BIG DIVE

TECH. CUSTOM EDITION

A project by TOP-IX
designed for Intesa Sanpaolo

#DataScience

aizoOn® AUSTRALIA EUROPE USA
TECHNOLOGY CONSULTING

aizoOn Group Profile

aizoOn is an independent technology consulting firm.
Focused on innovation.
aizoOn has achieved record growth with offices across Australia, Europe and USA.


Gartner.

G00300966

Cool Vendors in Managing Operational Technology in a Digital Business, 2016

Published: 25 April 2016

aizoOn Organisation: Markets & Technology

TeMI AUTOMATION, DURABLE GOODS, ENTERTAINMENT, MEDIA, TELCO, TRAVEL	**E2&R** CHEMICAL, ENERGY, ENVIRONMENT, MINING, OIL&GAS, UTILITIES	**DASS** AERONAUTICS, AEROSPACE, DEFENCE, NAVAL	**Pa_S** CENTRAL PA, LOCAL PA, HEALTH	**DDI** DATA DRIVEN INNOVATION	**CYSE** CYBER SECURITY
FIN INSURANCE, INVESTIMENT BANKING, REAL ESTATE, RETAIL BANKING	**F&F** FASHION, FOOD	**LAIF** AUTOMOTIVE, INFRASTRUCTURE, LOGISTIC, METRO, RAIL	**Scientiae** BIOTECHNOLOGY, HEALTH CARE, LIFE SCIENCES, PHARMACEUTICAL	**SW&IT** SOFTWARE DEVELOPMENT & ARCHITECTURES	**ITS** INTELLIGENT THINGS & SYSTEMS

About me

linkedin.com/in/paolobajardi

- PhD Physics of Complex Systems → useful for critical thinking
- Master in Epidemiology → useful for Stat
- Master in Data Engineer → useful for Tech (and CV)

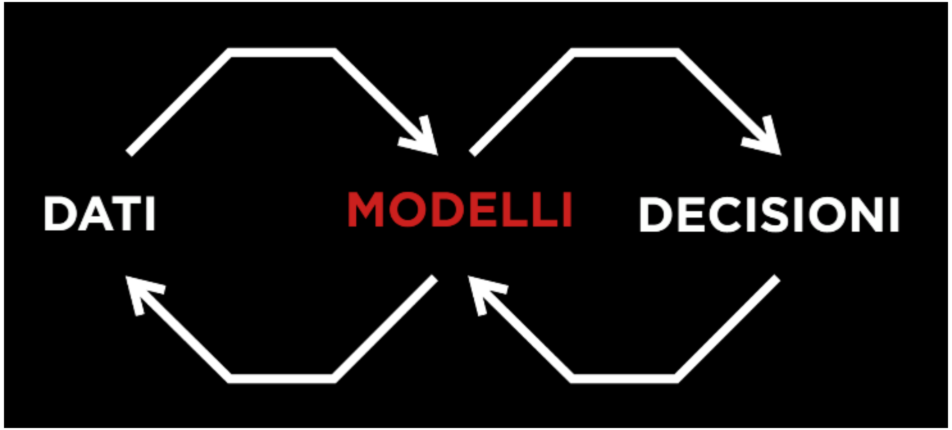
Moved from Reasearch to Business

Statistics & Data Analysis

AUSTRALIA
EUROPE
USA

TECHNOLOGY CONSULTING

Models



- Mathematical model
- Dynamical model
- Statistical model
- Machine learning model
- Agent-based model
- Data model
- ….

Data ..*..

“**Data analysis** has been generally used as a way of *explaining some phenomenon* by *extracting interesting patterns* from individual data sets with well-formulated queries.

Data science, on the other hand, aims to discover and extract *actionable knowledge* from the data, that is, knowledge that can be used to make *decisions and predictions*, not just to explain what’s going on.”

– I. Wladawsky-Berger

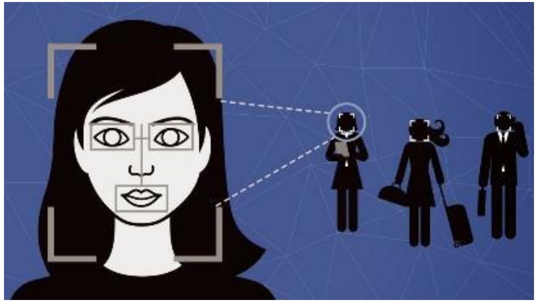
Actionable Knowledge

1. Scoprire patterns nascosti nei dati

The New York Times

What Wal-Mart Knows About Customers' Habits

2. Automatizzare decisioni su larga scala



Data-analytic thinking

“ It is important to understand data science even if you never intend to apply it to yourself. **Data-analytic thinking enables you to evaluate proposal for data mining projects**.

[…] If a potential business application extracting knowledge from data is proposed, you should be able to assess the proposal systematically and decide whether it is sound or flawed. “

- Foster Provost

Data Science ‘complementary skills’

- Capacità di scomporre un problema di data-analytics in parti per cui esistono strumenti e tecniche note da applicare

  → non reinventare la ruota!

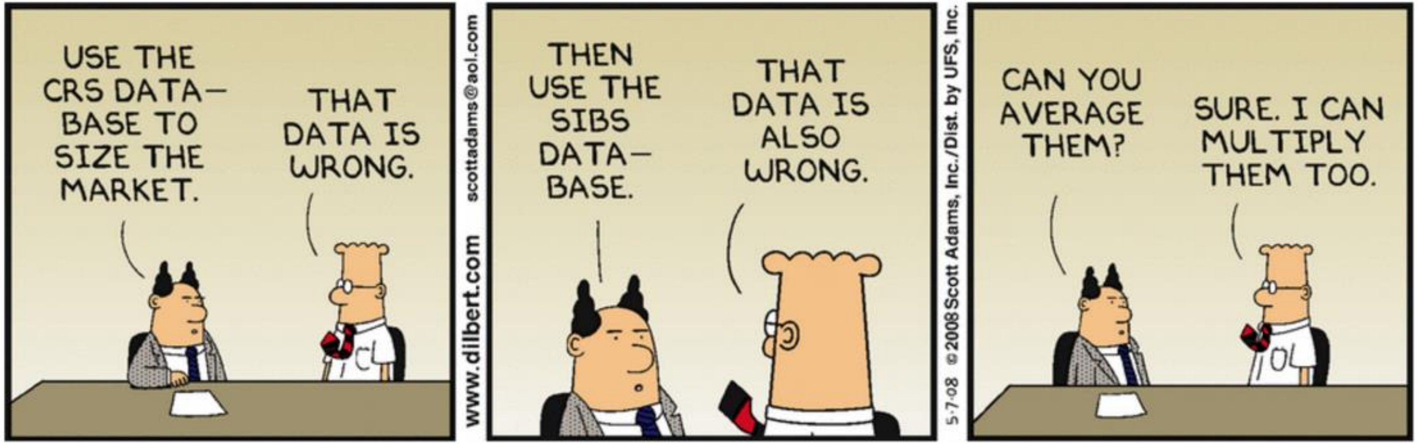
- Intuizione e immaginazione per sviluppare test e analisi in grado di validare o smentire le ipotesi generate

  → Non sempre è possibile fare test in laboratorio!

- Capacità di comunicazione dei risultati e ascolto dei requisiti di business

  → Actionable Knowledge!

Data Quality

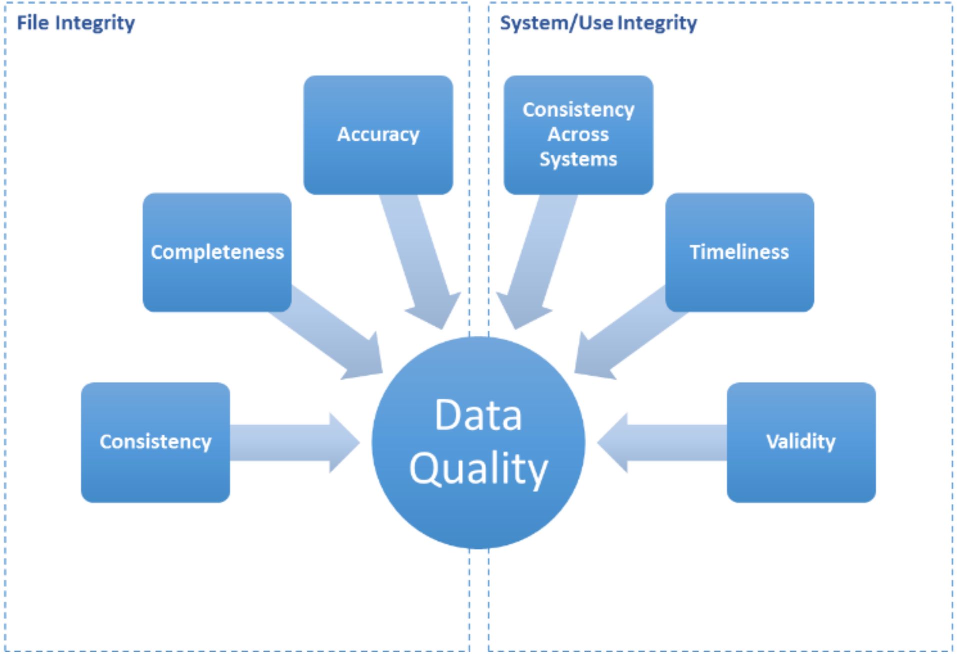


Garbage in, garbage out (**GIGO**) refers to the fact that computers, since they operate by logical processes, will unquestioningly process unintended, even nonsensical, input data ("garbage in") and produce undesired, often nonsensical, output ("garbage out").

- Wikipedia

Data Quality

Esistono molteplici definizioni e dimensioni per valutare la qualità dei dati.
Ad esempio:



Data Quality

Data Quality

Fitness-for-use
(Wang & Strong,1996)

Suggerimento: verificare i dati per integrità, orizzonte temporale e spaziale, modalità di raccolta e pattern di dati mancanti

Data Science: Il ruolo della statistica

Data Exploration

Model Testing

Prediction

Statistica descrittiva

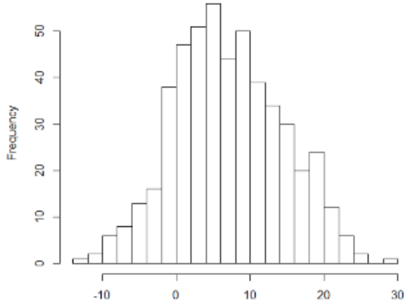
Dati

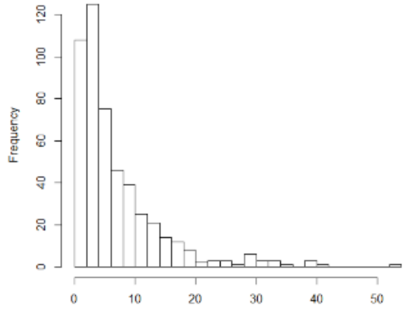
2	########	0:00:01	231	29	165	0.4	1.4	3.9	14.4	75.5	1021.8
3	########	0:10:01	234	28	158	0.4	1.3	3.4	14.4	75.5	1021.7
4	########	0:20:01	211	19	176	0.4	1.3	3.4	14.1	75.7	1021.7
5	########	0:30:01	211	25	187	0.4	1.2	2.7	14.3	76	1021.7
6	########	0:40:01	225	28	188	0.3	1.2	2.9	14.3	75.6	1021.7
7	########	0:50:01	226	30	177	0.4	1.3	3.3	14.3	75.3	1021.8
8	########	1:00:01	224	20	196	0.4	1.4	3.5	14.4	74.6	1021.9
9	########	1:10:01	231	29	174	0.3	1.3	3.4	14.4	74.5	1021.9
10	########	1:20:01	217	15	196	0.3	1.2	3.1	14.4	74.8	1021.9
11	########	1:30:01	232	17	187	0.4	1.3	3.3	14.5	74.3	1021.9
12	########	1:40:01	230	23	179	0.4	1.2	3.1	14.5	74.5	1022
13	########	1:50:01	237	21	171	0.3	1.3	3.2	14.6	74.4	1021.9
14	########	2:00:01	224	25	171	0.5	2.2	5.5	14.4	77.9	1021.9
15	########	2:10:01	230	32	182	0.4	1.9	5.2	13.8	82.8	1021.8
16	########	2:20:01	235	25	197	0.5	1.6	4	13.8	82.5	1021.9
17	########	2:30:01	230	21	174	0.5	1.6	3.6	13.9	82.1	1021.8
18	########	2:40:01	215	20	173	0.4	1.5	4.5	14	81	1021.6
19	########	2:50:01	227	26	197	0.5	1.6	3.8	14.1	80.2	1021.8
20	########	3:00:01	266	28	165	0.4	1.7	3.4	14.3	79.1	1021.7

Indicatori statistici

Full-Time Statistics by Geographical Region	Average	Standard Deviation
Texas	$60,632	$13,375
Austin, Texas*	$52,500	$11,529
Dallas, Texas*	$60,635	$10,041
Houston, Texas*	$66,996	$13,646
Other Texas*	$49,360	$16,406
Mid-Atlantic	$54,200	$22,451
Midwest	$60,770	$8,920
Northeast	$68,069	$15,595
South	$55,220	$11,670
Southwest	$64,600	$12,178
West	$62,341	$16,173
Other	$63,588	$5,074

Visualizzare i dati





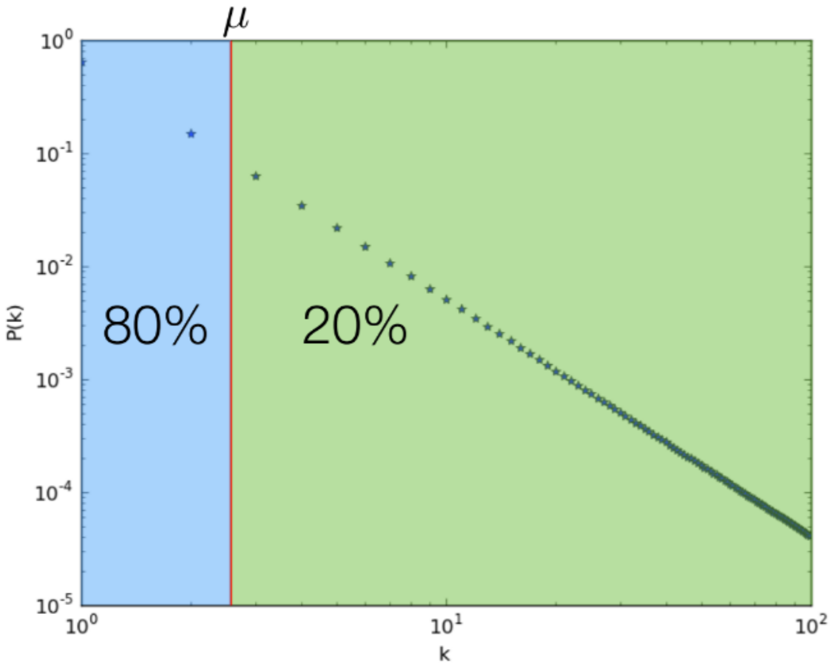
Outliers & heavy tail distribution

- Distribuzioni con una lunga coda: **quasi tutti sono sotto la media.**

- Attenzione agli **outliers**: se Bill Gates entra in un bar, in **media** tutti gli avventori sono milionari…

1 1 1 **2** 2 2 1000

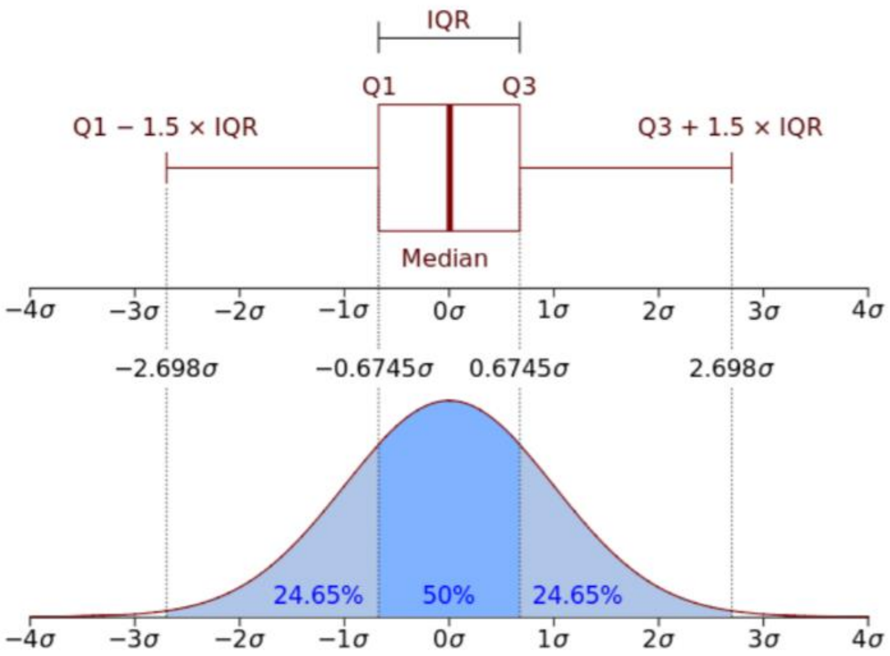
- ….ma la **mediana** resta invariata!



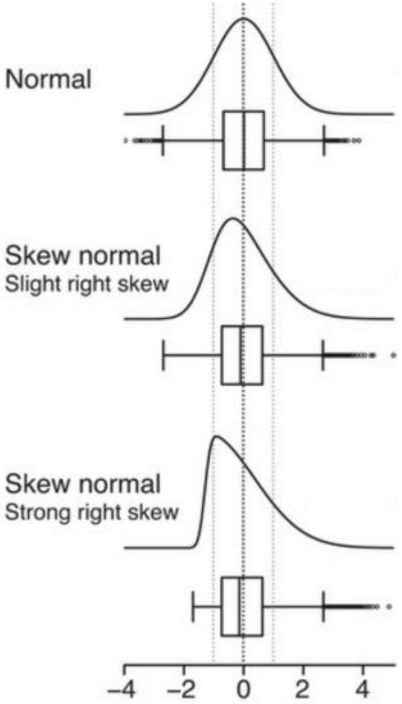
Quantili

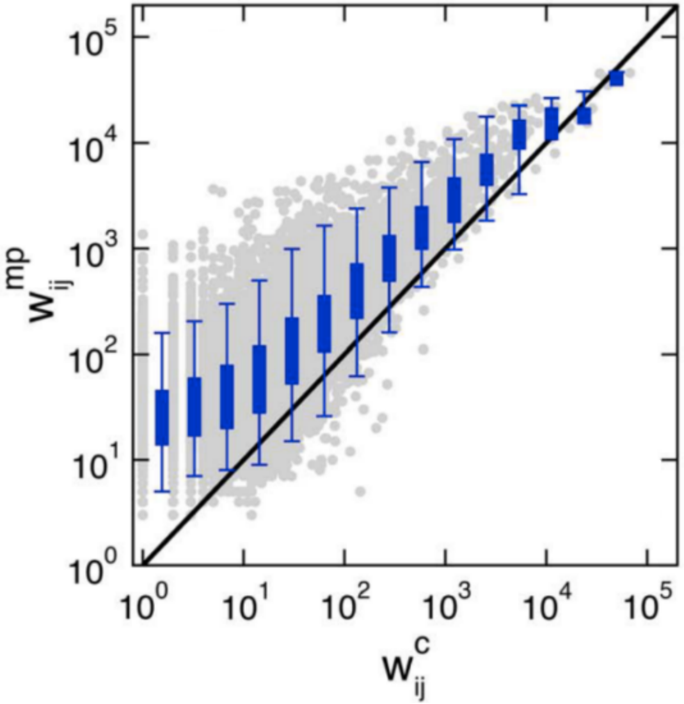
Quantili: punti presi a intervalli regolari sulla distribuzione cumulativa

- Quartili
- Percentili
- …



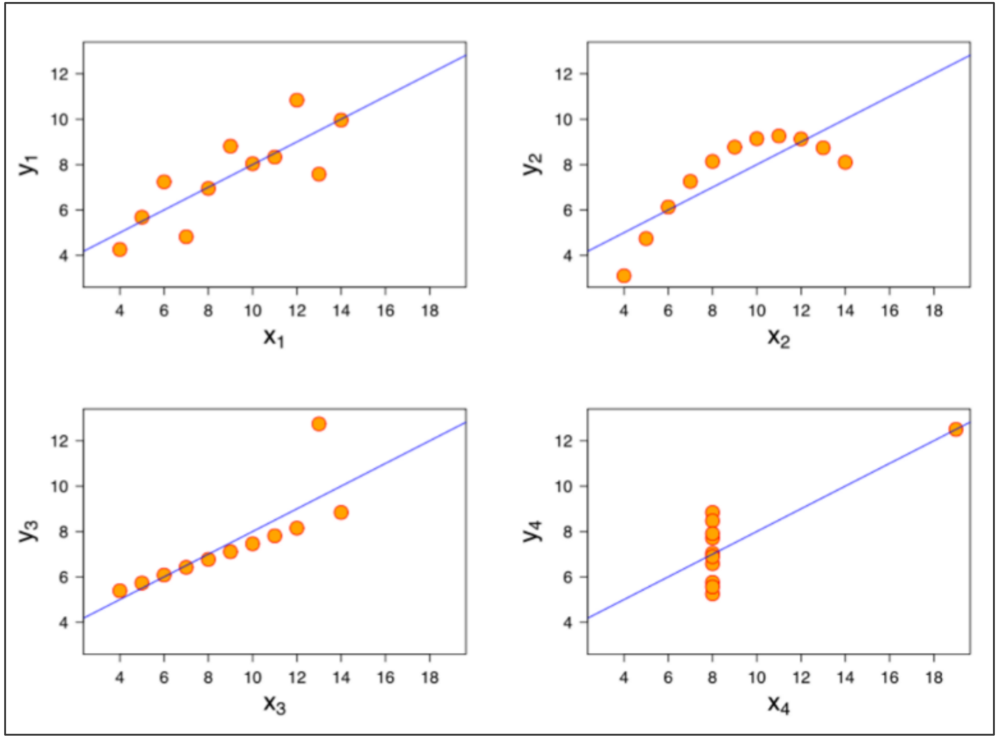
Boxplot





- Show the variation for each data bin
- More informative than averages or medians
- Useful to summarize measurements or simulated results when fluctuations are important

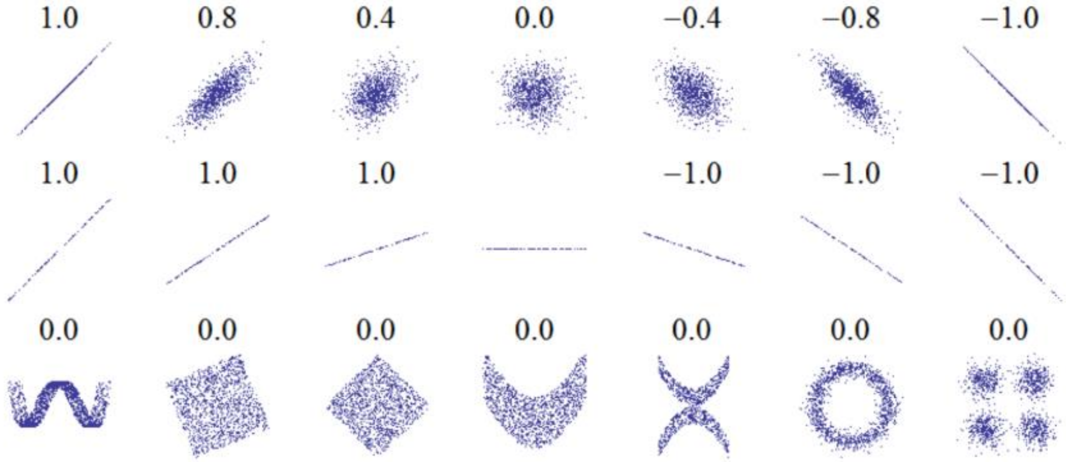
Statistica descrittiva



NOTA: tutte **le variabili** x_i hanno la **stessa media e stessa varianza**.
Lo stesso vale per tutte le variabili y_i.

Coefficiente di Pearson

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y}$$



credits: Wikimedia Commons

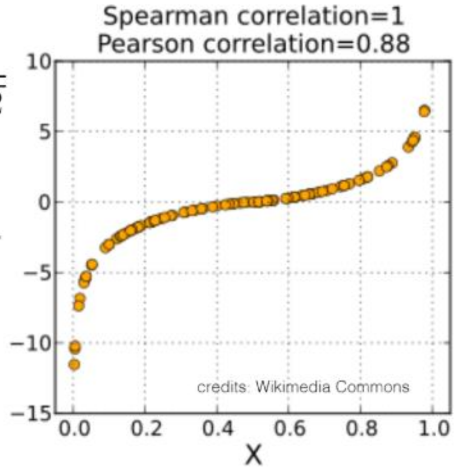
<u>Assunzioni</u>:

- Correlazione lineare
- Variabili continue
- Variabili con distribuzione normale (no outliers)
- Le due variabili formano una distribuzione normale bivariata
- Omoschedasticità dei dati

Coefficiente di Spearman

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}$$

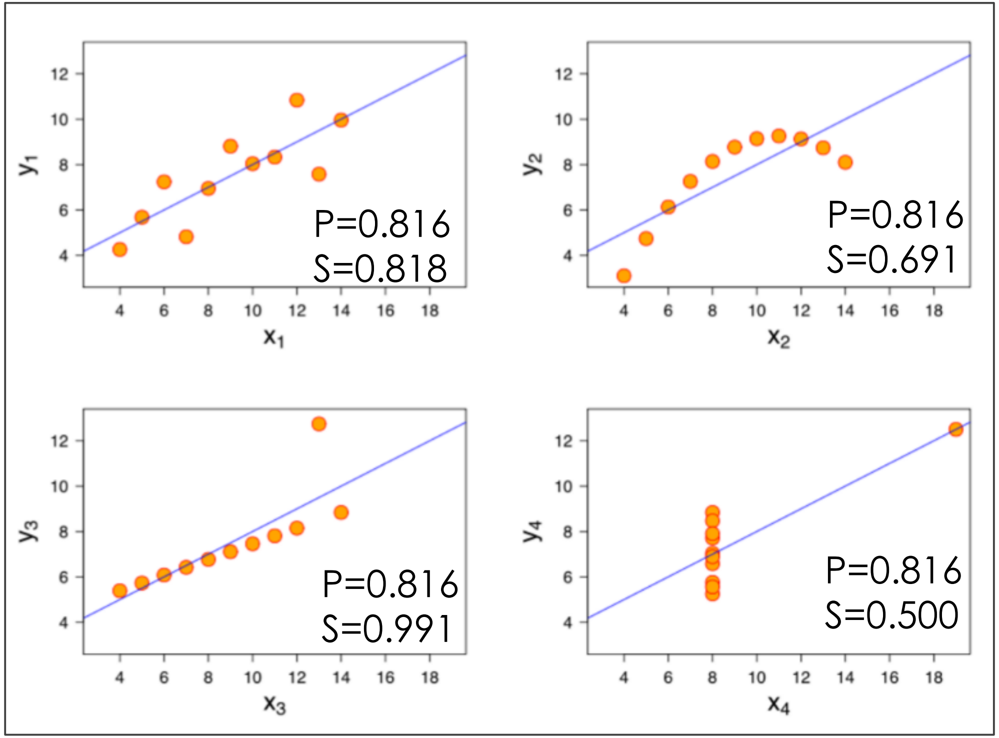


<u>Assunzioni</u>:

- Correlazione monotona
- Variabili continue o ordinali

→ Test non parametrico

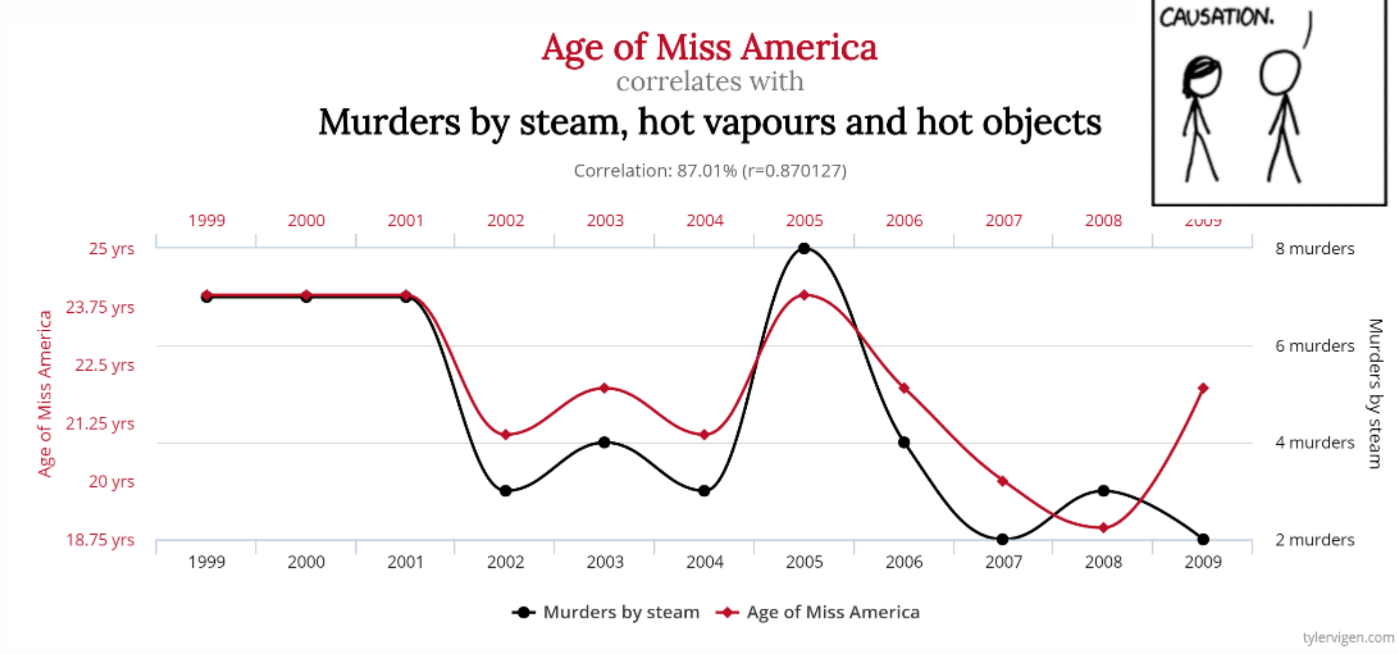
Correlazione

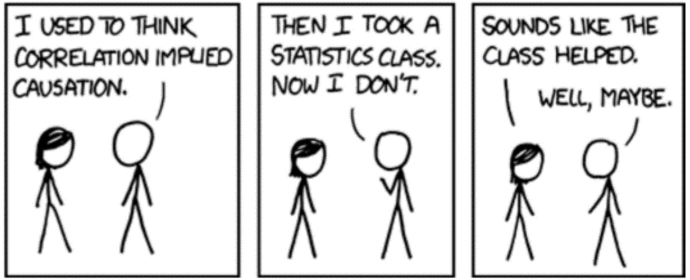


Tutti le variabili mostrate in figura hanno lo **stesso valore di correlazione lineare**.

(P: Pearson, S: Spearman)

Correlazione vs. Causalità





Correlazione vs. Causalità

Immaginiamo di osservare che due grandezze A e B siano correlate

$A \longrightarrow B$

‘Correlation is not causation but it sure is a hint.’
– E. Tufte

$B \longrightarrow A$

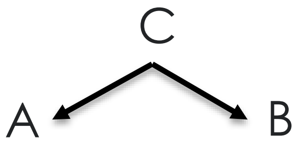
CAUSALITA’ INVERSA: Es. più pompieri vengono inviati e maggiori sono i danni registrati a causa dell’incendio.

$A \longleftarrow C \longrightarrow B$

A E B SONO CAUSATE DA UN FATTORE COMUNE C: Es. all’aumento del consumo di gelato, aumentano le morti per annegamento (“estate” è il fattore C).

.....

Confondimento e variabili proxy



- Nello stimare la forza di associazione tra A e B, la variabile C '**confonde**' la stima. → **BAD**

  Soluzioni possibili:

  1) randomizzazione
  2) misurare C

- La variabile A (misurabile) è un **proxy** della variabile C che è associata a B. → **GOOD**

Causality

Hill's criteria for causation, are a group of guidelines that can be useful for providing evidence of a causal relationship between a putative cause and an effect:

- **Strength**: A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.
- **Consistency**: Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect.
- **Temporality**: The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).
- **Plausibility**: A plausible mechanism between cause and effect is helpful
- **Analogy**: The effect of similar factors may be considered.

Causality

Hill's criteria for causation, are a group of guidelines that can be useful for providing evidence of a causal relationship between a putative cause and an effect:

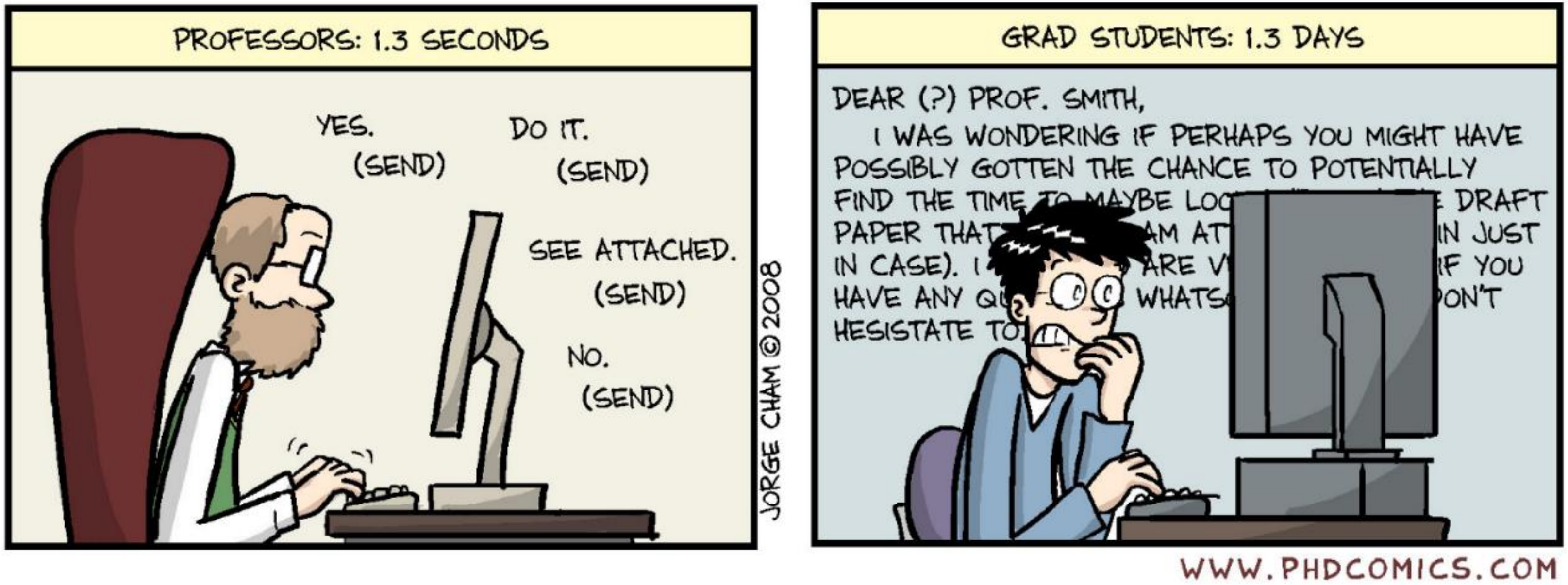
- **Strength + Consistency + Temporality + Plausibility + Analogy …**

- **Specificity**: Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.
- **Coherence**: Coherence between inferred and laboratory findings increases the likelihood of an effect.
- **Experiment**: "Occasionally it is possible to appeal to experimental evidence".

Feature Engeneering – Misurare l'effetto

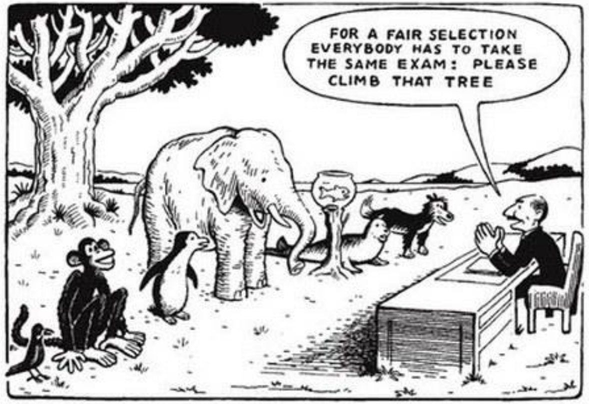
• Valore medio	→ effetti lenti e reversibili
• Valore cumulativo	→ effetti cumulativi e irreversibili
• Durata dell'esposizione	→ effetti cumulativi e irreversibili
• Valore di picco	→ effetto reversibile e acuto

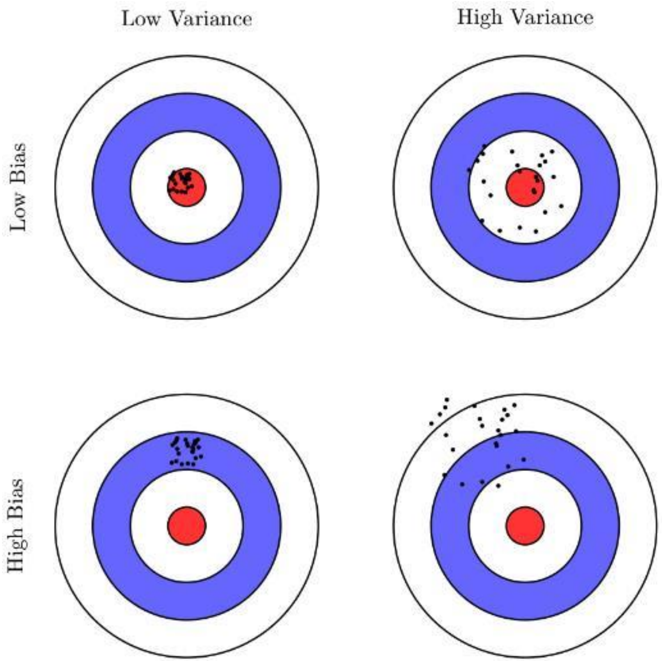
E.g. Average Time Spent Composing 1 email



Bias

Un pregiudizio (conscio od inconscio) o un errore di disegno delle analisi possono portare a misure distorte.





Diverse tipologie di bias

Es. Sbagliare la misura d'effetto può introdurre misclassificazione e solitamente causa un 'bias verso il valore nullo'.

A Catalog of Biases in Questionnaires

Bernard C.K. Choi, PhD, Anita W.P. Pak, PhD

Es. *Belief vs behavior:*

1. Do you think that it is a good idea to have everyone's chest regularly checked by X-ray?

2. Have you ever had yours checked?

Es. *Framing*

Which operation would you prefer?

[ ] An operation that has a 5% mortality.

[ ] An operation in which 90% of the patients will survive.

Simpson’s Paradox

‘il ***paradosso di Simpson*** *indica una situazione in cui una relazione tra due fenomeni appare modificata, o perfino invertita, dai dati in possesso a causa di altri fenomeni non presi in considerazione nell'analisi (variabili nascoste).’*

- wikipedia

vudlab.com/simpsons/

Simpson's Paradox

	# applied		# admitted		% admitted	
departments	men	women	men	women	men	women
"Easy"	1,372	294	856	234	38% 62%	20% 80%
"Hard"	1,319	1,541	338	408	26% 74%	27% 74%
Combined	2,691	1,835	1,194	642	44% 56%	35% 65%

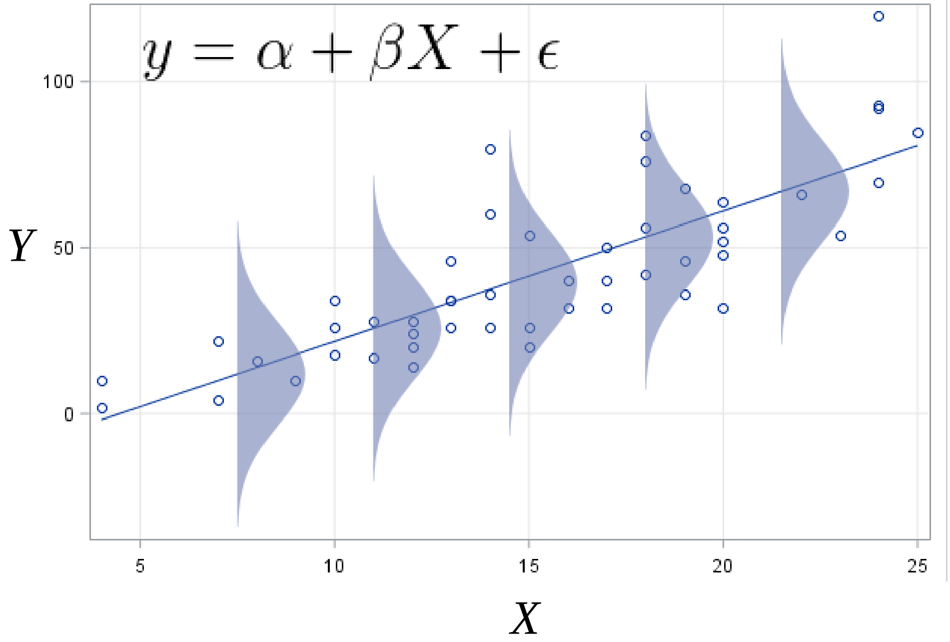
Simpson's paradox? yes

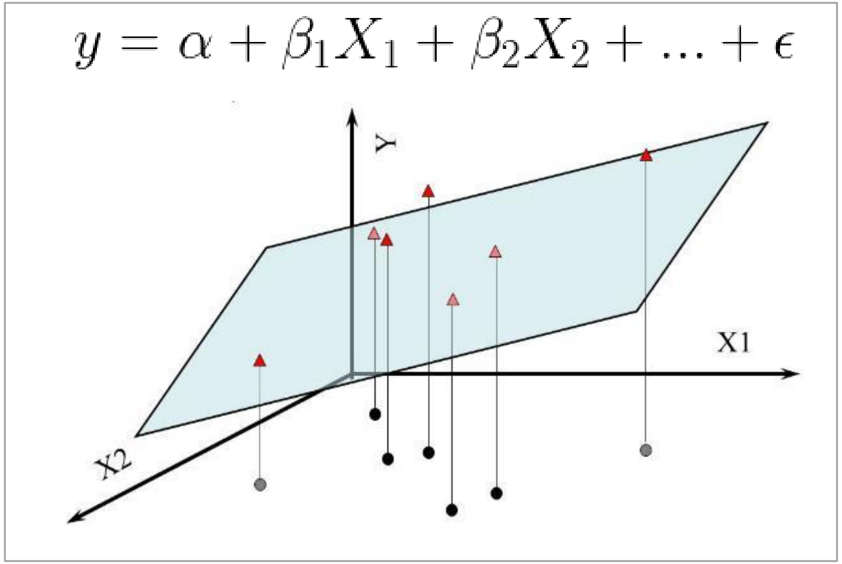
Simpson's Paradox

	# applied		# admitted		% admitted	
departments	**men**	**women**	**men**	**women**	**men**	**women**
"Easy"	1,372	294	856	234	62% 38%	80% 20%
"Hard"	1,319	1,541	338	408	26% 74%	27% 74%
Combined	2,691	1,835	1,194	642	44% 56%	35% 65%

Simpson's paradox? yes

Inference of association: regression – basic concepts





Inference of association: regression – basic concepts

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon$$

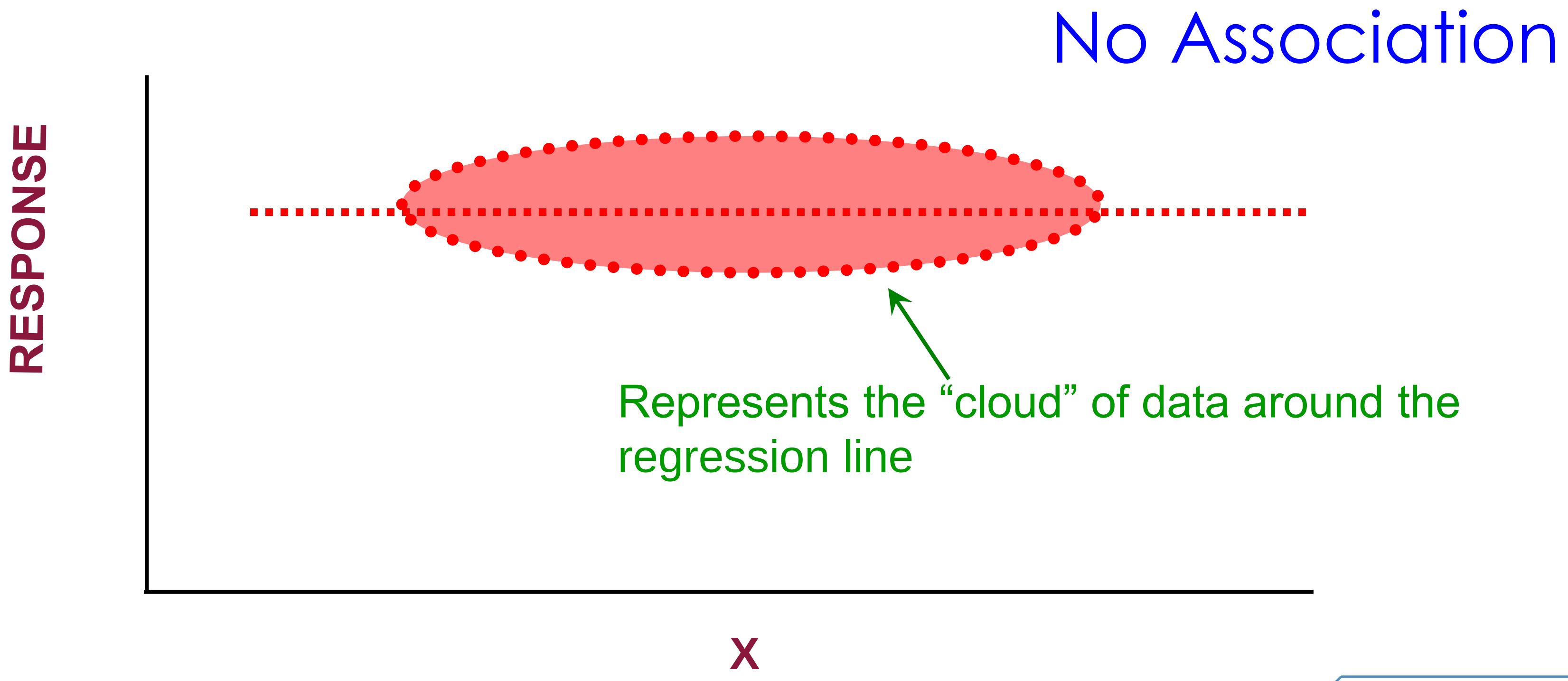
- Assunzioni da verificare:
  - Legame lineare
  - Variabili con distribuzione normale (no outliers)
  - Le due variabili formano una distribuzione normale bivariata
  - Omoschedasticità dei dati
  - Assenza di autocorrelazione
  - Tipologia variabili
  - multicollinearità

Inference of association: regression – basic concepts

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \epsilon$$

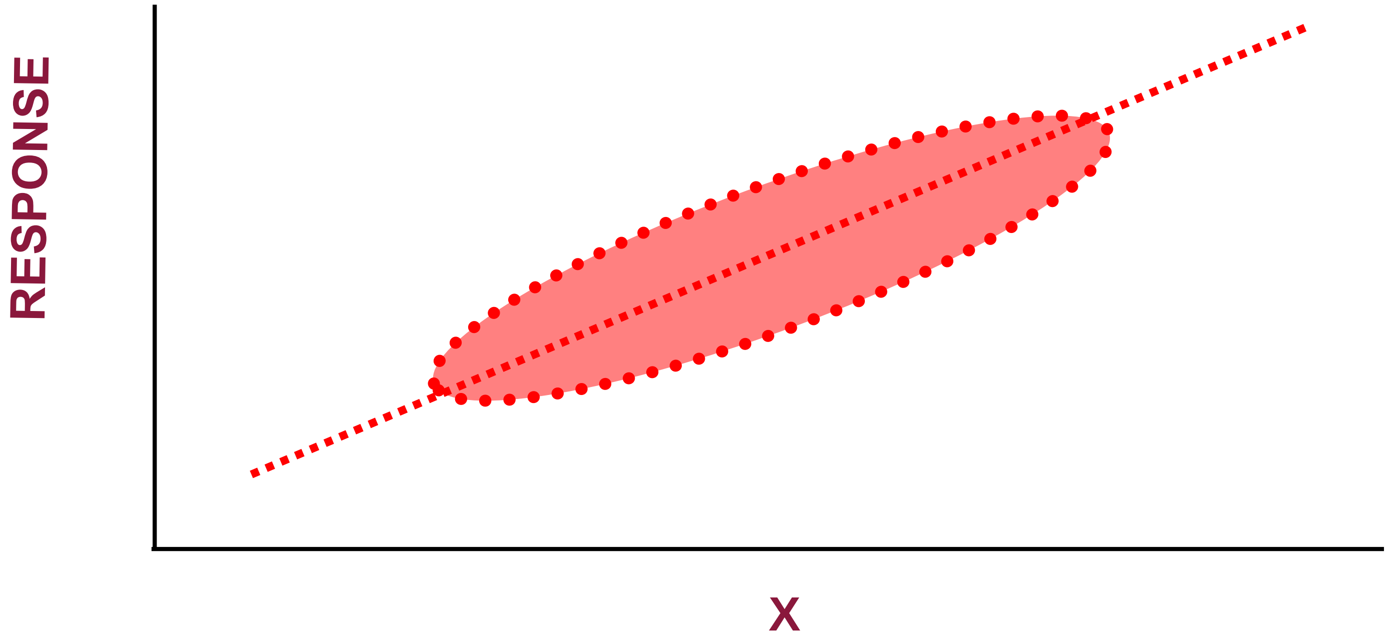
- Valutare se il modello ha una 'buona performance' (adjusted R-squared, RMSE, F-test, …..)
- β_i rappresenta la **forza di associazione** tra la variabile target y e la variabile indipendete X_i considerando i valori delle altre variabili.
  - Verificare 1) la forza di associazione della variabile e 2) se l'associazione è statisticamente significativa
- Valutare se esistono modelli migliori → Akaike Information Criterion (AIC)

Inference of association: regression – basic concepts



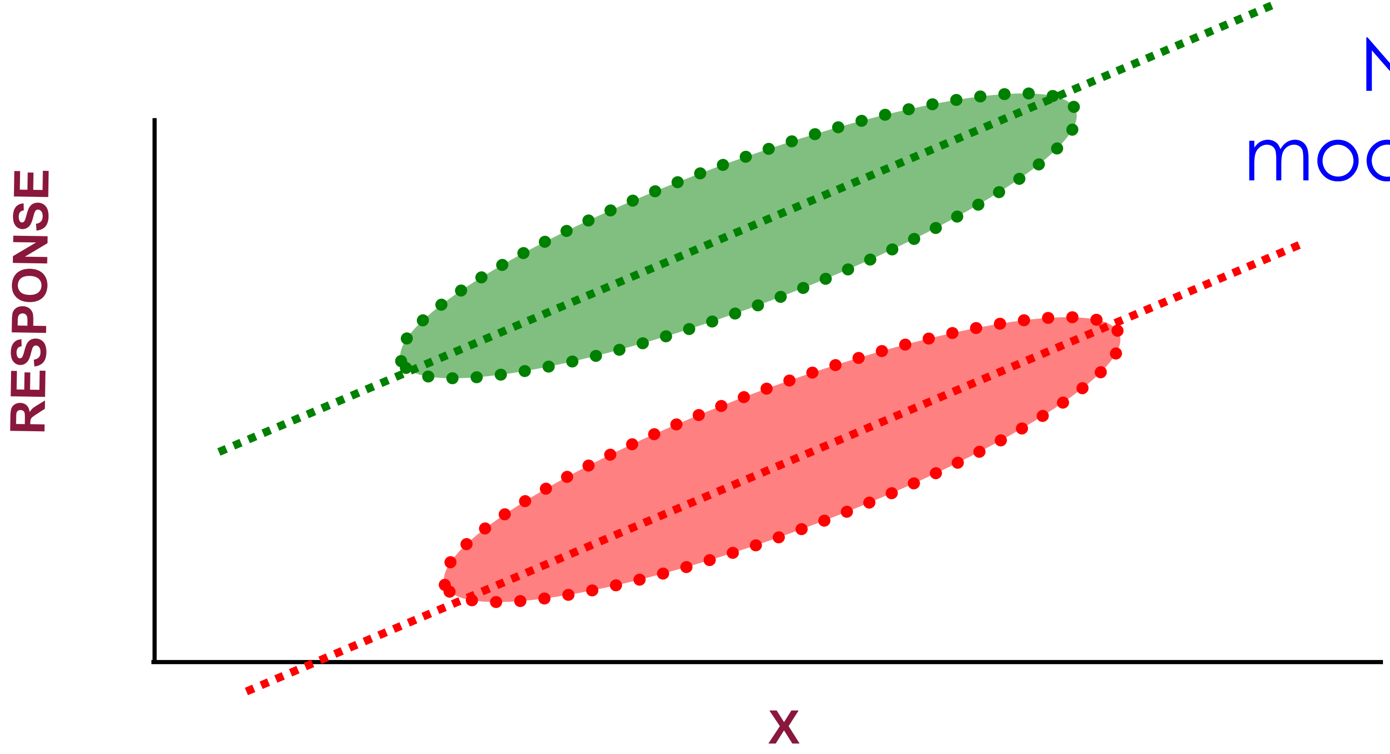
Inference of association: regression – basic concepts

Positive Association

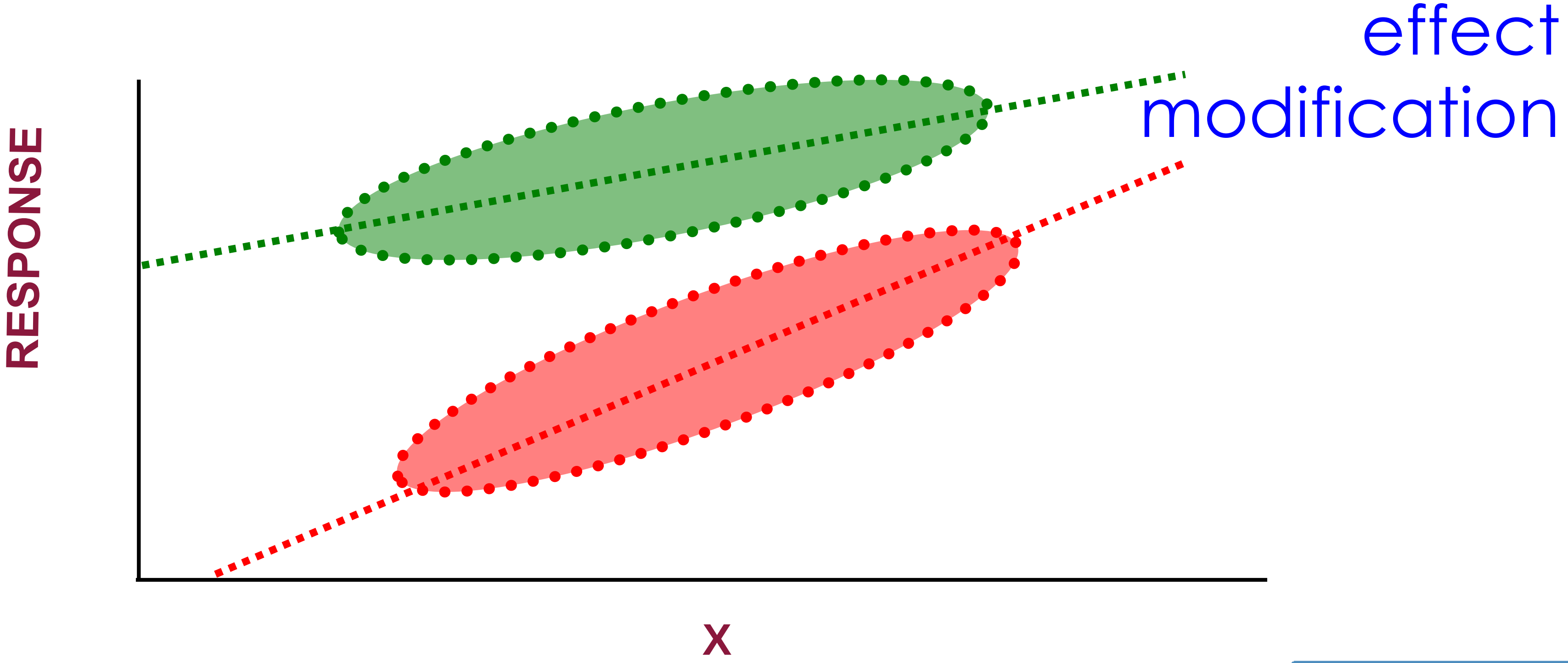


Inference of association: regression – basic concepts

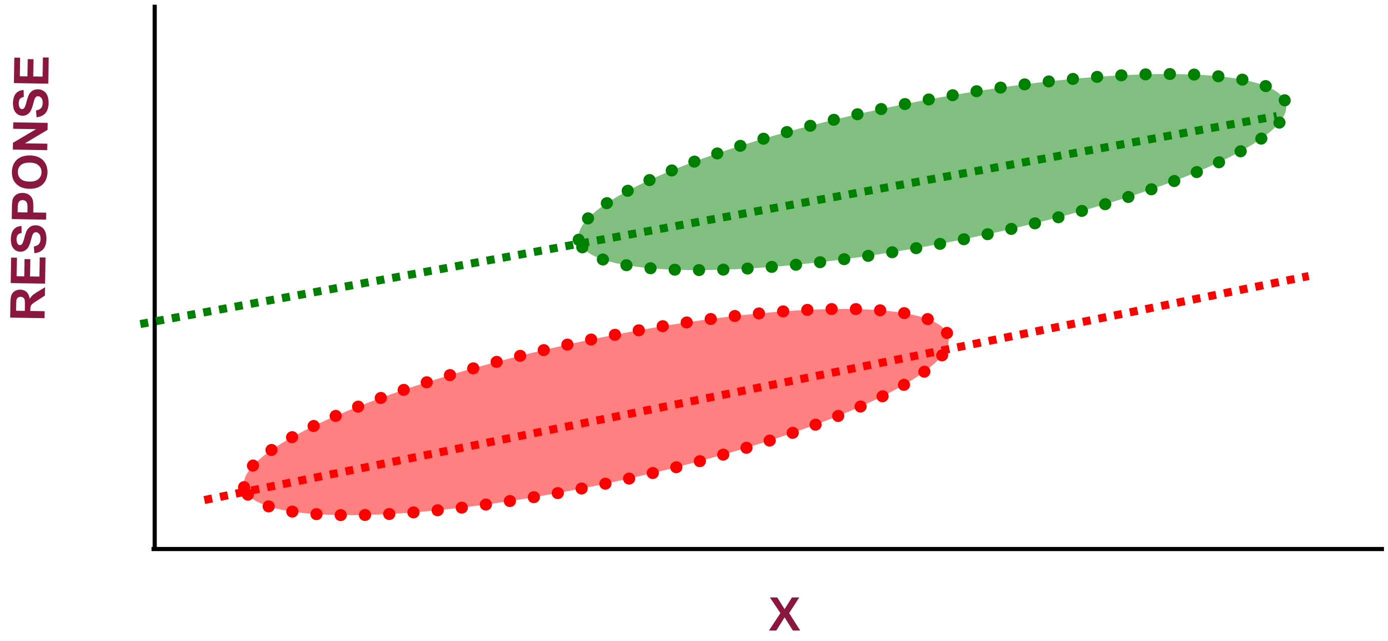
No effect modification



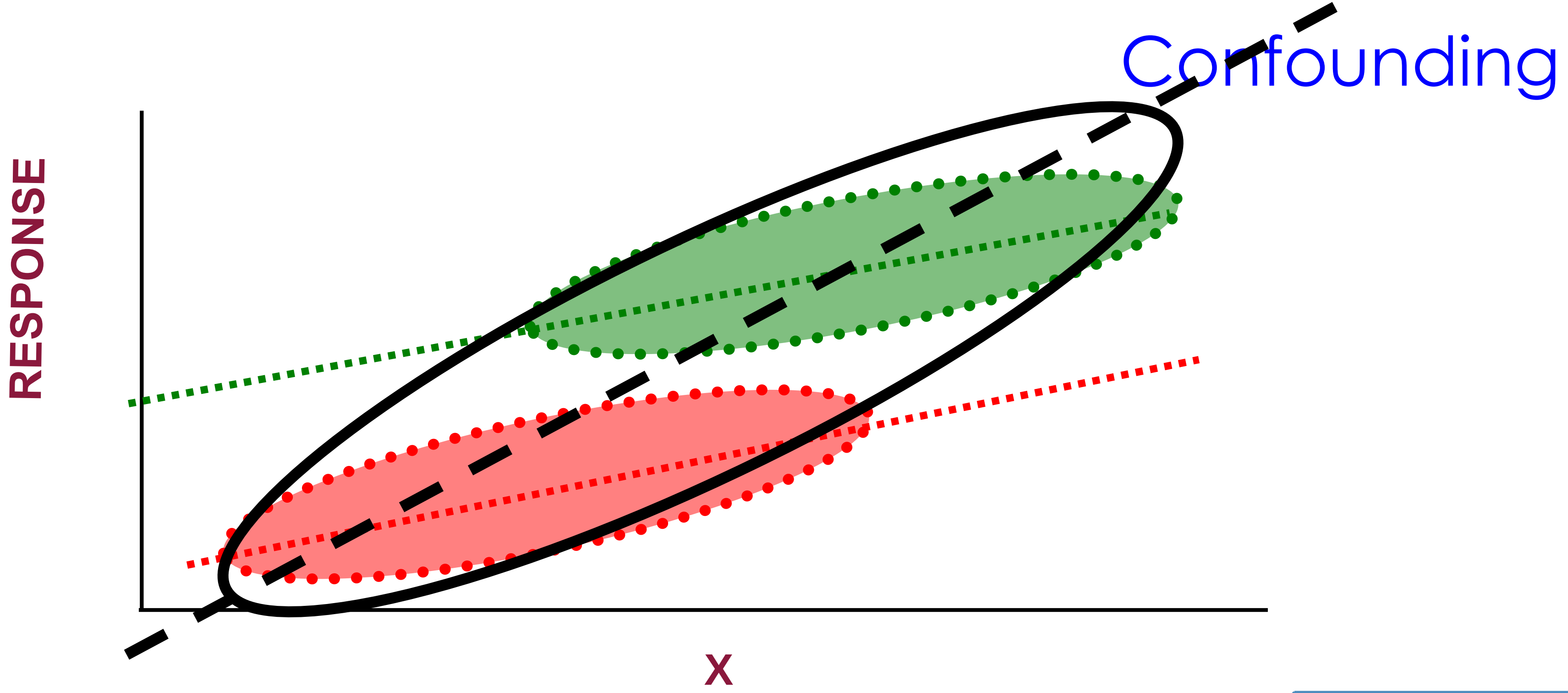
Inference of association: regression – basic concepts



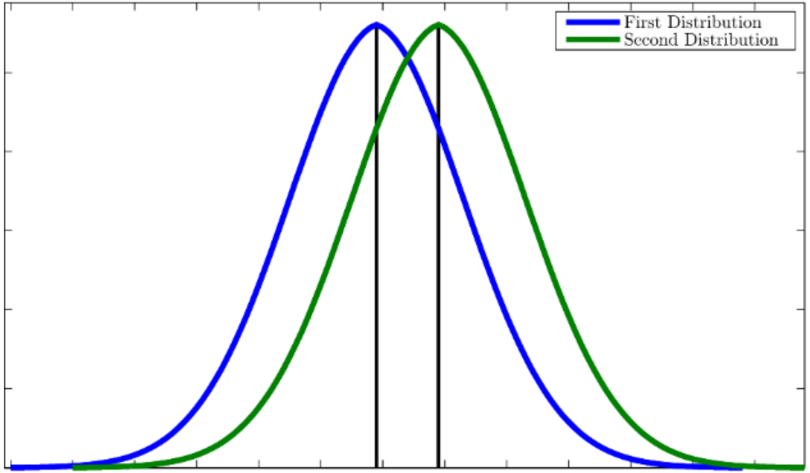
Confounding

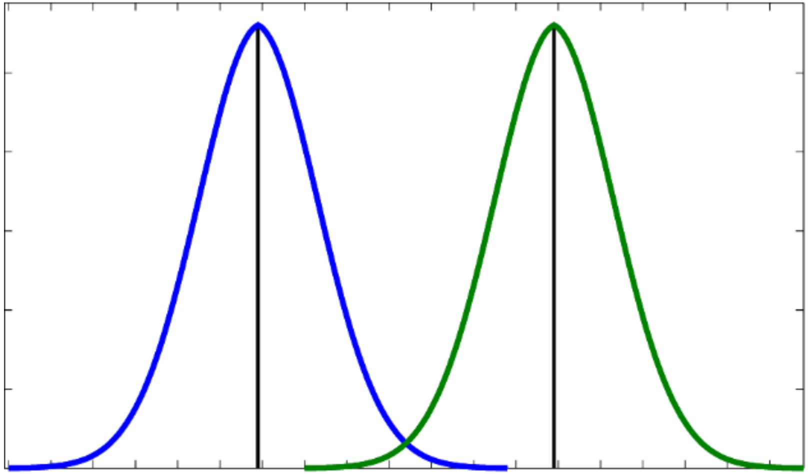


Inference of association: regression – basic concepts



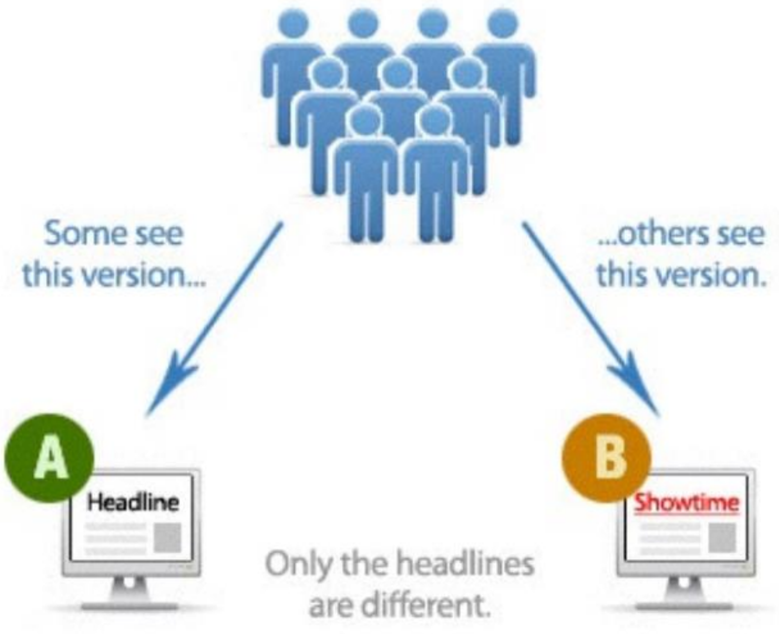
Confronto fra gruppi





Testare se le differenze osservate sono ‘statisticamente significative’

Confronto fra gruppi – A/B Testing



p_A p_B

- Gli utenti vengono divisi in due gruppi (A e B)

- Per ogni gruppo viene misurata una metrica di interesse (ad es. probabilità di conversione)

- Viene testata **l'ipotesi nulla** che le metriche misurate nei due grupppi (controllo e trattati) non siano diverse:

$$\boxed{H_0 = p_A - p_B \leq 0}$$

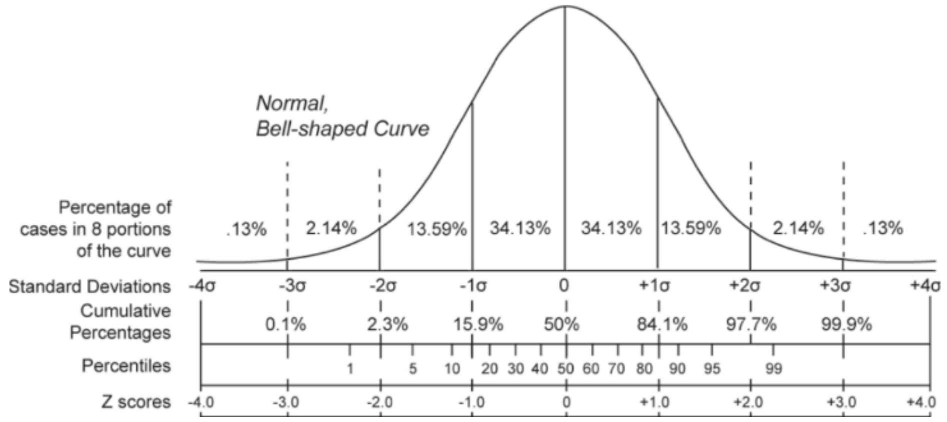
Confronto fra gruppi – A/B Testing

- Se l'outcome di interesse è un **processo binomiale**
- L'**errore standard** è:

$$SE = \sqrt{\frac{p\,(1-p)}{N}}$$

- Con **Z-score** (i.e. numero di deviazioni standard dalla media):

$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

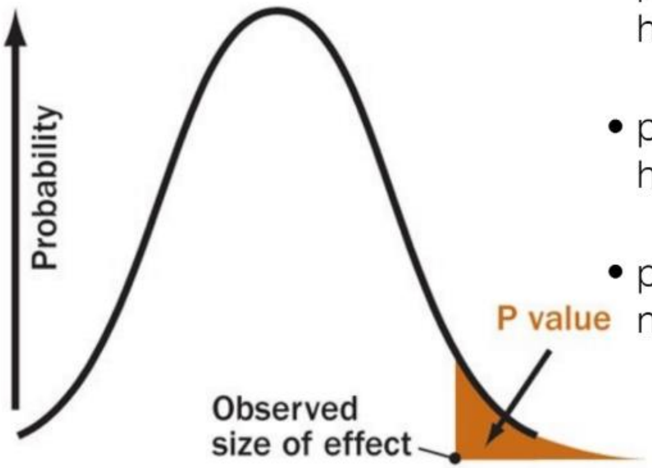


- Fissiamo un **livello di confidenza** (es. 95%)

se Z-score > 1.65, allora possiamo rigettare l'ipotesi nulla

P-Value

- Calculate the probability of an event more extreme that the observation under the “null hypothesis”



- $p < 0.05$ Moderate evidence agains null-hypothesis
- $p < 0.01$ Strong evidence against null-hypothesis
- $p < 0.001$ Very strong evidence against the null-hypothesis

- The smaller the p-value the better.

A/B test tool

http://abtestguide.com/calc

AB **Testguide**

Is your test result significant?

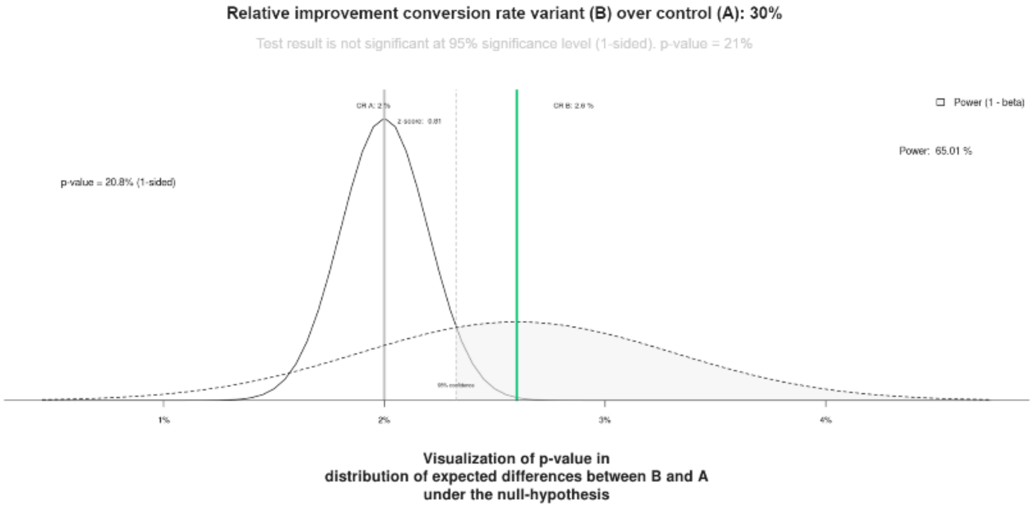
Does it have enough power?

Play with the controls and get a better feel for how a lower confidence level will boost the power or how an increase in test size can make a small CR-difference significant!

Pre-test calculation or post-test evaluation?

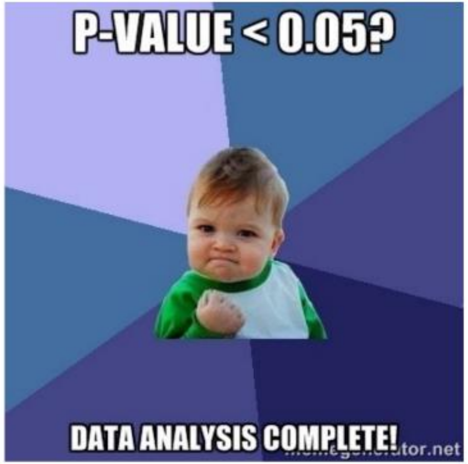
- ( ) Pre-test analysis
- (•) Test evaluation

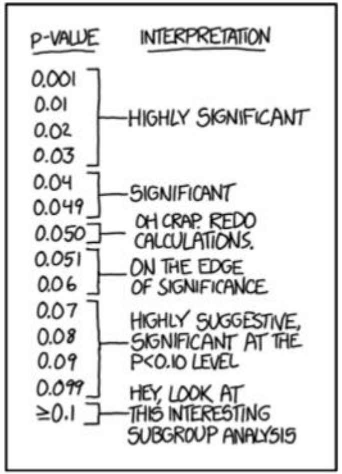
Visitors A:	Conversions A:
5000	100
Visitors B:	**Conversions B:**
500	13

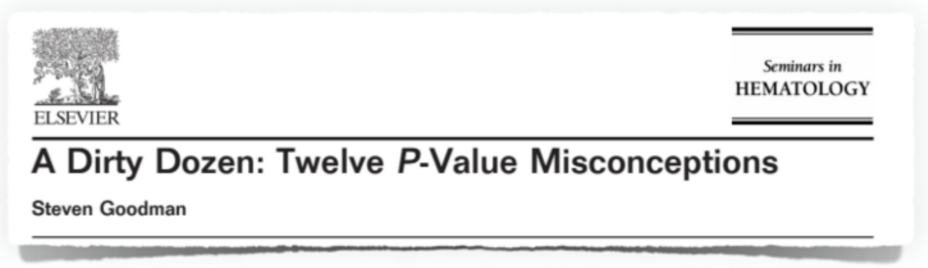


P-Value (misuse)

credits: xkcd






ELSEVIER

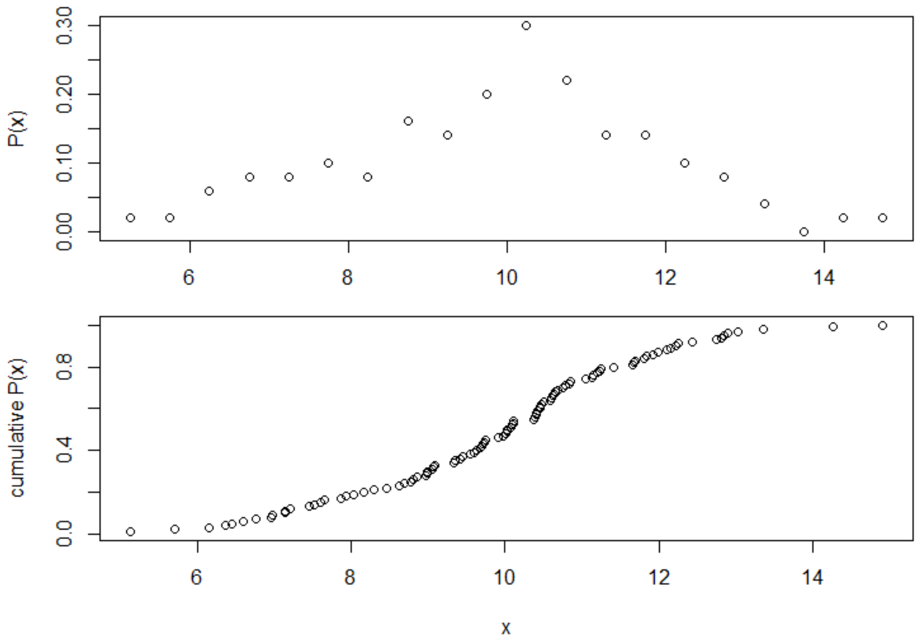
Seminars in HEMATOLOGY

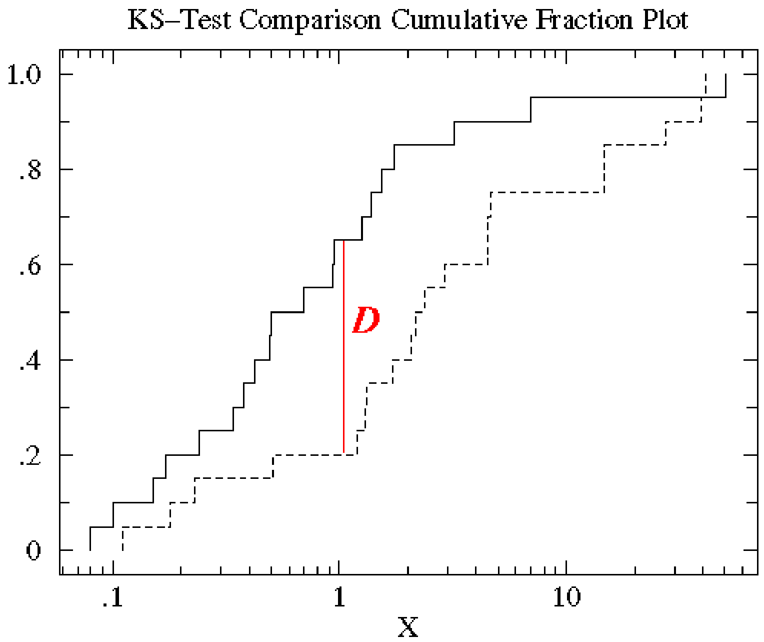
A Dirty Dozen: Twelve *P*-Value Misconceptions

Steven Goodman

Kolmogorov-Smirnov test

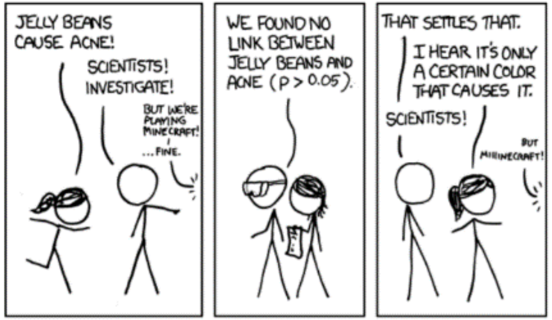
KS è un test **non parametrico** per verificare se due gruppi differiscono significativamente.

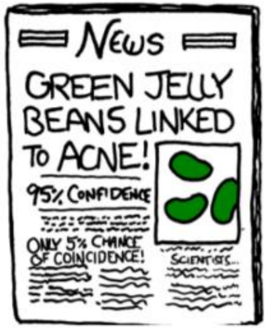




Confronto fra gruppi - Bonferroni

ATTENZIONE: se vengono effettuati controlli multipli, la probabilità di trovare differenze dovute al caso aumenta!





Confronto fra gruppi – A/B Testing

Un esempio ‘big data’:

PNAS

CrossMark
click for updates

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

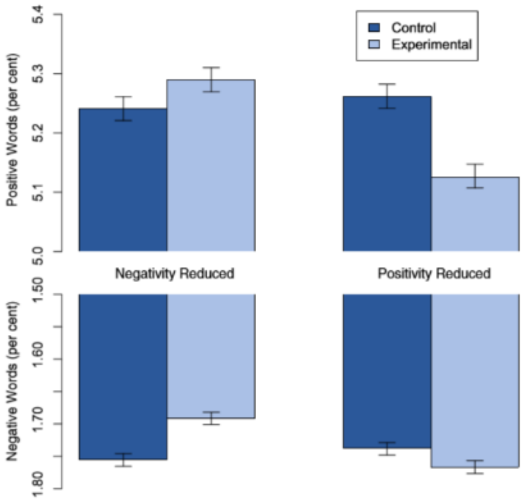
Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks. This work also suggests that, in contrast to prevailing assumptions, in-person interaction and nonverbal cues are not strictly necessary for emotional contagion, and that the observation of others' positive experiences constitutes a positive experience for people.

computer-mediated communication | social media | big data

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook's "News Feed" product (8). Because people's friends frequently produce much more content than one person can view, the News Feed filters posts, stories, and activities undertaken by friends. News Feed is the primary manner by which people see content that friends share. Which content is shown or omitted in the News Feed is determined via a ranking algorithm that Facebook continually develops and tests in the interest of showing viewers the content they will find most relevant and engaging. One such test is reported in this study: A test of whether posts with emotional content are more engaging.

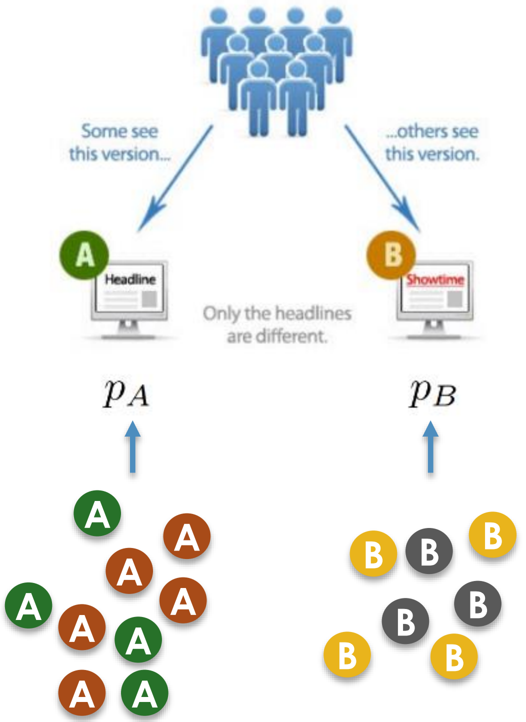
The experiment manipulated the extent to which people (N = 689,003) were exposed to emotional expressions in their News Feed. This tested whether exposure to emotions led people to



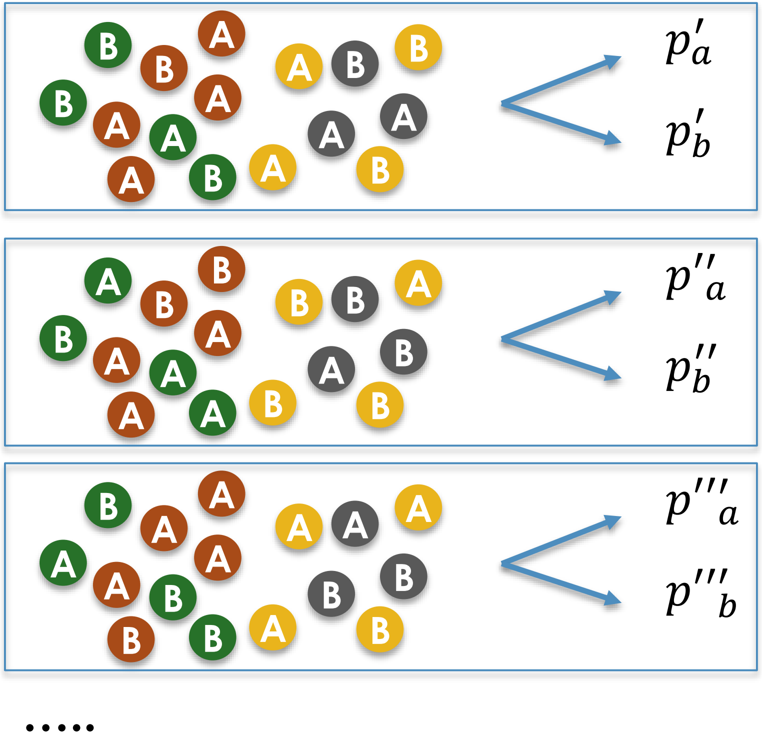
AUSTRALIA
EUROPE
USA
TECHNOLOGY CONSULTING

Computational approach for testing hypothesis

Keywords: bootstrap, resampling

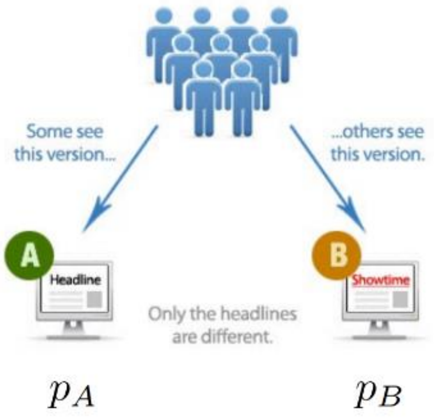


Randomize labels: Keep 'A' & 'B' sample size

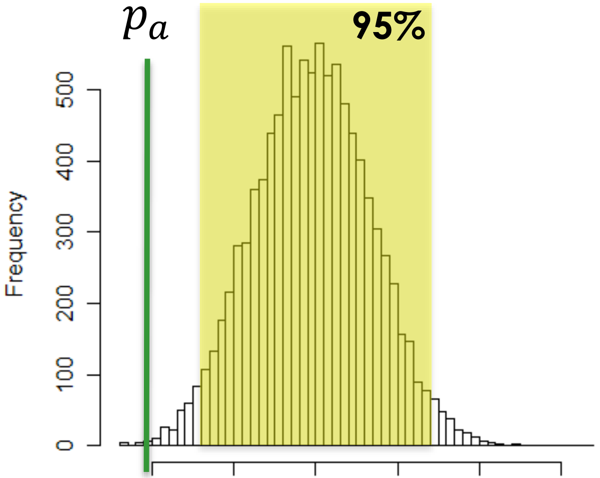


Computational approach for testing hypothesis

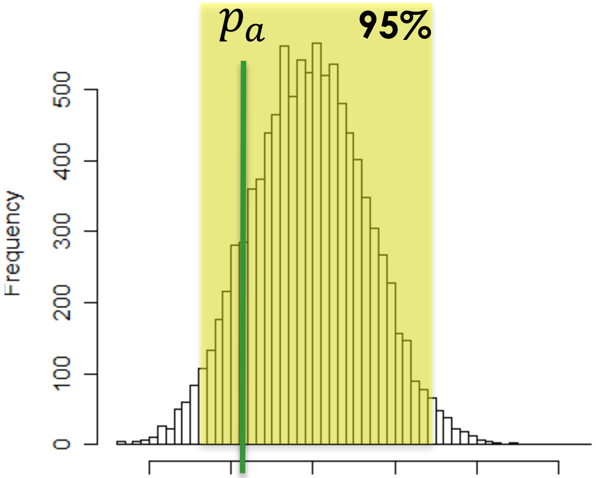
Keywords: bootstrap, resampling



Si verifica se il valore misurato di p_a è compatibile con valori di p_a ottenuti da una **distribuzione casuale** di utenti con outcome positivo/negativo.



L'effetto di A vs. B impatta **significativamente** l'outcome



Effetto **non significativo:** esistono molti casi in cui il valore p_a può essere effetto del caso

Q & A