & North Africa	High income	4	330	3 145	3 134 067	9 630 966
7	Argentina	ARG	Latin America & Caribbean	Upper middle income	3	4 386	8 811	36 870 796	44 494 502
8	Armenia	ARM	Europe & Central Asia	Upper middle income	3	346	521	3 069 597	2 951 741
9	Antigua and Barbuda	ATG	Latin America & Caribbean	High income	4	0	6	76 007	96 282
10	Australia	AUS	East Asia & Pacific	High income	4	23 276	53 610	19 153 000	24 982 688
11	Austria	AUT	Europe & Central Asia	High income	4	6 577	12 362	8 011 566	8 840 521
12	Azerbaijan	AZE	Europe & Central Asia	Upper middle income	3	155	761	8 048 600	9 939 771
13	Burundi	BDI	Sub-Saharan Africa	Low income	1	2	21	6 378 871	11 175 379
14	Belgium	BEL	Europe & Central Asia	High income	4	9 723	15 688	10 251 250	11 427 054
15	Benin	BEN	Sub-Saharan Africa	Lower middle income	2	43	228	6 865 946	11 485 035
16	Burkina Faso	BFA	Sub-Saharan Africa	Low income	1	48	252	11 607 951	19 751 466
17	Bangladesh	BGD	South Asia	Lower middle income	2	440	3 135	127 657 862	161 376 713
18	Bulgaria	BGR	Europe & Central Asia	Upper middle income	3	1 653	3 311	8 170 172	7 025 037
19	Bahrain	BHR	Middle East & North Africa	High income	4	83	322	664 610	1 569 440
20	Bahamas, The	BHS	Latin America & Caribbean	High income	4	2	20	298 045	385 635
21	Bosnia and Herzegovina	BIH	Europe & Central Asia	Upper middle income	3	75	704	3 751 176	3 323 929
22	Belarus	BLR	Europe & Central Asia	Upper middle income	3	1 170	1 180	9 979 610	9 483 499
23	Belize	BLZ	Latin America & Caribbean	Lower middle income	2	1	9	247 310	383 071
24	Bolivia	BOL	Latin America & Caribbean	Lower middle income	2	39	103	8 418 270	11 353 140
25	Brazil	BRA	Latin America & Caribbean	Upper middle income	3	12 783	60 148	174 790 339	209 469 320

https://data.worldbank.org/indicator/IP.JRN.ARTC.SC?end=2018&start=2000

1.3.18 Loading data

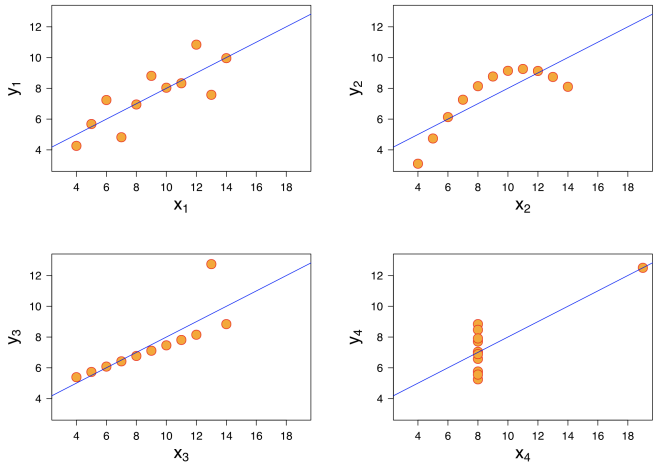
Matlab has specialized routines to load data in the workspace

MATLAB

- *X=0 :1 :20* creates the vector $X = [0, 1, 2, \ldots, 20]$
- *X=[. . .]* creates a vector with specific values
- *Y=repmat(X,h,w)* creates a matrix copying X, h times vertically w times horizontally
- *T = table(var1, ..., varN, Name, Value)* creates a table
- *T = readtable(FileName.xls','Sheet',SheetName)* loads a table from Excel

1.3.19 ANSCOMBE'S QUARTET

- Four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
- Each dataset consists of eleven (x,y) points
- Constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it
- Intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

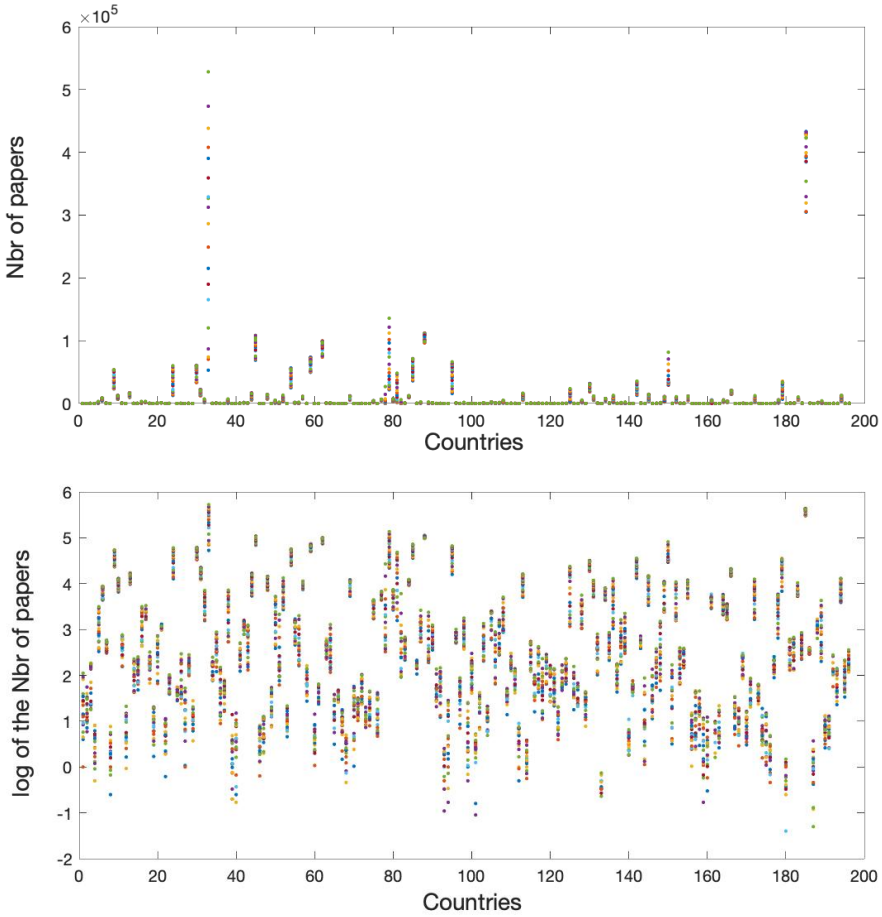


1.3.20 Visual analysis

- To detect patterns, aberrations, etc.
- The change of the metric is also interesting

Matlab

- *plot(x, y)*
  plot(x, y, LineSpec)
  plot(x_1, y_1,. . ., x_n, y_n)
  plot(. . .,Name,Value)
- *bar(x, y)*
  bar(. . ., Width)
  bar(. . ., Style)
  bar(. . ., Name, Value)



1.3.21 Sorting data

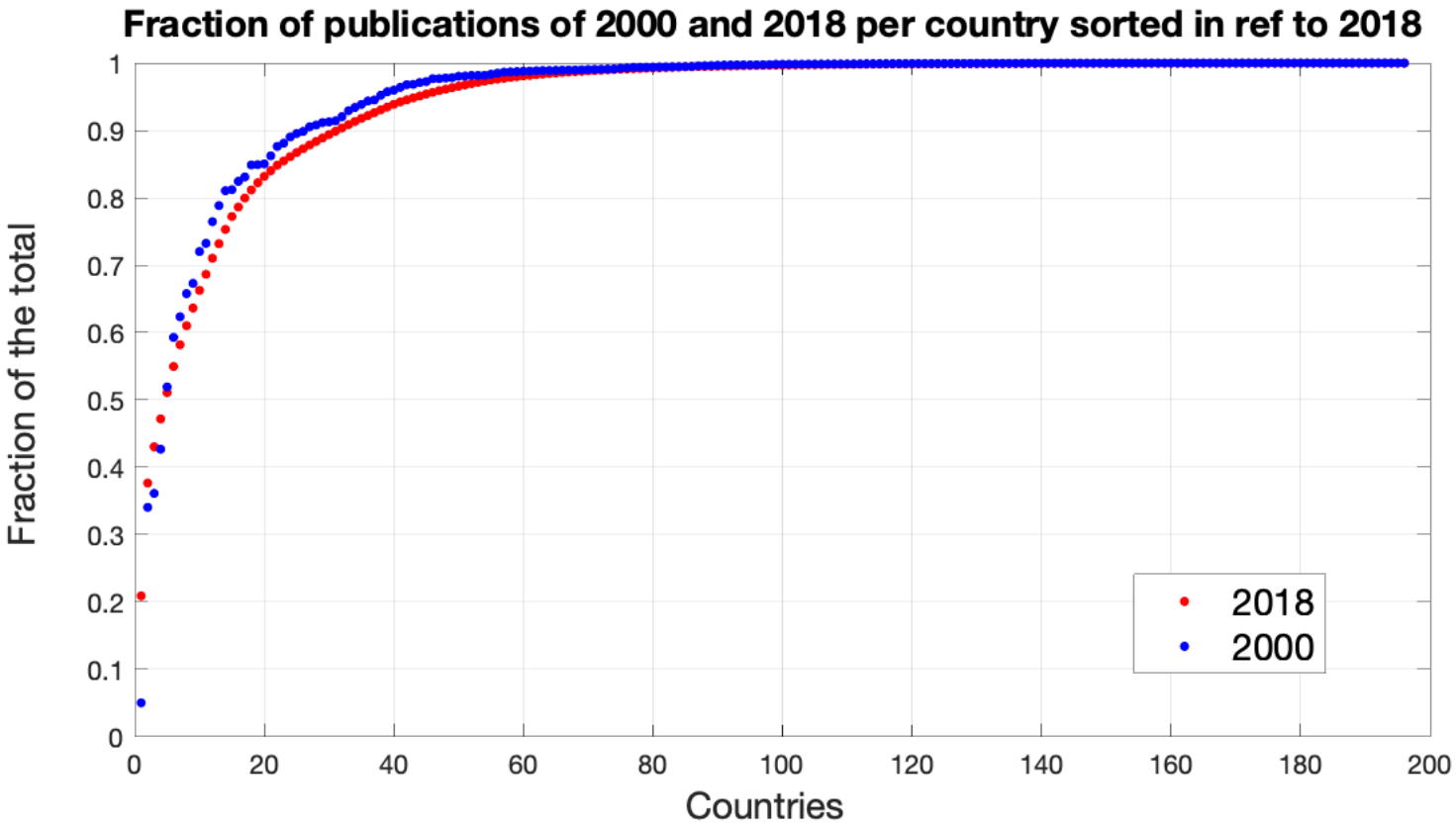
$$x_1\ x_2\ \ldots x_N \quad \Rightarrow \quad x_{(1)} < x_{(2)} < \cdots < x_{(N)}$$

MATLAB

- *[B,Index]=sort(A,dim,direction)*
- *[B,Index]=sortrows(A,col,direction)*
- *[tblB,Index]=sortrows(tblA,col,direction)*

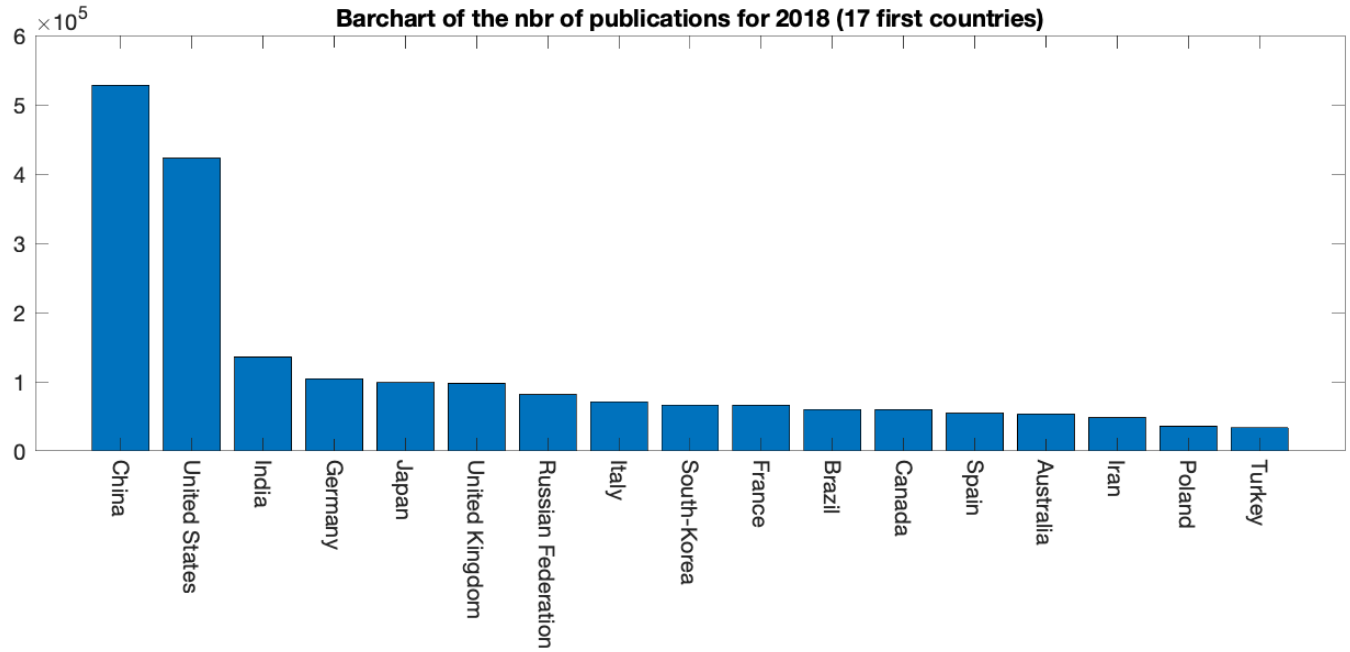
A, B : matrices
tblA, tblB : table
dim : dimension to realize the sorting (1,2, ...)
col : column of reference for sorting
direction : 'ascend' or 'descend'

1.3.22 Plot of the sorted data



1.3.23 Pareto principle

80% -20% principle : A maximum of things are related to a minimum of causes



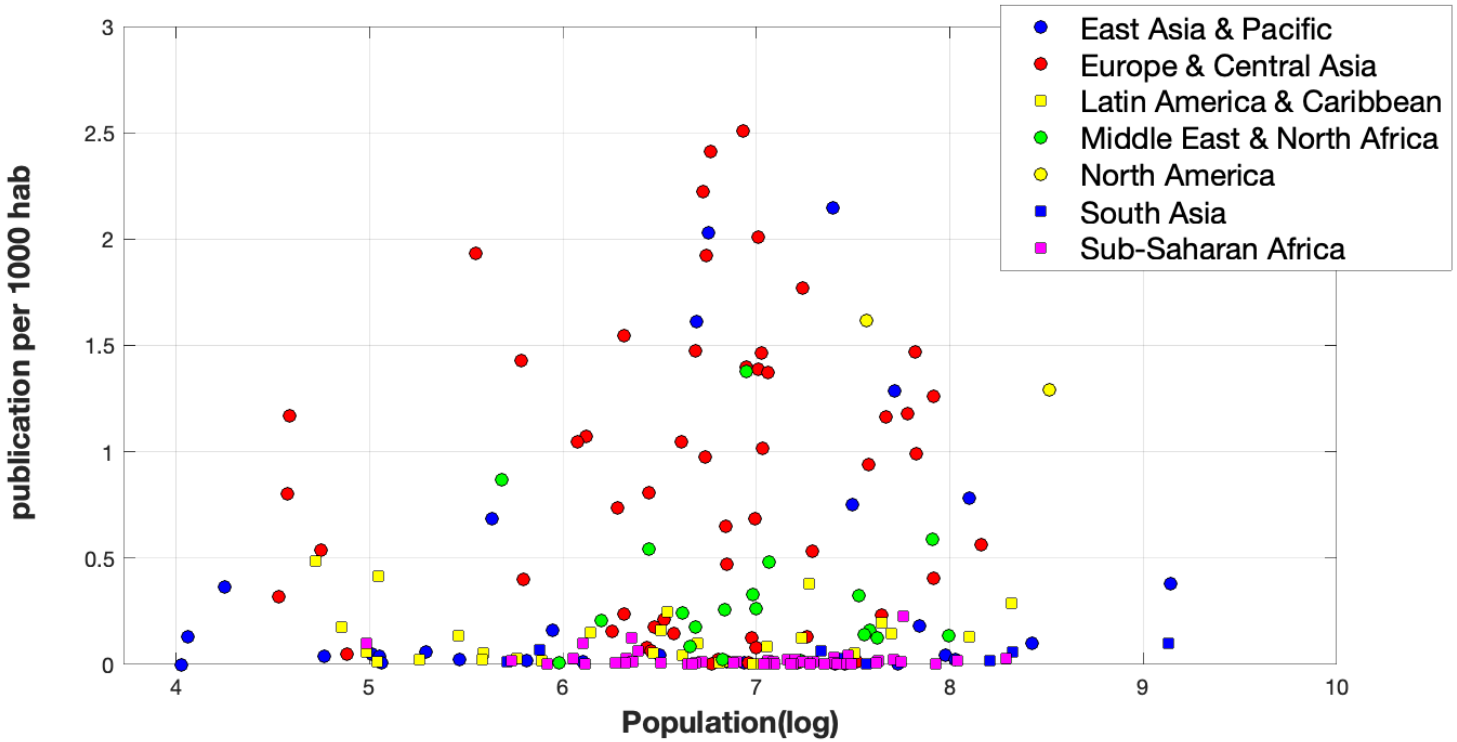
1.3.24 Dealing with categories

Defining a column of a table as *categorical* allows you to perform some computation by categories

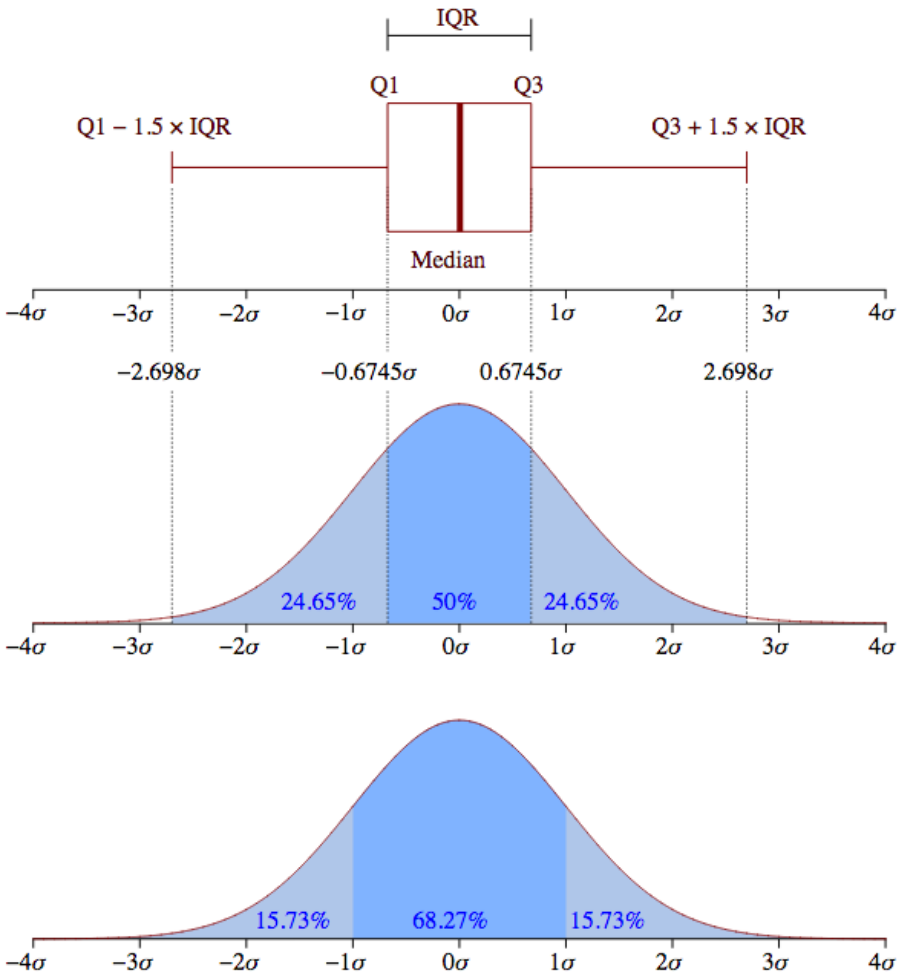
MATLAB

- *B=categorical(A)*
  tbl.var = categorical(tbl.var)
- *statarray = grpstats(tbl,group,stats)*
  stats=grpstats(tbl,'var1',{ 'min','max' },'tblVars','var2')
- *gscatter(x,y,group)*
  A,B : array or column of a table
  tbl : table
  group : the variable(s) of the groupe
  stats : statistics to compute such as min, max, etc.

1.3.25 Scatter plot

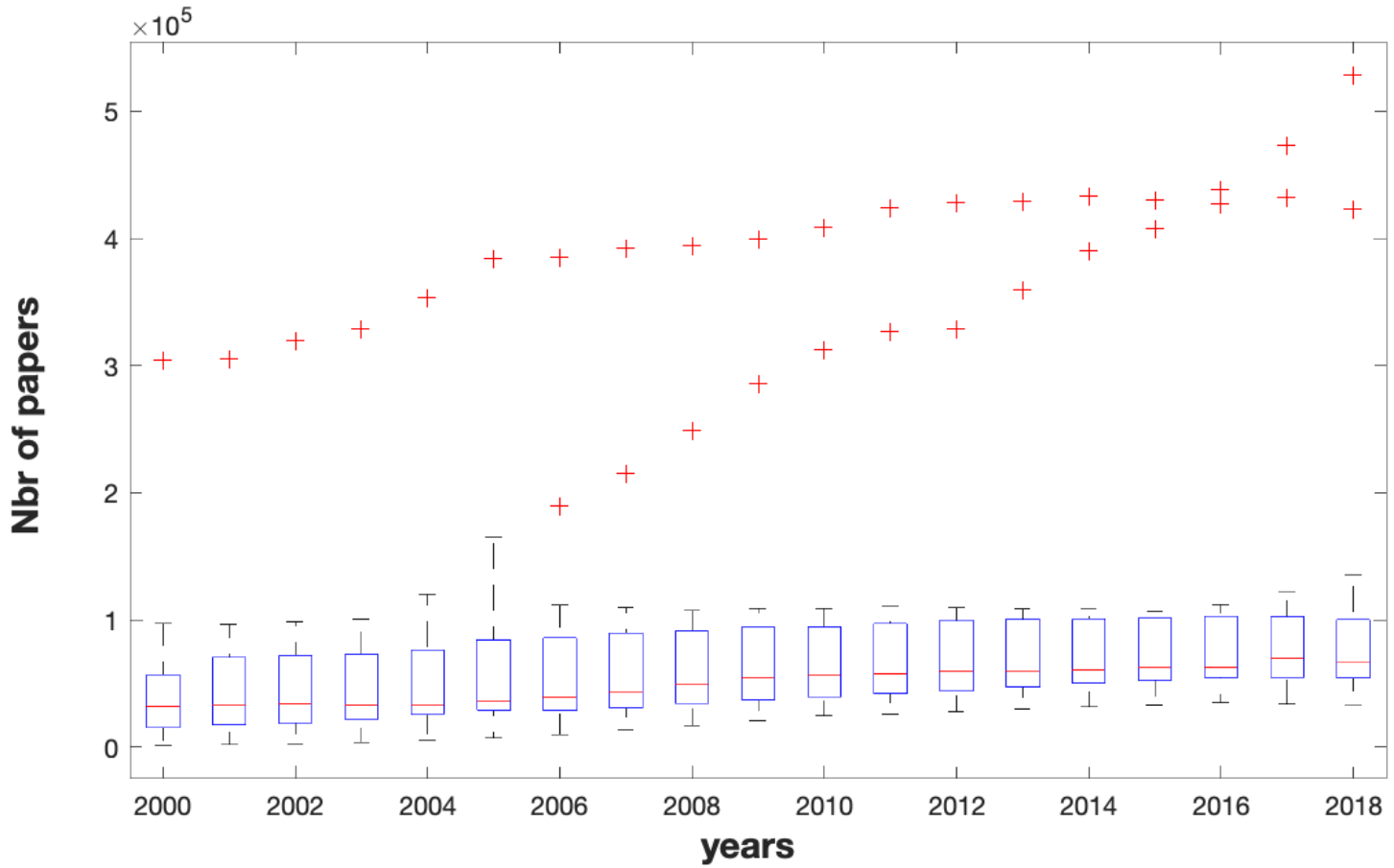


1.3.26 Box-plot (definition)



- *Outliers* are points placed at more than 1.5 IQR at the left of Q_1 or at the right of Q_3
- *Wiskers* are drawn at the minimum or maximum of the data points after the exclusion of the outliers

1.3.27 Box-plot



1.3.28 Location and range

$$x_{(1)}, x_{(2)}, \ldots, x_{(N/4)}, \ldots, x_{(N/2)}, \ldots, x_{(3N/4)}, x_{(N-1)}, x_{(N)}$$

- Median
- Average
- Range
- Standard deviation
- Variance
- Quartile
- Percentile

MATLAB

$M = median(A)$
$M = mean(A)$
$R = range(A)$
$S = std(A)$
$v = var(A)$
$Y = quantile(X,p)$

1.3.28 Location and range

$$m_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(N number of data point)

$$var(x) = s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$$

$$M = \begin{cases} \frac{1}{2}(x_{(N/2)} + x_{(N/2+1)}) \\ x_{((N+1)/2)} \end{cases}$$

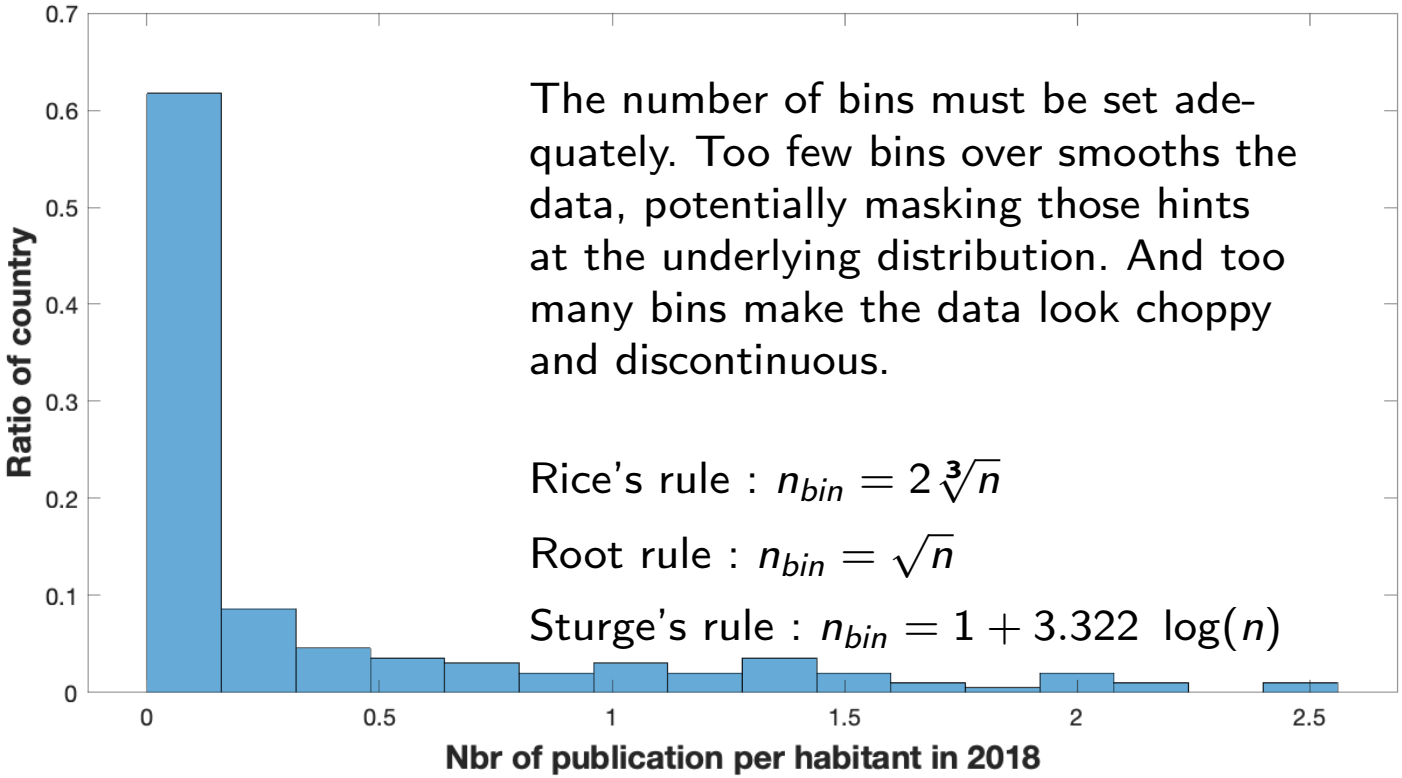
$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2}$$

$$R = x_{(N)} - x_{(1)}$$

1.3.29 Quartiles

	Mean	Deviation	Variance	Q25	Median	Q75
2000	0.18	0.36	0.13	0.0025	0.018	0.11
2018	0.37	0.57	0.33	0.0128	0.08	0.48

1.3.30 Histogram



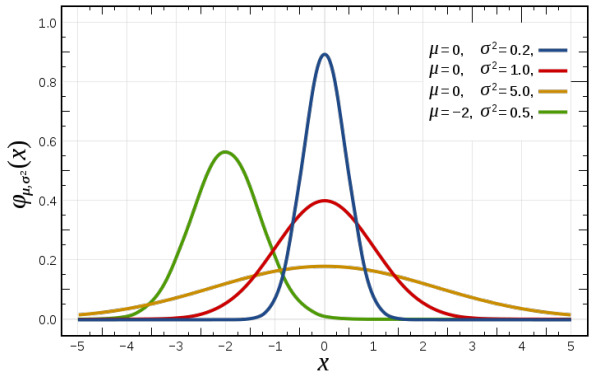
The number of bins must be set adequately. Too few bins over smooths the data, potentially masking those hints at the underlying distribution. And too many bins make the data look choppy and discontinuous.

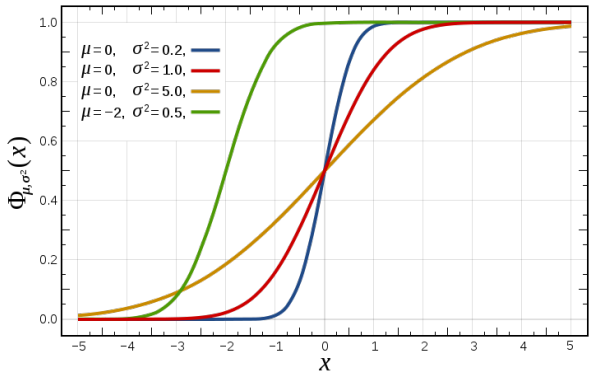
Rice's rule : $n_{bin} = 2\sqrt[3]{n}$

Root rule : $n_{bin} = \sqrt{n}$

Sturge's rule : $n_{bin} = 1 + 3.322 \ \log(n)$

1.3.31 Normal distribution $\mathrm{N}(\nu, \sigma)$





$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The Normal distribution is characterized by its bell-shaped curve, which is symmetric around its mean. It describes how values of a variable are distributed, with most of the data clustering around the central mean μ and decreasing in frequency as they move further away.
The spread of the distribution is determined by its standard deviation, σ.
The Normal distribution is fundamental due to its natural occurrence in many real-world phenomena and its properties, such as the central limit theorem, which states that the sum of a large number of random variables tends to follow a normal distribution, regardless of the original distribution of the variables.

1.3.31 Normal distribution $\mathbf{N}(\nu, \sigma)$

MATLAB

- Random number generation
  X=randn(Ni,Nj)
- Probability density function
  p=pdf('Normal',x,mu,sigma)
- Cumulative density function
  p=cdf('Normal',x,mu,sigma)
- Inverse cumulative distribution function
  x=icdf('Normal',p,mu,sigma)

1.3.32 Data vs distributions

- Original data : $Y_i \sim N(\mu_i, \sigma_i)$ avec $0 \leqslant i \leqslant n$
- Linear : $a_j = \sum_i x_{ij} Y_i \sim N\left(\mu = \sum x_{ij}\mu_i,\ \sigma = \sqrt{\sum x_{ij}\sigma_i^2}\right)$
- Average : $\sqrt{n-1}\left(\frac{\bar{Y}-\mu}{s}\right) \sim T(\nu)$
- Quadratic function : $(a_j)^2 \sim \chi^2(w_j)$
- Quadratic function quotient : $\frac{(a_j)^2}{(a_i)^2} \sim F(w_j, w_i)$

1.3.33 Student's T_ν distribution and CI

- Published by William Gosset, 1908, (Guinness)

  If the observations X_i are IID [1] then $\left(\frac{\overline{X}-\mu}{s/\sqrt{n}}\right) \sim T_{n-1}$

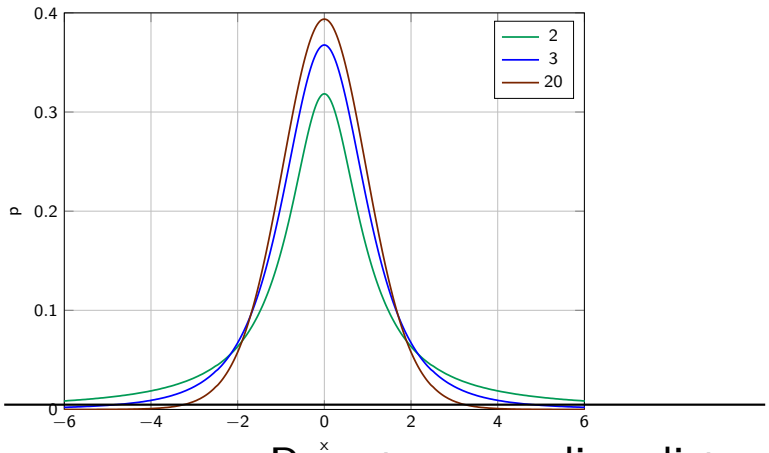
- Confidence interval theorem :

$$P\left(x \in \left[\bar{x} - t^{n-1}_{\alpha/2}\sqrt{\frac{s^2}{n}}\ ,\ \bar{x} + t^{n-1}_{\alpha/2}\sqrt{\frac{s^2}{n}}\right]\right) = 1-\alpha$$

  with t^ν_α the value of t at which $\int_0^t T_\nu(t')\,dt' = \alpha$

1. independent and identically distributed

1.34 The Student's T_ν distribution



	Parent dist.	sampling dist. for $\bar{y}$
Mean	η	η
Variance	σ^2	$\frac{\sigma^2}{n}$
Std dev.	σ	$\frac{\sigma}{\sqrt{n}}$
Form	$\sim$ any	more nearly Normal

The Student's t-distribution is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

The t-distribution is parameterized by degrees of freedom $\nu = n - p$ (n being the sample size and p, the number of parameters) ; as the degrees of freedom increase, the t-distribution approaches the normal distribution.

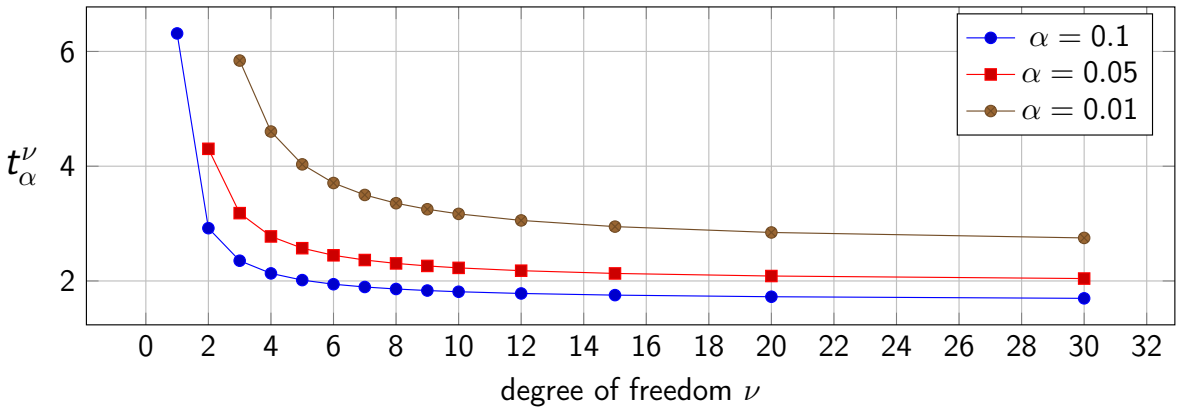
It is commonly used in hypothesis testing to determine if there is a significant difference between sample means.

1.3.35 Compute with the Student's distribution

MATLAB

- Generation of random number following t_ν
  X=trnd(nu,Ni,Nj)
- Probability density function of t_ν
  p=tpdf(x,nu)
- Cumulative density function of t_ν
  p=tcdf(x,nu) *p=tcdf(x,nu,'upper')*
- Inverse cumulative distribution function :
  x=tinv(p,nu)

1.3.36 Confidence and degrees of freedom

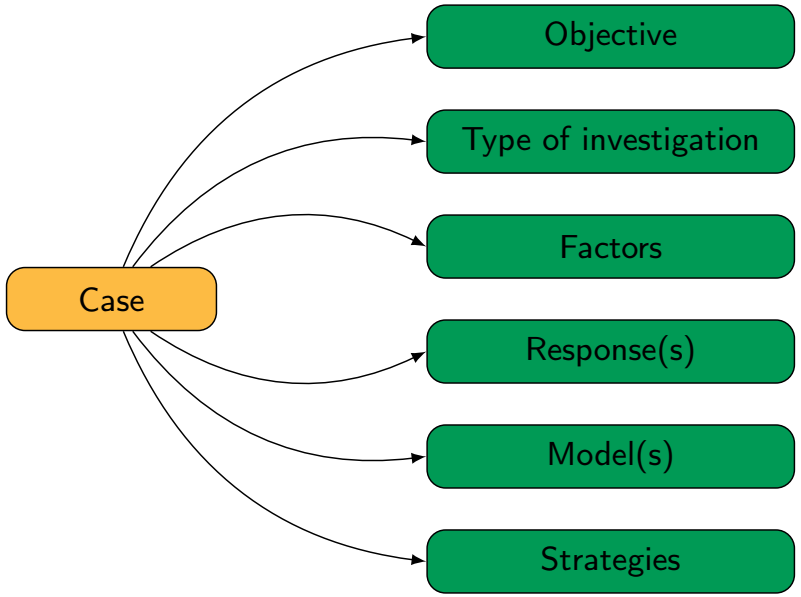


After about 10 degrees of freedom, the Student's t-distribution begins to closely resemble the normal distribution, and further increases in degrees of freedom result in only minor changes to the shape of the distribution. This means that, for practical purposes, you would need a much larger sample size (a change in the order of magnitude) to observe a significant reduction in the width of the confidence interval. Similarly, for confidence levels below 95%, the size of the confidence interval does not vary dramatically.

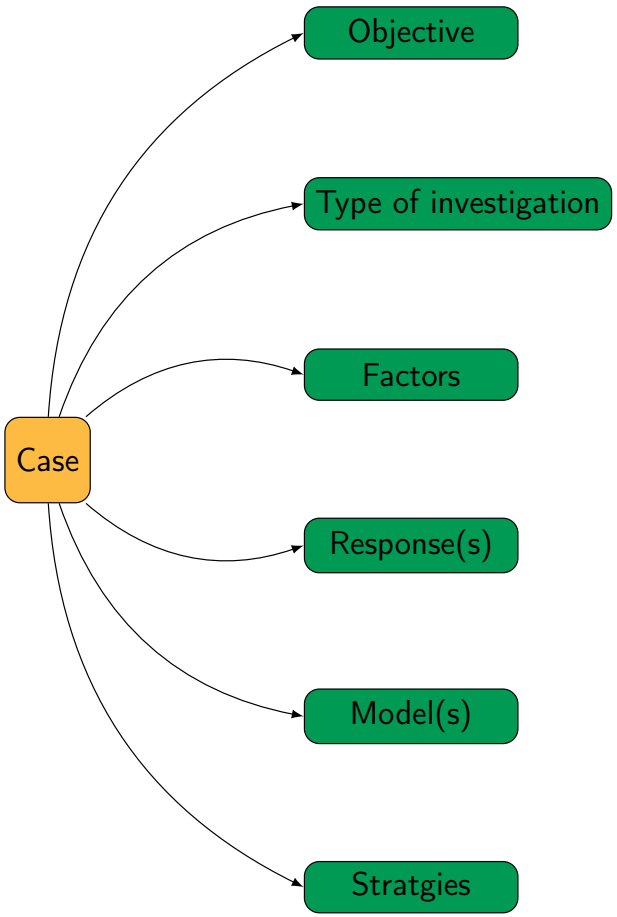
Therefore, using **10 degrees of freedom** and a **95% confidence** level is a reasonable standard for many statistical analyses, as it balances precision with practicality.

1.4.0 Mindmap of a case

- To do at the start of the project
- To maintain all along the project
- By hand or with a dedicated application
- Some available applications :
  - Freemind
  - iMindmap
  - Mindjet
  - MIRO



1.4.1 Building the mindmap



Usually, it concists in determinating the dominant factors, or the optimal conditions.

Indicate the type of investigation : observationnal vs experimental (controled), randomized or not, with the purpose to place the project in the scale of proof.

Make the list of the factors with their range and their essential characteristics (discreet vs continuous, hard-to-change, nested, etc.)

Make the list les responses to be modelled.

Make the list of the models of interest which can be theoretical or empirical.

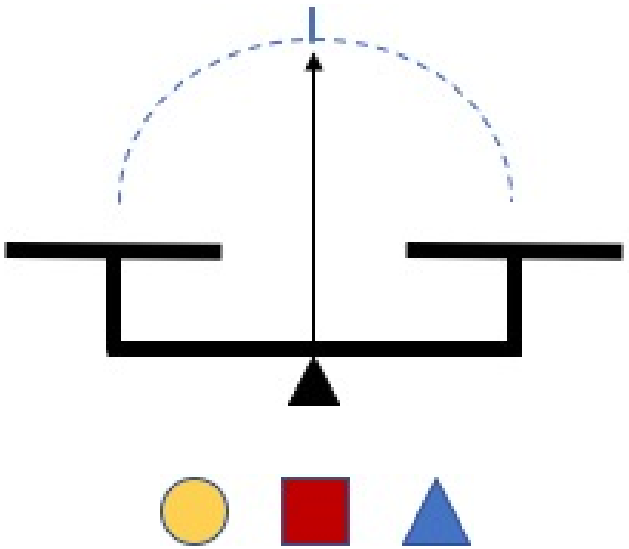
Make the list of the strategies possible to get to the objective. Indicate the number of experiments for each strategy, that will help to take decisions. By linking the elements of the strategy it is possible to define a scenario.

1.4.2 Weighing three objects with a two-pan scale

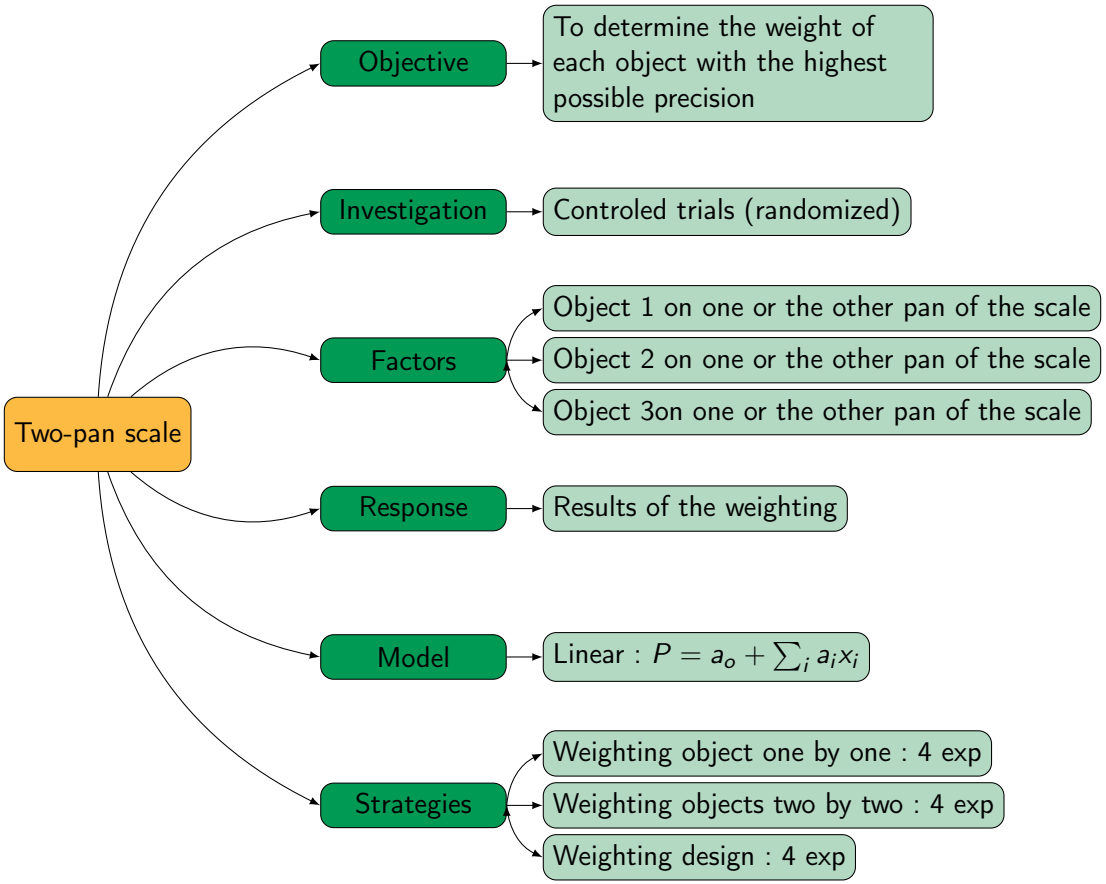
Description of the problem

To measure the weight of three objects with the best accuracy for a reasonable cost :

- The weight of the objects are of the same order of magnitude
- The instrument is a two-pan scale

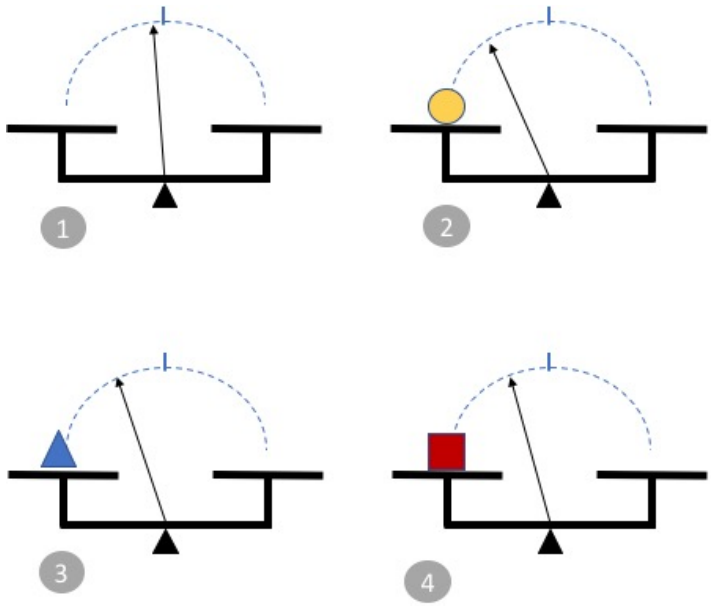


1.4.3 Weighing three objects



1.4.4 Strategy 1 : Weighting the objects one by one

- Four measurements
- One measurement without any object to determine the offset of the scale
- One object only at a time on one of the two pans

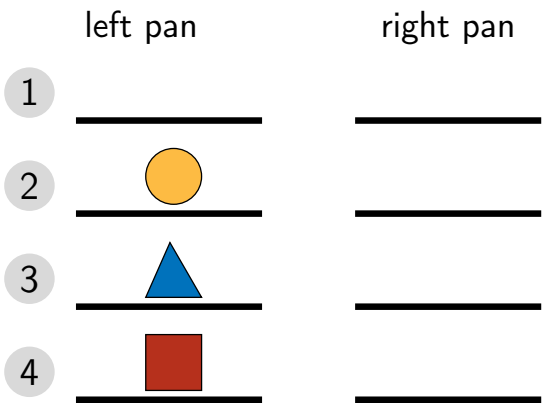


Questions :

- What is the weight of each object ?
- What is the accuracy of the results ?

1.4.4 Strategy 1 : Weighting the objects one by one

- What is the weight of each object ?
- What is the accuracy of the results ?

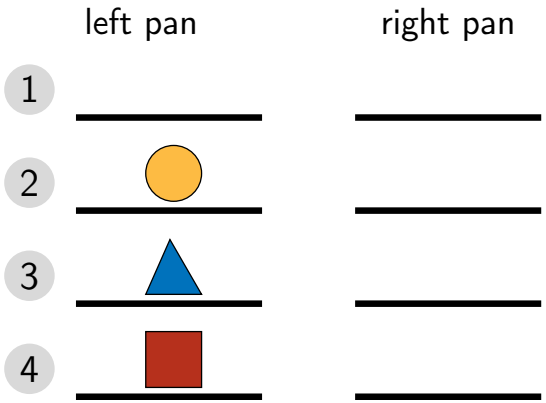


$$\begin{bmatrix} R_o \\ R_1 \\ R_2 \\ R_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} m_o \\ m_1 \\ m_2 \\ m_3 \end{bmatrix}$$

$$\begin{cases} m_o = R_o \\ m_i = R_o - R_i \end{cases}$$

1.4.4 Strategy 1 : Weighting the objects one by one

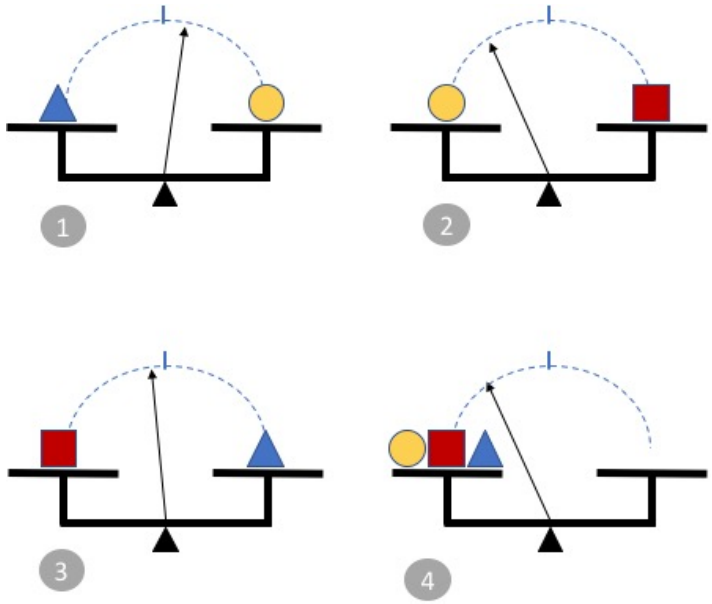
- What is the weight of each object ?
- What is the accuracy of the results ?



$$\begin{cases} var(m_o) & = var(R_o) = \sigma^2 \\ var(m_i) & = var(R_o - R_i) \\ & = var(R_o) + var(R_i) \\ & = 2\sigma^2 \end{cases}$$

1.4.5 Strategy 2 : Weighting objects 2 by 2

- Four measurements
- Three measurements with one object per pan
- One measurement with the three objects on one pan

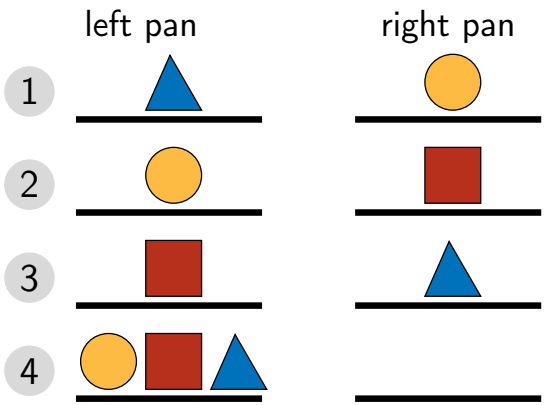


Questions :

- What is the weight of each object ?
- What is the accuracy of the results ?

1.4.5 Strategy 2 : Weighting objects 2 by 2

- What is the weight of each object ?



$$\begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} m_o \\ m_1 \\ m_2 \\ m_3 \end{bmatrix}$$

$$\vec{R} = X\vec{m} \quad \Rightarrow \quad \vec{m} = X^{-1}\vec{R}$$

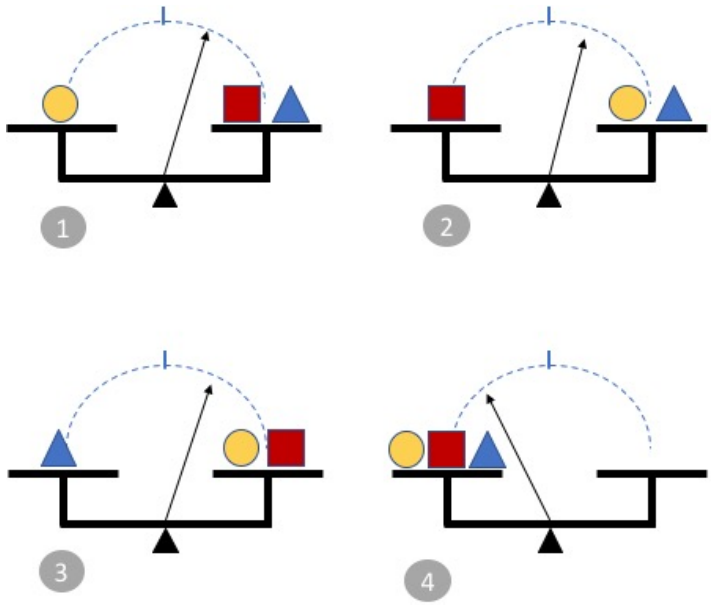
1.4.5 Strategy 2 : Weighting objects 2 by 2

- What is the accuracy of the results ?
- $$\begin{bmatrix} m_o \\ m_1 \\ m_2 \\ m_3 \end{bmatrix} = \frac{1}{9} \begin{bmatrix} 3 & 3 & 3 & 0 \\ 4 & -2 & 1 & -3 \\ -2 & 1 & 4 & -3 \\ 1 & 4 & -2 & -3 \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix}$$
- $var(\vec{m}) = var(X^{-1}\vec{R}) = (X^T X)^{-1} var(\vec{R}) = D\ var(\vec{R})$
- $var(m_i) \approx D_{ii}\ \sigma^2$
- $D_{00} = 1/3 \quad D_{11} = D_{22} = D_{33} = 10/27$

see the comprehensive variance calculation in the document " "

1.4.6 Strategy 3 : Weighting objects 3 by 3

- Four measurements
- For three measurements, two objects are weighted against a third one
- For one measurement the three objects are weighted together

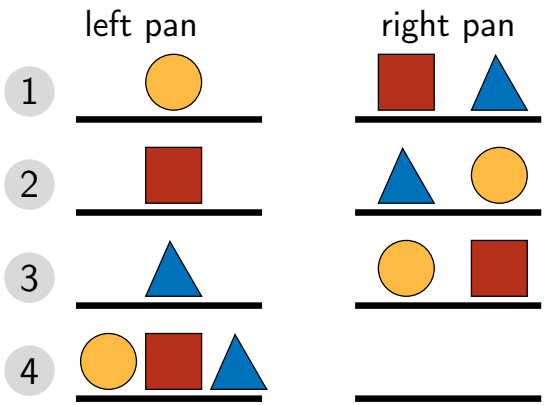


Questions :

- What is the weight of each object ?
- What is the accuracy of the results ?

1.4.6 Strategy 3 : Weighting objects 3 by 3

- What is the weight of each object ?



$$\begin{bmatrix} R_o \\ R_1 \\ R_2 \\ R_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} m_o \\ m_1 \\ m_2 \\ m_3 \end{bmatrix}$$

$$\vec{R} = X\vec{m} \quad \Rightarrow \quad \vec{m} = X^{-1}\vec{R}$$

1.4.6 Strategy 3 : Weighting objects 3 by 3

- What is the accuracy of the results ?
- $$\begin{bmatrix} m_o \\ m_1 \\ m_2 \\ m_3 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} R_o \\ R_1 \\ R_2 \\ R_3 \end{bmatrix}$$
- $var(\vec{m}) = var(X^{-1}\vec{R}) = (X^T X)^{-1} var(\vec{R}) = D\ var(\vec{R})$
- $var(m_i) \approx D_{ii}\ \sigma^2$
- $D_{00} = D_{11} = D_{22} = D_{33} = 1/4$

1.4.7 Conclusion

- DOE invented in the 20's by Fisher
- Importance of the visual check of data
- Beware of cognitive biases
- Follow the relation between the mathematical and the causal model
- Make a mind-map at the beginning and maintain it along the project
- Sorting data is an easy way of performing visual comparison
- A lot of functions available on Matlab (Python) for data analysis
- The weight of 3 objects : a paradigm of DOE