«Se torturi i numeri abbastanza a lungo confesseranno qualsiasi cosa»

Stelvio Cimato

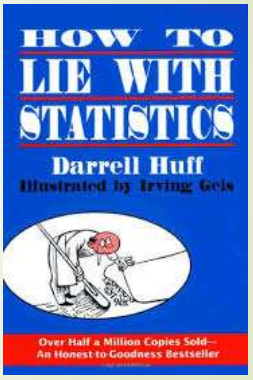
Università degli studi di Milano

Ronald Coase

- **"If you torture the data long enough, it will confess to anything."**
- premio Nobel per l'economia nel 1991.
- Critica alla «econometrics»
  - Utilizzo della statistica in ambito economico
  - Analisi dei fenomeni economici basandosi sulle osservazioni dei dati
  - Trovare relazioni nei dati per supportare un modello economico

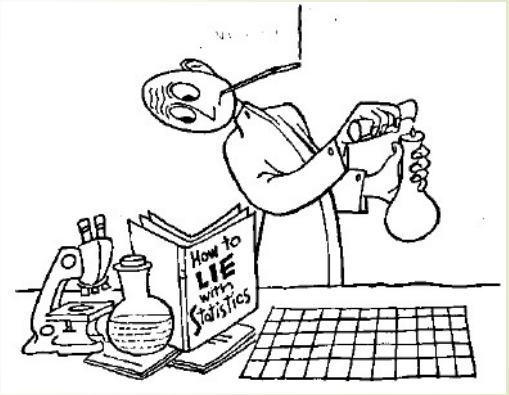
Citazioni celebri sulla statistica

- ***Me spiego: da li conti che se fanno secondo le statistiche d'adesso risurta che te tocca un pollo all'anno: e, se nun entra ne le spese tue, t'entra ne la statistica lo stesso perché c'è un antro che ne magna due.***
  (Trilussa)
- ***Se vuoi ispirare fiducia, dai molti dati statistici. Non importa che siano esatti, neppure che siano comprensibili.***
  Basta che siano in quantità sufficiente.
  (Lewis Carroll)
- ***Le statistiche dicono che uno su quattro soffre di qualche malattia mentale. Pensa ai tuoi tre migliori amici. Se stanno bene, vuol dire che sei tu.***
  (Rita Mae Brown)



How to lie with statistics – D. Huff (1954)

- Statistica è necessaria
- ..ma senza:
  - Onestà degli scrittori
  - Comprensione dei lettori
- ..si ottengono risultati non veritieri
- «A well-wrapped statistics is better than Hitler's big lie»
- A differenza della matematica
  - Statistica non fornisce risultati «certi»



Statistica: storia ed etimologia

- L'etimologia della parola "Statistica" deriva dal vocabolo italiano "Stato"
  - inteso come raccolta d'informazioni organizzate e gestite dallo "Stato".
  - Ghislini nel 1859, indica la Statistica come "descrizione delle qualità che caratterizzano e degli elementi che compongono uno Stato".
- L'evoluzione storica della Statistica nasconde due *anime*
  - La **Statistica** come una scienza che rappresenta uno strumento essenziale per la scoperta di leggi e relazioni tra fenomeni.
  - nel pensare comune dei non specialisti e, quindi, nel linguaggio dei mass-media le *statistiche* significano dati, tabelle, grafici, indici, medie.

Data mining

- Il processo di utilizzare diverse techniche di apprendimento automatiche per analizzare ed estrarre conoscenda dai dati (Roiger and Geatz, 2003).
- Utilizzando machine learning, statistica e tecniche di visualizzazione
- Trovare pattern nascosti nei dati per spiegare dei fenomeni

Produzione di dati

Big data

- Rivoluzione dei dati:
  - open & big data
- Dati strutturati e non strutturati
  - Tweets, conversazioni estratte dai social network
- Elevata variabilità nel tempo dei dati da analizzare, che richiede complesse analisi statistiche in contesti quasi-realtime
- La crescita del numero dei dati disponibili è impressionante:
  - ogni giorno le nostre attività su Internet generano circa 1 Exabyte (10 elevato alla 18) di dati.
  - Internet of Things potrebbe portare i dati generati quotidianamente a volumi nell'ordine dei Brontobyte (10 elevato alla 27).

Piramide della conoscenza

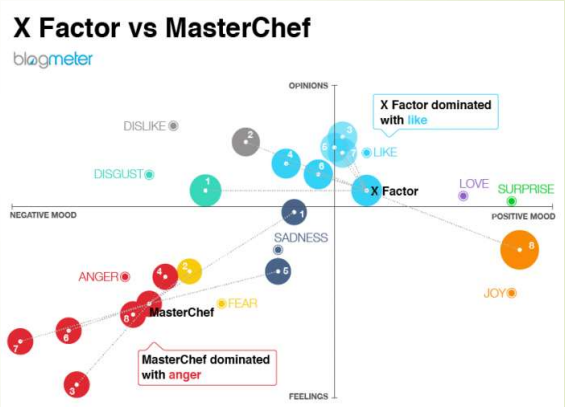
- La rilevanza dei dati per il mondo delle aziende:
  - capire meglio quel che è successo in passato
    - Evitare errori
  - comprendere tutte le caratteristiche del contesto presente
    - prendere decisioni "informate"
  - prevedere l'andamento futuro di eventi ancora in corso
    - o l'emergere di nuovi fenomeni

Applicazione di data mining: recommender system

Amazon.com increased sales by 15%, using data/text mining generated purchase suggestions

Applcazione di data mining: Sentiment Analysis

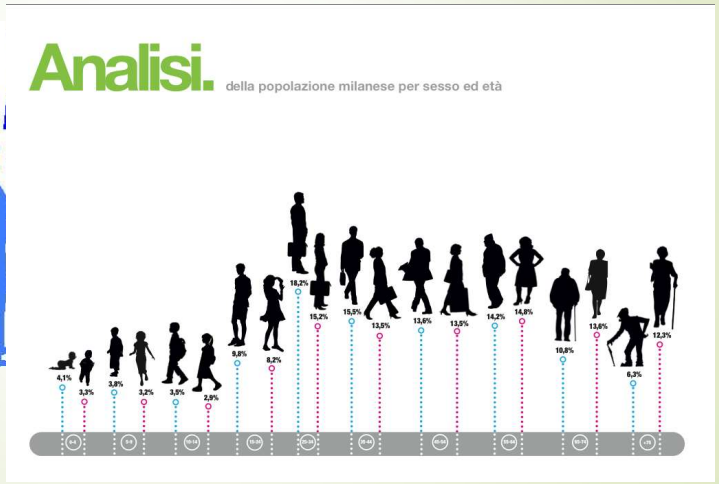
- Natural Language Processing classifica le opinioni espresse in rete
- Il processo di **comprensione automatica delle opinioni** estrae il mood positivo e negativo dai messaggi
- ..eassocia, inoltre, un punteggio che descrive l'intensità con cui l'opinione è espressa nel documento (High, Medium, Low).



Statistica descrittiva

- utilizzo di alcune tecniche statistiche per comunicare ad altre persone brevemente, con logica ed ordine, le principali caratteristiche dei dati raccolti
- Trovare tendenze nell'evoluzione dei dati
- evidenziare tendenze inattese a priori che possono suggerire analisi non previste inizialmente o anche nuovi esperimenti o campionamenti
- Trarre alcune conclusioni

Analisi dei dati



Esempio analisi

Materia	Italiano	Storia	Geografia	Matem.	Scienze	Ed. Fisica	Totale
Maschi	5	4	4	2	6	5	26
Femmine	3	7	2	3	4	5	24

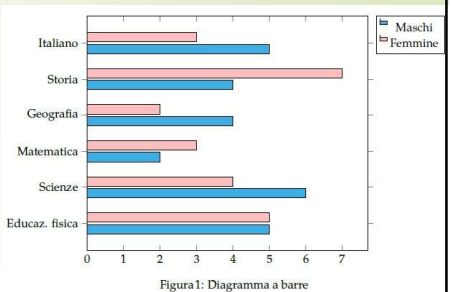


Figura 1: Diagramma a barre

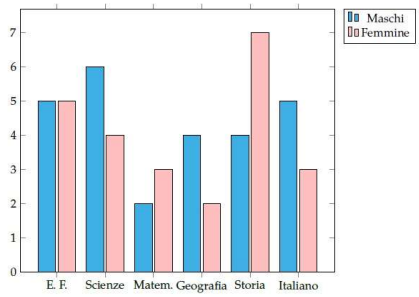
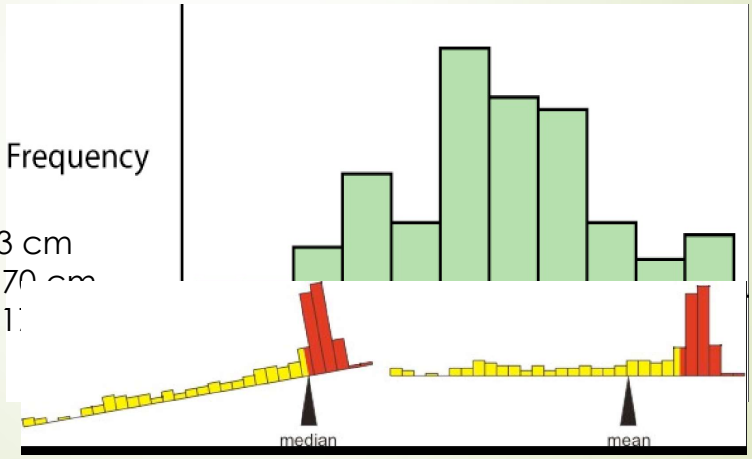


Figura 2: Diagramma a colonne

Media, Mediana, Moda

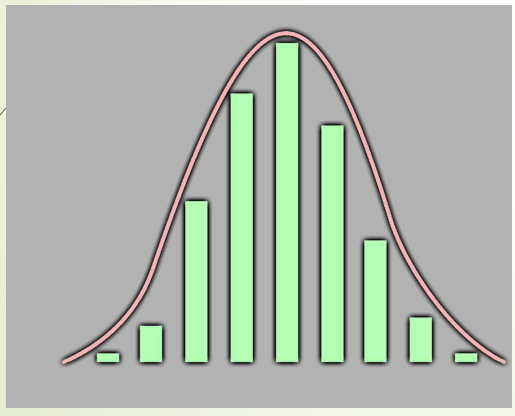
Classe di 28 studenti

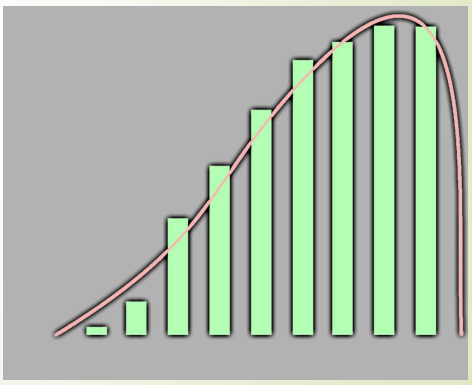
Media169.3 cm
Mediana 170 cm
Moda 165-1



Curva Normale

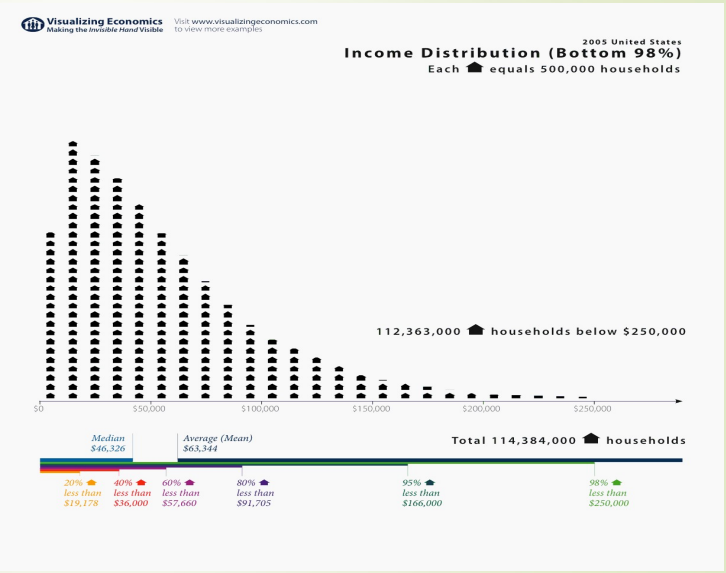
- Media e mediana coincidono





Reddito dei Possessori di case in USA nel 2005

- Median $46,326
- Mean $63,344
- Mode $5000-$9999



Esempio 1-Huff «How to lie with statistics»

- 3 soci di un'azienda con 90 operai
- 198.000$ di stipendi
- 11.000$ ciascuno per i soci in salari

Problema: come dividere 45.000$ di profitti?

Media degli operai **2.200**$

Media salari dei soci **26.000**$

Dei 45.000$ dividi fra i cosi
30.000 sono bonus
15.000 come profitti

Calcolo la media includendo i soci:

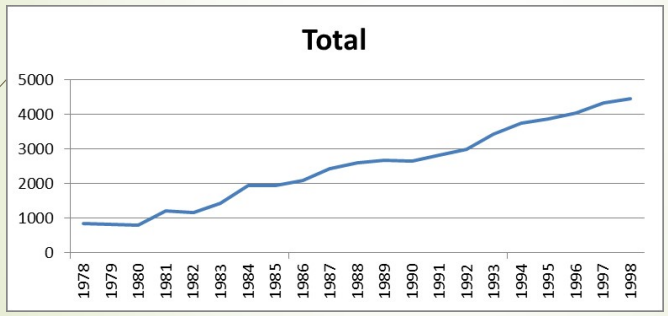
(198.000+33.000+30.000)/93

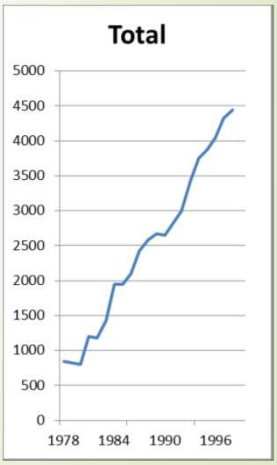
Media dei salari **2.806,45**$

Media profitti dei soci **5.000**$

Basta mostrarla nel verso giusto…

Crescita della popolazione

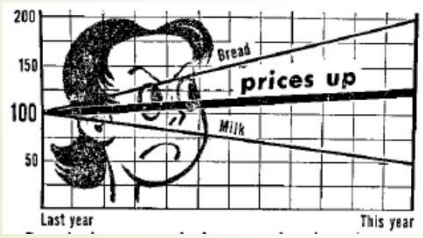




Esempio 2- Huff

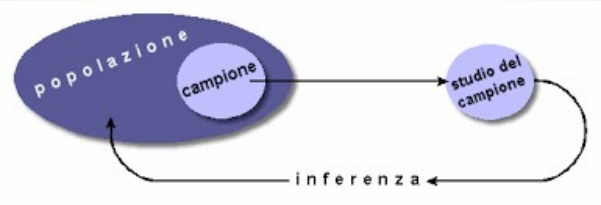
Costo della vita?

- Latte costa la metà
- Pane è raddoppiato
- Si parte dal presente:
  - Latte costava il doppio
  - Pane aumentato del doppio



Statistica inferenziale

- Scopo: generalizzare le osservazioni su uno o più campioni
- L'inferenza statistica è il processo attraverso il quale i risultati campionari vengono utilizzati per trarre conclusioni sulle caratteristiche di una popolazione
- E' necessario però anche aggiungere con quale grado di sicurezza, la stima o generalizzazione sia corretta



Campionamento

- Quanto grande il campione?
- Quanti campioni devo prendere?
- Qual è l'affidabilità dei risultati osservati?

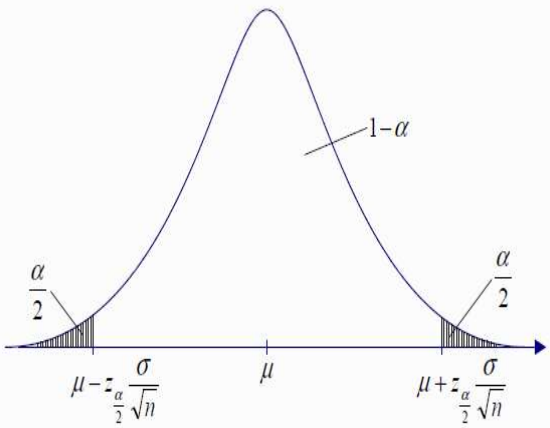
Proprietà di un buon campione

- In un campione casuale ogni membro deve avere la stessa probabilità di essere scelto
- Selezione indipendente
- Selezione random
- Abbastanza grande

- La distribuzione campionaria di una stima ci informa sulla precisione di una singola stima
- Teoria statistica dice come sono fatte le distribuzioni campionarie delle stime e delle statistiche in generali
- **Errore standard**
- **Intervallo di confidenza:**
  - Misura l'incertezza di una stima
  - E' un intervallo, definito intorno alla stima di un parametro, che ha alta probabilità di contenere il parametro (ossia, il valore vero)

Risultati del campionamento

- per *n*>30 osservazioni la distribuzione campionaria della media è con buona approssimazione normale
- Nel caso di campioni grandi uso la varianza campionaria
- Nel caso di campioni piccoli uso un'altra distribuzione



Esempio 3- Huff - Indagine statistica

- Indagine basata su questionario porta a porta:
  - Quale rivista c'è in casa?
    - Risultati Harper's e poche True Story
    - Dati contrari a quelli dell'editore
- Indagine sul reddito medio dei laureati a Yale classe '24
  - Risultati 25.000$
  - Campione
    - solo quelli raggiungibili
      - Quelli non di successo non interpellati
    - Risposte biased: mostrare reddito più alto o più basso
  - Il risultato vale per il campione: gruppo di quelli che vogliono esporsi e dichiarare il proprio status

Elezioni presidenziali in USA nel 1936

Alf Landon Campaign Poster, 1936

- 2.4 milioni risposte
- Basato su questionari spediti a 10 milloi di persone scelti dall'elenco telefonico, e dale liste dei club
- Previsione Landon wins
  - Landen 57% Roosevelt 43%

Risultati

- Bias nelle risposte:
- Il campione era format da gente che aveva un telefono o era iscritto al club
- Erano ricchi
- Il voto è correlato alla ricchezza personale

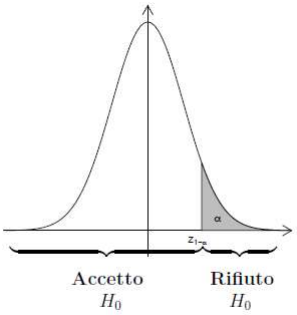
Test delle ipotesi

- Definite diverse ipotesi si cerca di identificare qual è l'ipotesi più adatta a spiegare i dati osservati
- In generale, si definisce un'ipotesi nulla H0 e un'ipotesi alternativa H1.
  - I dati sono sempre confrontati con quelli previsti dall'ipotesi nulla.
    - È l'ipotesi di totale casualità dei risultati
  - Se sono troppo diversi da quelli previsti dall'ipotesi nulla, si favorisce l'ipotesi alternativa.
  - Altrimenti, si conclude che i dati sono compatibili con l'ipotesi nulla.
- Una VERIFICA DI IPOTESI è una procedura statistica che permette di utilizzare le informazioni campionarie per saggiare un sistema di ipotesi, ossia per decidere se accettare H0 (e quindi rifiutare H1) o se rifiutare H0 (e quindi accettare H1).

Esempio ipotesi

Dati dell'ultimo censimento: altezza media della popolazione italiana pari a 173 cm.

- Ho un campione di 8 individui con altezza media 175 cm
- ? il fatto di aver osservato $\bar{x} = 175 > 173$
  - incertezza legata alla stima campionaria?
  - O altezza media aumentata rispetto all'ultimo censimento?
- a livello di significatività
  - a =0,05 indica che vi sono 95 probabilità su 100 che il risultato ottenuto sul campione non sia casuale.
- calcolo l'esatta probabilità di ottenere per solo effetto del caso il risultato osservato nel campione (H0 sia vera)
  - Ottengo una quantità p-value
    - Se p <0,05, il caso è una spiegazione improbabile
    - Se p >0,05, il caso è invece una possibile spiegazione



$$\begin{cases} H_0: & \mu = 173 \\ H_1: & \mu > 173 \end{cases}$$

Statistica correlazionale

- Si pone il problema di studiare la relazione fra due variabili casuali coefficiente di correlazione.
  - Il coefficiente può essere positivo o negativo e varia da -1.00 a 1.00.
  - Il valore dà la forza e la direzione della correlazione
- Indice chi-quadro
  - Valuto le variabili sotto l'ipotesi di indipendenza

$$\begin{cases} H_0: & X \text{ e } Y \text{ sono indipendenti} \\ H_1: & X \text{ e } Y \text{ non sono indipendenti} \end{cases}$$

al livello di significatività α.

Esempio

- un **campione** di studenti iscritti al III anno di **quattro licei scientifici**
- relazione tra il **rendimento scolastico** e il **livello di istruzione del capofamiglia**.
- La ricerca svolta su un campione di **180** studenti, ha dato una tabella di contingenza

	Lic media	Diploma	Laurea	Totale
Scarso	16	12	6	**34**
Sufficiente	30	55	13	**98**
Buono-ottimo	3	23	22	**48**
Totale	**49**	**90**	**41**	**180**

Esempio

- Valori attesi in caso di indipendenza
- Valore chi= 30,82
- Valore Pearson= 0.17
- Dipendenza!

	Lic media	Diploma	Laurea	Totale
Scarso	9,3	17,0	7,7	**34**
Sufficiente	28,7	49,0	22,3	**98**
Buono-ottimo	13,1	24,0	10,9	**48**
Totale	**49**	**90**	**41**	**180**

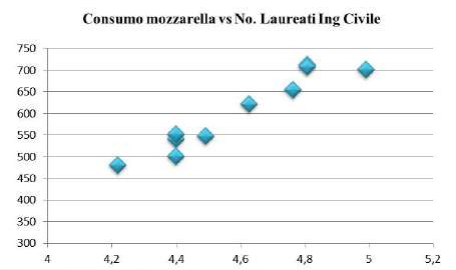
Correlazione non è causalità

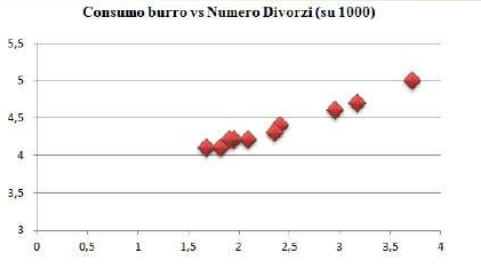
- Evidenza di una relazione non garantisce causalità!
- Considera numero di visite dal medico e Stato Socio-Economico SSE
  - La relazione è: Più alto il SSE, più visite
  - Possibili spiegazioni:
    - (SSE→visite) Persone con alto SSE possono pagare le visite
    - (visite→SSE) Persone che mantengono la propria salute, sono più efficienti ed hanno un alto SSE
    - C'è una terza variaile nascosta X che influenza sia le visite che SSE (dimensione della città in esame)
      - (SSE←X→visite)Città grandi hanno più cliniche e più opportunità lavorative

Esempi di correlazione

- Numero di volte che un grillo canta è legato alla temperatura
- Crimini e agenti di polizia in una zona
- Numero di gelati venduti e assassini a New York
  - In realtà la temperatura correla con entrambi

Insidie della correlazione





Paradosso di Simpson

- Classe A
- 1 ragazza media 28
- 5 ragazzi con 27,26,25,24,23.

- In totale
  - 3 ragazze media 23
  - 6 ragazzi media 24

- Classe B
  - 2 ragazze con 21, 20
  - 1 ragazzo con 19.

On the Predictive Power of University Curricula – A. Azzini, P. Ceravolo, E. Damiani, N. Scarabottolo

- Indagine sul curriculum di 65 studenti universitari
- Correlazione fra performance e attività scelte
  - **H1.** Cases significantly different in terms of process development are likely to be significantly different in terms of performances
  - **H2.** Cases significantly different in terms of process development are likely to be significantly different in terms of activity order.
  - **H3.** Cases significantly different in terms of process development are likely to be significantly different in terms of activity type.

Huff – Suggerimenti per evitare inganni

- Chi lo dice?
  - Cerca bias voluto o non voluto
- Come fa a saperlo?
  - Guarda la selezione del campione
- Cosa manca?
  - Quanti casi esaminati? (sondaggi, correlazione)
  - Manca un fattore determinante (es. Pasqua)
- Qualcuno ha cambiato l'oggetto di discussione?
  - Più casi denunciati di una malattia/crimine, non sono più casi di malattie/crimine
  - Si parla di qualcosa come causa di un'altra

Conclusioni

- Lo straordinario aumento degli scambi e delle interrelazioni genera problemi complessi
- La missione della statistica ufficiale è pertanto quella di fornire informazioni comprensibili e confrontabili
  - prodotte sulla base di definizioni e metodi rigorosi
  - fondate su dati di fatto e non su pregiudizi e conoscenze episodiche.
- Tecniche di data mining avanzate
  - Consentono di estrarre maggiore conoscenza dai dati

Ronald Coase e Mark Twain

- «Se torturi i numeri abbastanza a lungo, la **natura** confesserà»
- Evitare gli inganni dell'uso dei dati
  - far credere che questi dati supportino una certa verità o interpretazione della verità
  - intento doloso o colposo per ingannare il destinatario del messaggio
  - finalità diversa da quella di informare
  - Chi legge a volta preferisce farsi ingannare, sentendo una storia che è più piacevole della verità.
- «Ricerche, sondaggi e statistiche sono come un lampione. Utile quando serve a fare luce. Ma non dobbiamo comportarci come l'ubriaco che ci si appoggia»

Credits

- How to lie with statistics – D. Huff
- Dr. Michael Whitlock, Department of Zoology, BIOL 300: Biostatistics http://www.zoology.ubc.ca/~whitlock/bio300/
- **Torturing Numbers** Dr. Jason S.T. Deveau Application Technology Specialist OMAF & MRA, Simcoe Station **A Grower's Guide to Descriptive Statistics**
- E. Di Nardo, Universita degli Studi della Basilicata, V Giornata Italiana della Statistica 2015
- **Pitfalls of Data Analysis**, Clay Helberg, M.S.
- P. Bortot – Corso di statistica – Lecture notes – Università di Bologna
- G. Bertorelle – Corso di biostatistica – Università di Ferrara
- P. Pasetti – Corso di statistica sociale – Università di Ferrara