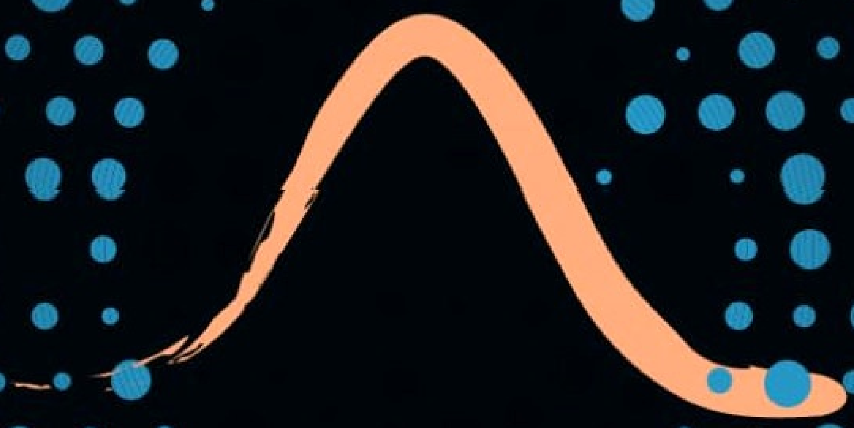


Modern Statistics

Intuition, Math, Python, R

Mike X Cohen

sincXpress

Modern Statistics - Kindle ebook

1 - Introduction to this book

La statistica riguarda l'uso dei dati per aiutare a prendere decisioni di fronte all'incertezza.

Molte procedure statistiche alla fine danno una probabilità "P".

1,2 - Statistica, scienza dei dati, apprendimento ..

- statistica = collezione, organizzazione, analisi, interpretazione, presentazione dei dati
- apprendimento = uso di "pattern" contenuti nei dati per predire o fornire in uscita una classificazione
- scienza dei dati = riguarda prevalentemente le applicazioni più recenti - prima ignorate - come immagini, serie temporali, testi

1,3 - Target audience

Libro non orientato esclusivamente alla matematica della statistica (statistica teorica), ma focus su concetti e applicazioni pratiche (Python).

1,4 - Prerequisites

- high school math / - Calculus & Linear Algebra //
  programming
- ChatGPT transcodifica con buona efficienza linguaggi in SAS, MATLAB, Julia

1,5 - Exercises

Sono forniti i sorgenti Python e R.
Al capitolo 20 brevi descrizioni degli esercizi.

1,6 - Learning from simulated data

- a partire da esposizioni nella letteratura, →

Spesso ciò proviene da letteratura "classica", non "computer-oriented", questo approccio può essere utile x statistica teorica.

- usare big data da letteratura, ma spesso avremo esempi poco correlati con ciò che ci interessa.
- usare Tutorial SW x implementare, questo va bene se già si conosce la statistica
- SIMULAZIONE dei DATI – è l'approccio di questo libro. Con esperienza e buon senso questo conduce a risultati che non sono troppo scorrelati dalla realtà.

1.7 - Using the code in this book

- è bene personalizzare i codici forniti. Se si producono cose nuove, utili, passarle al blog dell'autore.

1.8 - Online resources

- vedi info x downloads
- ChatGPT-4 – il suo utilizzo è concesso nel libro.

— FINE CAPITOLO 1 —

2 - What are (is?) "data"?

- plurale o singolare? meglio plurale.

2.2 - Where data come from, what do they mean?

- i dati in forma numerica che usiamo "non sono la realtà", ma una sua rappresentazione. Sono misure che abbiamo fatto sulla realtà!
- le qualità e la precisione dei dati sono fondamentali.

2.3 - What do data look like?

- i dati sono memorizzati in un computer. Spesso sono organizzati come tabelle (righe, colonne).

riga = osservazione colonne = misure

esempio:

Giorno	Coffee	Gradimento
lunedì	Iced-cappuccino	10/10
martedì	Espresso	8/10
mercoledì	Iced-mocha latte	6/10
—	—	—
—	—	—

	misura A	B	C		
osservazione 1	15,84	23,60	-15,76	—	—
— 2	18,52	32,15	4,15	—	—
— 3	37,86	-11,67	43,46	—	—
—			—	—	—
—	—	—	—	—	—

2.4 - Limitations of data

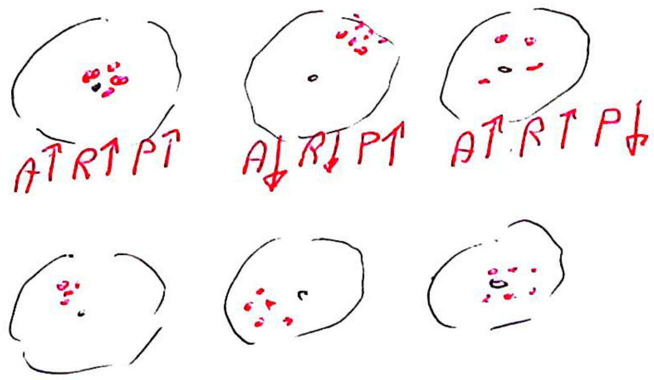
- i dati provengono da misure imperfette (necessariamente), es. la temperatura di una camera di scoppio di un motore non può essere misurata direttamente.

- nelle malattie virali conosciamo solo i casi dichiarati in modo ufficiale, il resto ci sfugge.
- i dati sono soggetti a "rumore"
- ci possono essere dati "fuori dalla ragionevolezza", "skew statistics" (quelli fuori dal 3σ)
- i dati sono espressi secondo unità di misura, $\exists$ metodi per trasformarli in numeri adimensionali.

2.5 - Accuracy, precision, resolution, range

- accuratezza (A) = molto legata alla qualità dello strumento di misura
- precisione (P) = capacità di dare misure simili a fronte di ripetizione della misura
- risoluzione (R) = la distanza numerica tra due successive misurazioni. Es. un termometro misura 200°, 225°, 250° ma non gli intervalli intermedi.

esempio grafico:



2.6 - Data Types

- non confondere con tipo dato del SW
- possono essere "numerical" o "categorical" (etichette)

esempio:

Family	Type	Description	Example
Numerical	Discrete	No arbitrary precision	Population
	Interval	Meaningful intervals, arbitrary precision	Temp. in °C
	Ratio	Interval but with meaningful zero	Height in cm
Categorical (labeled)	Nominal	Non-sortable, discrete	Movie genre
	Ordinal	Sortable, discrete	Education level

- ciascuna di queste caselle, dentro il PC e' tradotta in numeri. Es. BMW=1, Mercedes=2, ...

2.7 - From anecdotes to populations

- speculazione, teoria - $N=\phi$
- aneddoto $N=1$
- case report $N=1$ - in medicina
- sample - se voglio peso medio di un animale, non posso pesare tutti quegli animali esistenti, ma un sottoinsieme
- pilot study N piccolo
- small scale / large scale
- observational study — ho dati senza manipolazioni o forzanti esterne
- convenience sample — dati raccolti con loro fatica (es. nucleo familiare, amici)
- population — ciò che si studia non riguarda il singolo

2.8 - Data management

Sistema di documentazione, archiviazione, memorizzazione. Backups.

2.9 - Ethics of making up data

- e' facile generare "fake data" - quello che conta e' saperli riconoscere -

— FINE CAPITOLO 2 —

3 - Visualizing data

3,1 - Why visualize data?

- per meglio capire, per sintetizzare, per trovare errori_

3,2 - How visualize data

- per dati poco numerosi, li si possono visualizzare tutti

Lun	Mar	Mer	Gio	Ven	Sab	Dom
19,30	19,12	18,47	20,04	21,25	21,10	18,28

3,3 - Bar plots

es. dove la gente prende le sue informazioni?

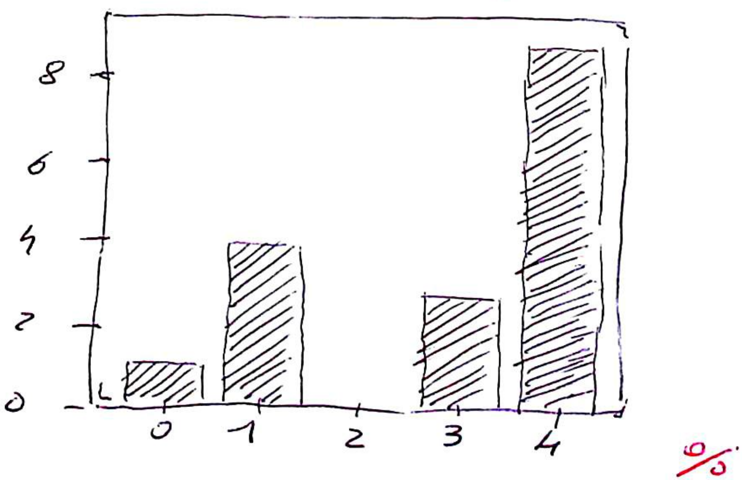
Fig. 3,3

sono equidistanti
ordine arbitrario

non necessita il colore

```
import matplotlib.pyplot as plt
Y = [1,4,3,9]   # altezza delle barre
X = [0,1,3,4]   # bar locations
plt.bar(X,Y)
```

Fig. 3,4



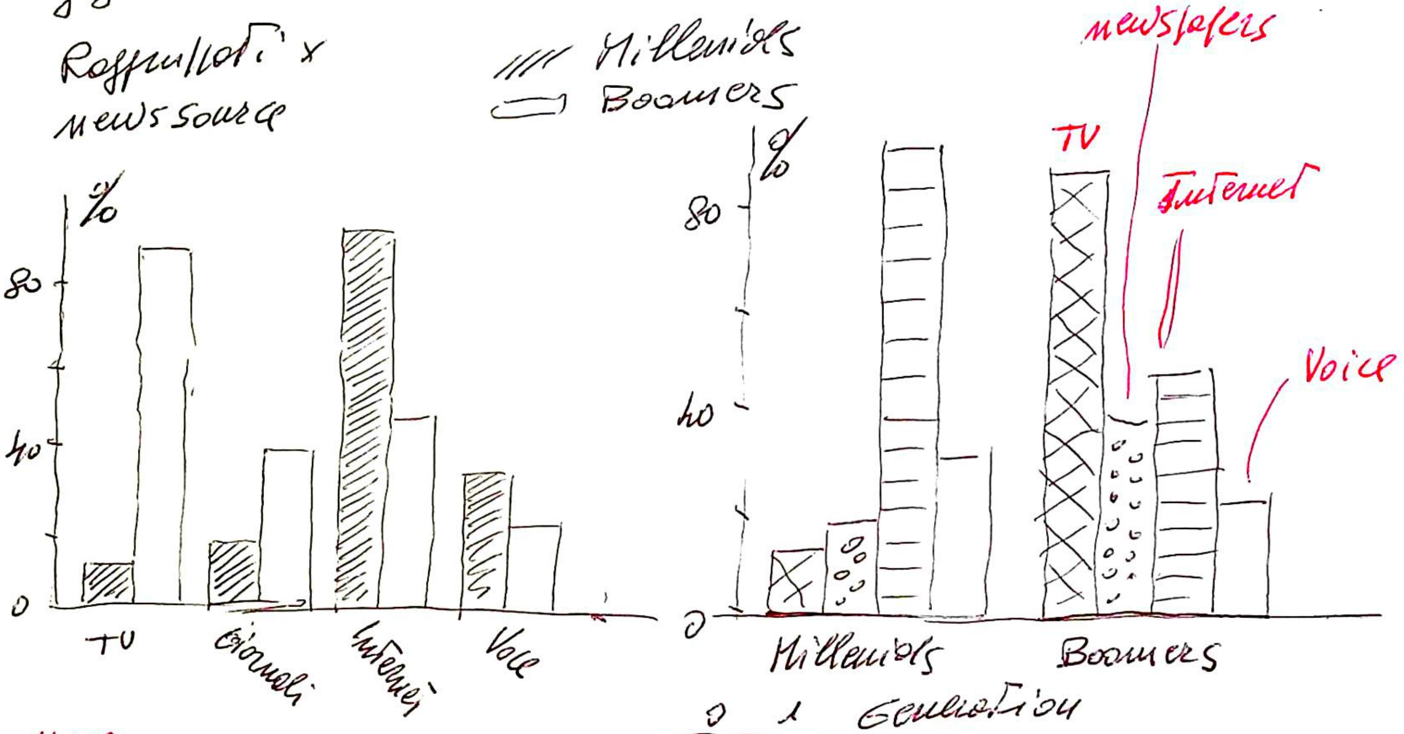
stats_ch03_visualization

— i dati possono essere rappresentati in modi differenti

news_sources = np.array([[12, 17, 95, 35], [90, 40, 50, 25]])

agegroups = ['Millennials', 'Boomers']

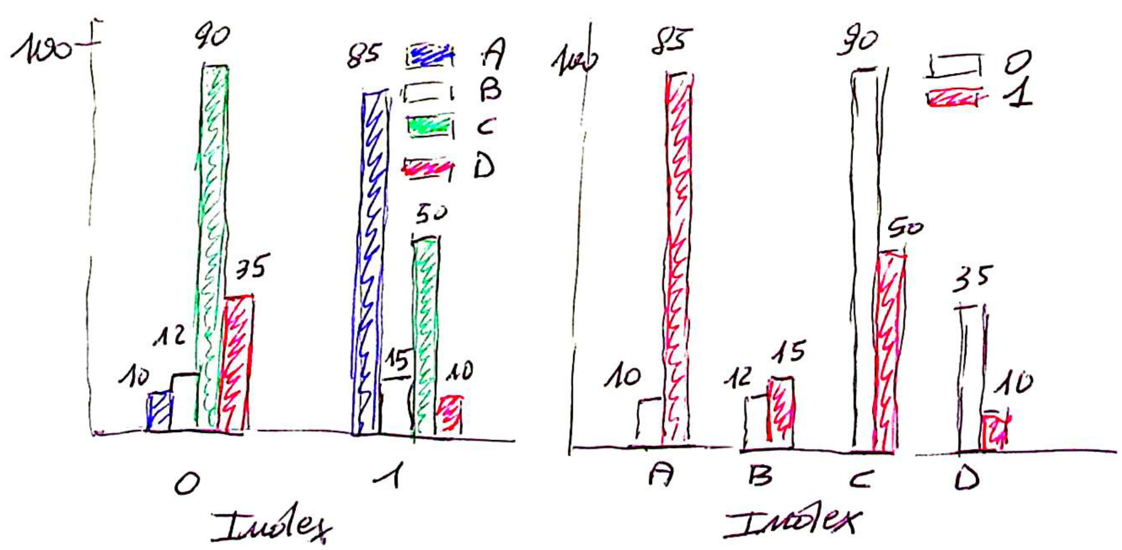
Rappresentati x news source



oppure

	A	B	C	D	
0	10	12	90	35	Observations
1	85	15	50	10	
	Features				

	0	1	Generation
Features	10	85	A
	12	15	B
	90	50	C
	35	10	D
	Observations		



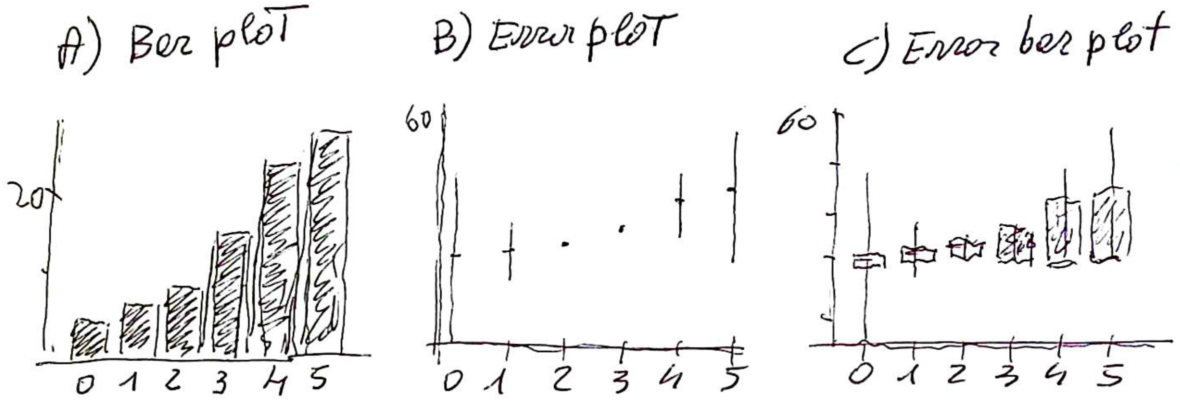
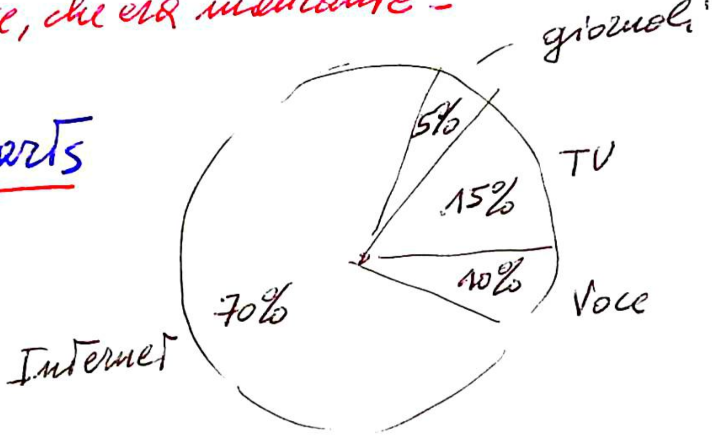


Fig. 3.7

l'errore non ha quindi una rilevanza fisica, è il risultato di una formula

sviluppato codice, che era mancante -

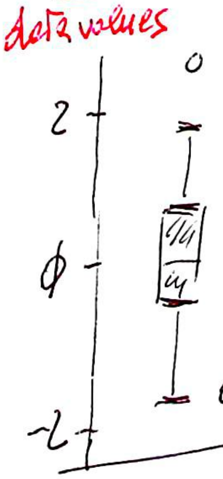


3.4 - Pie charts

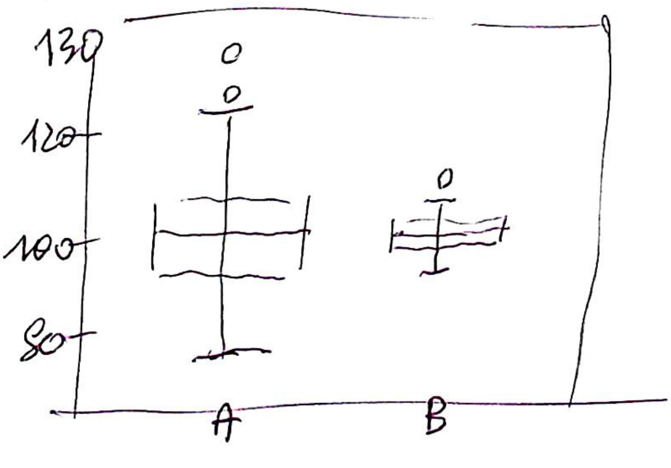
NB $\Sigma\% = 100\%$

3.5 - Box plots

plt.tight_layout()



f = outliers
e = massimo valore non considerato outlier
c = 75% dei dati
a = mediana (50% dati)
b = 25% dei dati
d = il minimo valore non considerato outlier



due distribuzioni di dati con uguale media e mediana, ma differente variabilità -

Possono essere confrontati

3.6 - Histograms

Appaiono simili a grafici a barre, ma sono molto diversi:

- bar plots = "categorical data"
- histograms = "numerical —"

es. X = [1, 2, 2, 3, 3, 5, 5, 5, 5, 6, 7, 7, 7, 8, 8, 9]

```
plt.hist(X, bins=len(set(X)), color='gray', edgecolor='k')
plt.xticks(np.arange(np.min(X), np.max(X)+1))
plt.tight_layout()
```

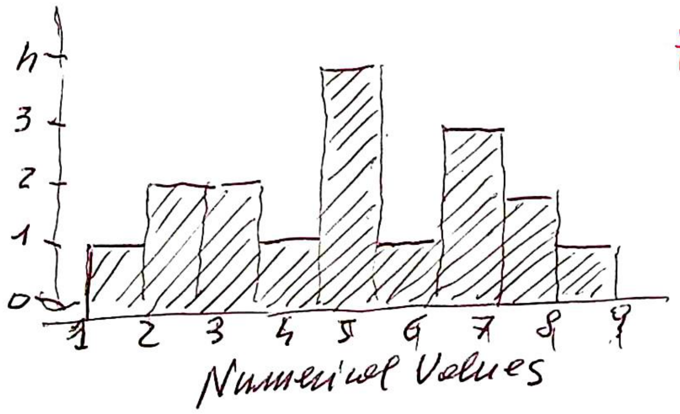


Fig. 3.11

i ticks mark non sono al centro della barra →

facciamo una modifica

```
plt.hist(X, bins=np.arange(.5, 9.51, step=1), color='-', edgecolor='-')
```

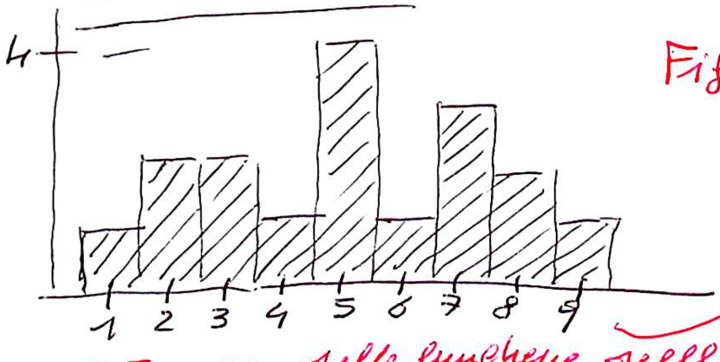


Fig. 3.12

molto importante x istogrammi multipli

es: istogramma delle lunghezze delle manguste

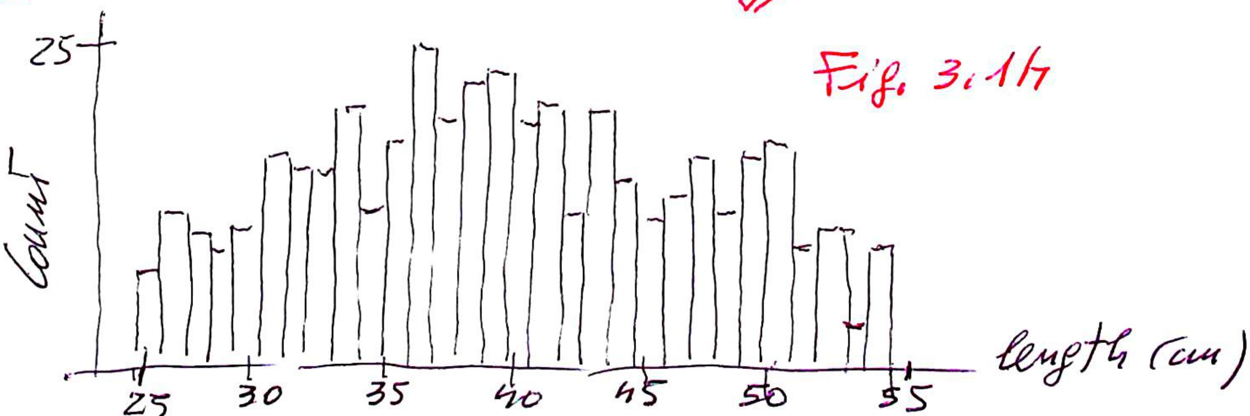


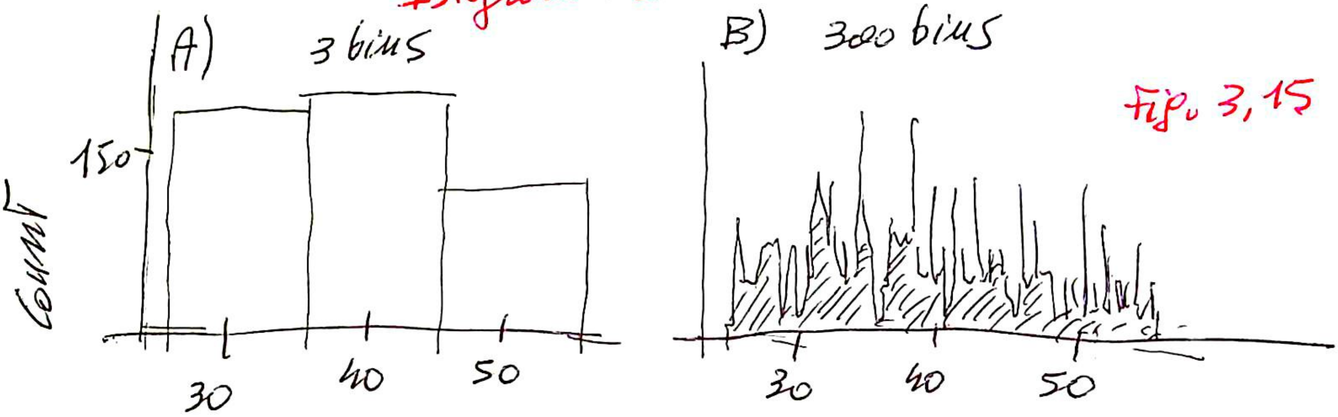
Fig. 3.14

Istogramma delle mangoste

mongooses = np.arctanh(np.random.uniform(-.75, .75, size=500))*15+40

plt.hist(mongooses, bins=30, color ———

Istogrammi con Bins differenti



quale il n° giusto di bins? — in concreto 30-40

Distribution tails

se vanno attorno allo zero → negative tail / positive —

— one tailed
— two tailed

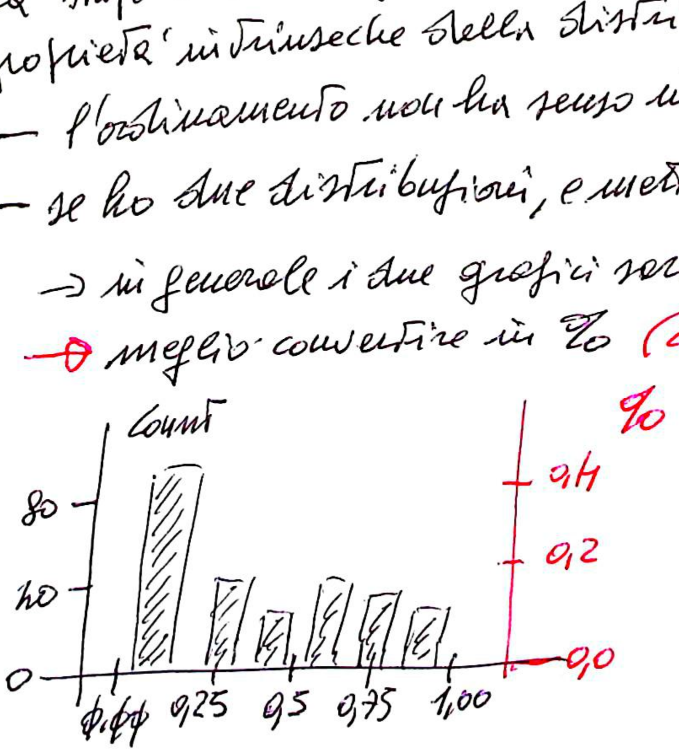
La "shape" di un istogramma — ordinando i dati lungo x — riflette proprietà intrinseche della distribuzione.

— l'ordinamento non ha senso in un "barplot" (labels)

— se ho due distribuzioni, e metto su y i dati grezzi →

→ in generale i due grafici saranno difficilmente confrontabili

→ meglio convertire in % (con questo cambia asse y)



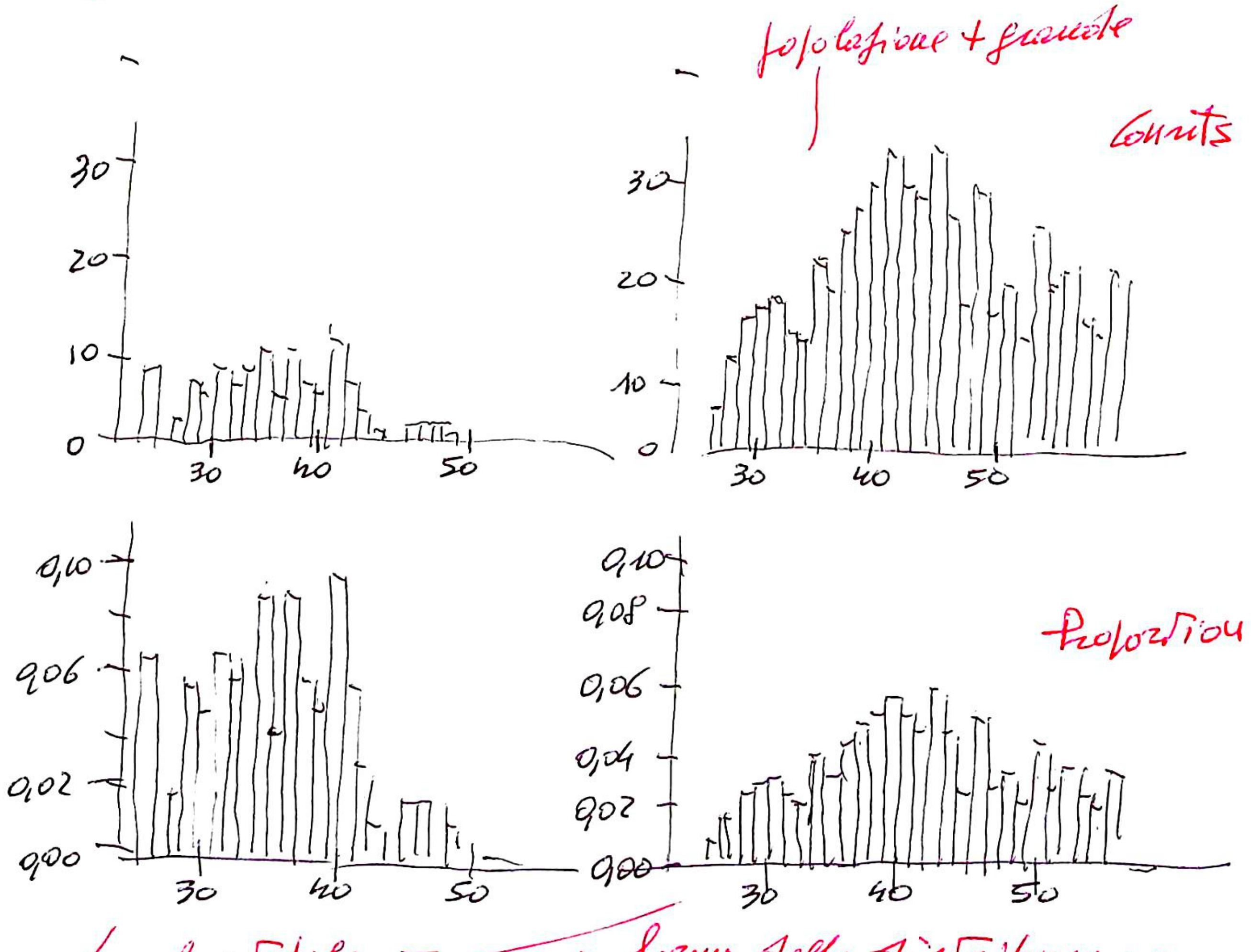
Raw counts	Proportion
Facile interpretazione	Sforzo x ricondurre ai dati
Difficile comparare campioni	Facile da comparare
Σ arbitraria	$\Sigma = 1$ (o 100%)
buono x ispezione qualitativa	buono x ispezione qualitativa

Fig. 3.18

Come convertire "counts" a "proporzioni"?

$$\tilde{b}_j = \frac{b_j}{\sum_{i=1}^{K} b_i} \qquad (3,2)$$

$\Sigma = 1$ — Comportamento dei due tipi di grafici



confrontabili $\Sigma = 1 \rightarrow$ forma della distribuzione

rivela le differenze di struttura

3,7- Lines vs, bars in a histogram

- se le barre sono sufficienti → ha senso tracciare una linea tra bins

3,8- Violin plots

— ci sono deformazioni

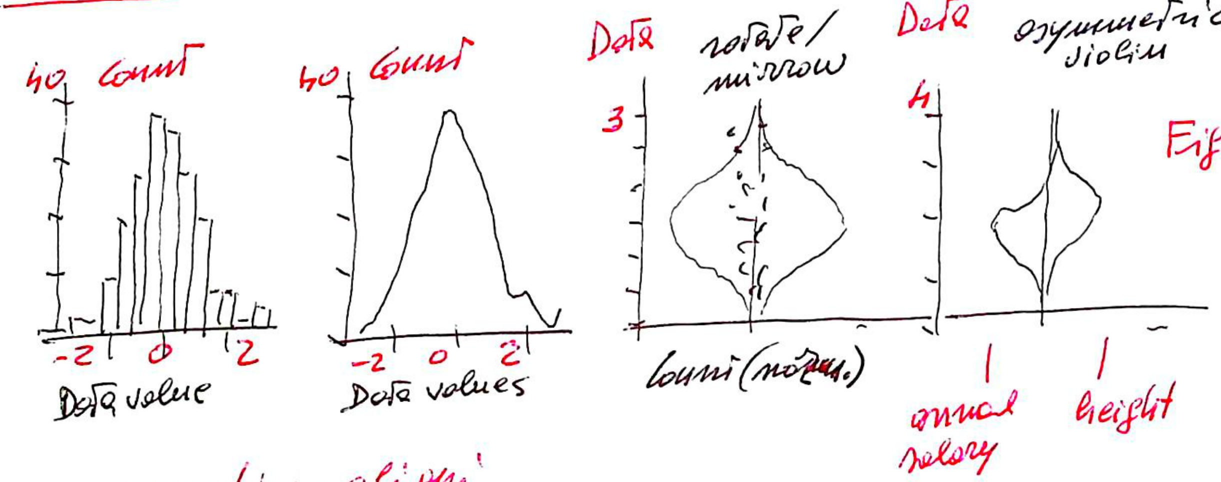
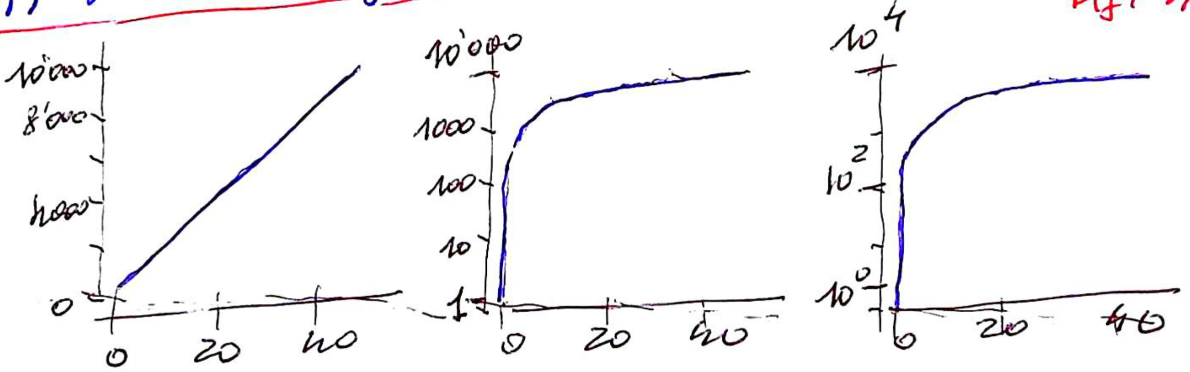


Fig. 3,21

- si scambiano gli assi

3,9- Linear vs. logarithmic axis scaling

Fig. 3,22



3,10- Discretizing continuous data

E' importante nel decidere se perseguire ANOVA o regressione, e nella visualizzazione in una "regression analysis" -

- per capire meglio visualizziamo la figura 3,23

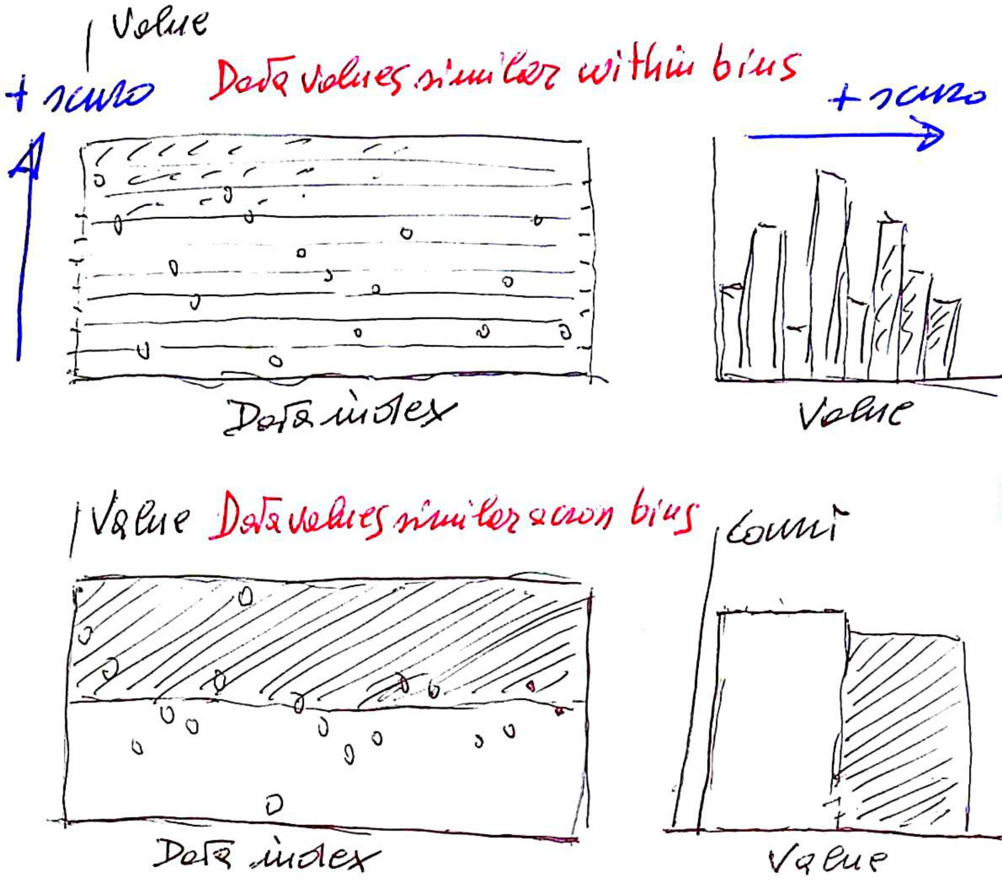


Fig. 3,23

- campionare è quindi una problematica complicata

3.11 - Radial plots

- utili quando sono "wrapping around"
- la distanza dal centro indica la magnitudine (equivalente a altezza)

High temps (°C) near Patagonia

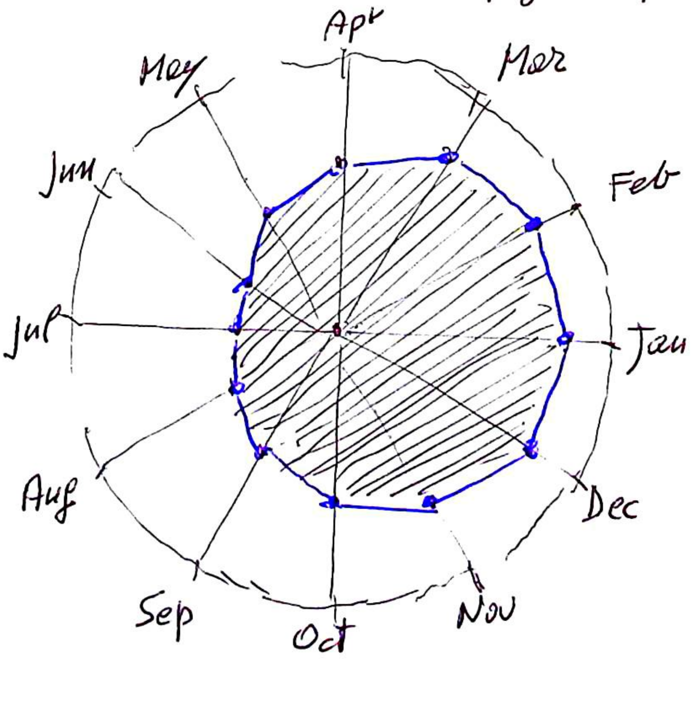


Fig. 3,24

```
ax = plt.subplot(111, polar=True)
ax.plot(theta, tempC, 'ko')
```
— pallini

```
ax.set_xticks(theta[:-1])
ax.set_xticklabels(months)
ax.set_yticks([10,20,30])
ax.set_ylim([0,30])
ax.set_title('________')
plt.tight_layout()
plt.show()
```

3.12 Color

in matplotlib → color = (R, G, B) — tra 0 e 1

- colore può dare chiarezza, ma anche confondere
- usare con cautela
- colori / se tenoti in "tonalità di grigio" riconoscibili

3.13 - Exercises

disegnare x righe e per colonne, data la

	0	1	2	3
A	0	3	6	9
B	1	4	7	10
C	2	5	8	11

creare random data x error bar plot
30 osservazioni, 6 misure, date da:

$Y_i \sim N(\mu_i, \sigma_i^2)$ num. casuali:

$\mu_i = (i+1)^2$ distr. norm. media μ

$\sigma_i = 30(2i/5-1)^2$ " " dev. std $=\sigma$

np.random.normal(mu, sigma)

ottengo la Fig. 3.7

Graficare – 60 persone che esprimono un gusto <

Vaniglia
Strawberries
Pistachio
chocolate

– è appropriata una torta?

– Riprodurre la Fig. 3,28

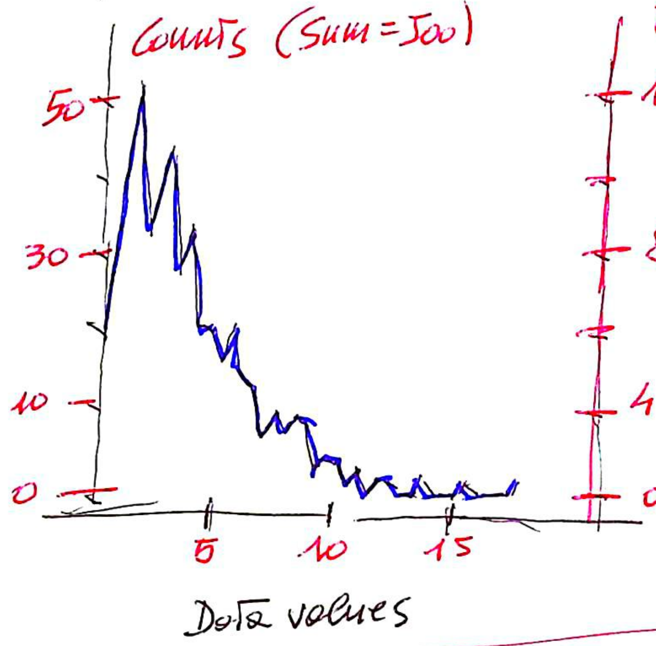


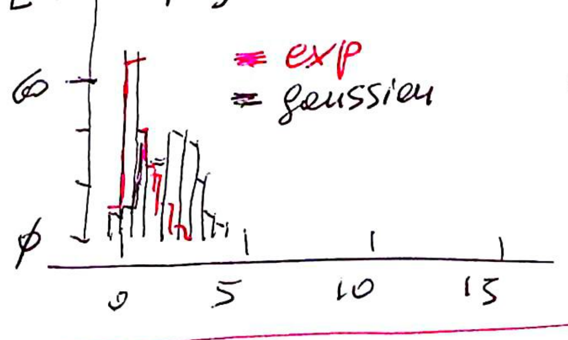
Fig. 3,28

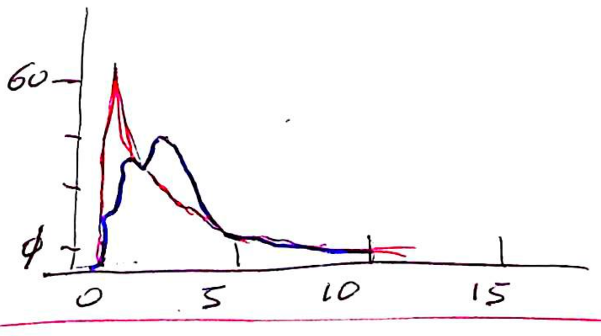
bins=40

– Barre o linee? N=200 < normale / esponenziale

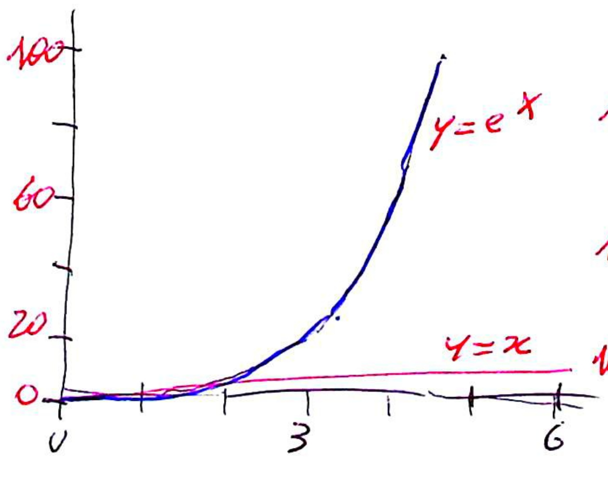
$G \sim \mathcal{N}(2,1)$ bins=30

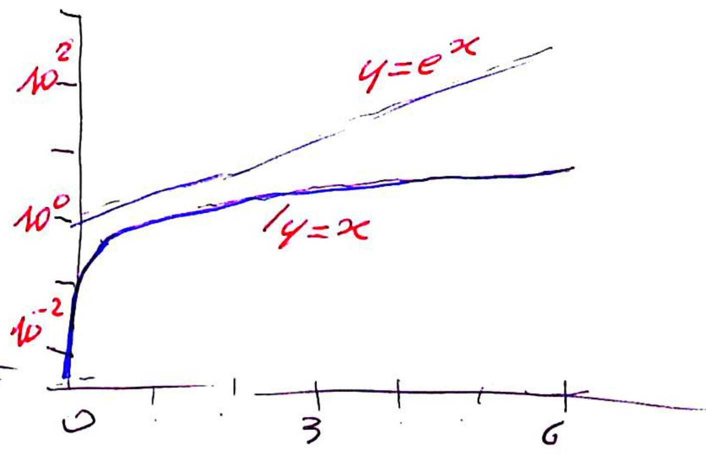
$E \sim \exp(\mathcal{N}(\phi,1))$ usare np.histogram()





– linear vs. logarithmic





Riprodurre la fig. 3,32 – 123 numeri random (normale, uniforme

memorizzati su 123 righe, 2 colonne → cioè due data sets, 123

np. random. uniform (0, 1, size = 123) df [[—], [—]]

=

sns. violinplot (data = df, palette = 'gray', ax = axs[0])

sns. stripplot (data = df, ax = axs[0], palette = 'dark:w')

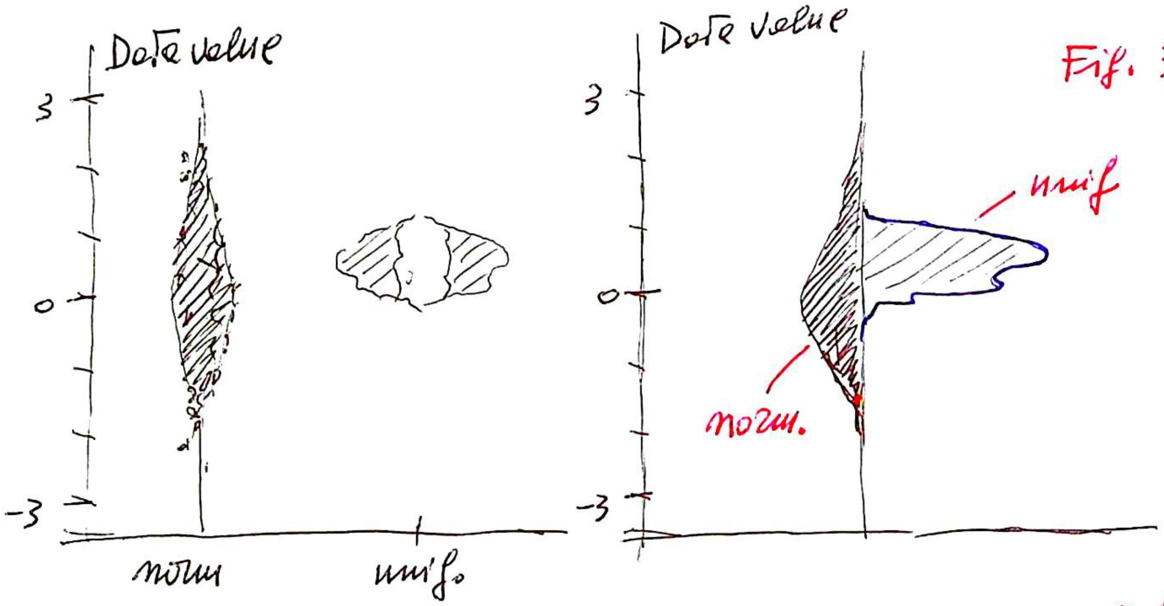


Fig. 3,32

p. ~~100~~

df = pd. DataFrame (np. hstack ((np. random. randn (123, 1), np. random. randn (123, 1))), columns = ['norm', 'unif'])

impila gli array in sequenza orizzontale (colonna)

0	-0,145	0,261
1	-0,845	0,888
2		
⋮	⋮	⋮
	norm	unif

(123 righe, 2 colonne)

(123, 2)

(123,)

non solo label, definiscono il tipo di distribuzione dobbiamo combinare le colonne in una

```
df_all = pd.DataFrame(pd.concat((df['norm'], df['unif'], axis=0)),
    columns=['y'])
```

→ df_all (246, 3)

nome della colonna di 246 righe

0	0,34	norm
1	0,61	norm
2	=	norm
3	=	=
	y	distr

```
df_all['distr'] = 'unif'
df_all['distr'][:len(df)] = 'norm'
```

ignorare warning

```
df_all[''] = '-'
print(df_all)
```

ora creiamo "split violin plot"

```
sns.violinplot(data=df_all, x='', y='y', palette='gray',
    ax=axs[1], split=True, hue='distr')
=
plt.tight_layout()
plt.show()
```

— Fine capitolo 3 —

4 - Descriptive Statistics

4.1 - Descriptive vs. inferential statistics

- Statistica descrittiva = numeri che caratterizzano un set di dati (media, mediana, varianza, inclinazione, spettro, covarianza)
- Statistica inferenziale = algoritmi applicati a uno o più set di dati per verificare se è probabile che le statistiche descrittive di quel set di dati si generalizzano in altri set di dati (F-value, t-value, ANOVA, regression)

La stat. inf. utilizza i dati per fare affermazioni che non abbiamo.

CONVENZIONE: Stat. descr. = "descriptive statistics"
Inf. stats = "statistics"

4.2 - Data distributions

300 persone cui si chiede: quante volte hai visionato , YouTube "Gangnam Style"?

- il video può essere visionato in parte
- anche lo $\emptyset$ è ammesso
- non hanno senso numeri $< \emptyset$

Data type Ratio

- Come visualizzare?

p. 103

(500,)

```
timesWatched = np.round(np.abs(np.random.randn(500)*20))/2
```

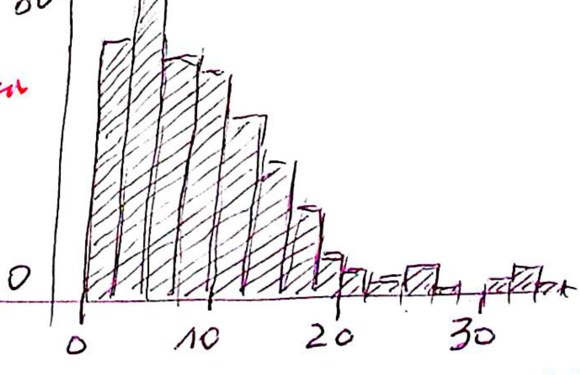
forzo un "outlier" non significativo

```
timesWatched[300] = 35
```

un grafico semplice è un istogramma

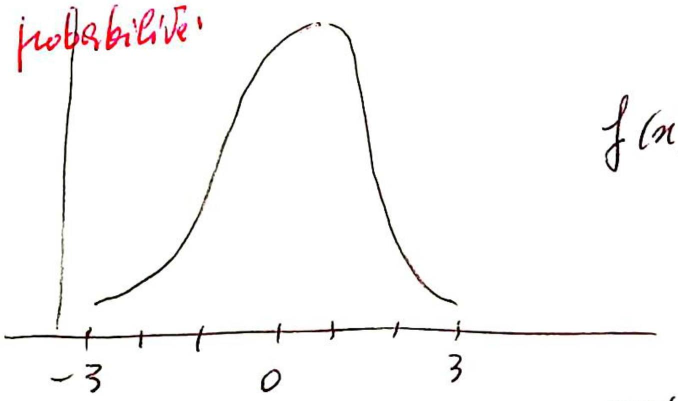
```
bins='fd'
```

Times watched



I dati possono provenire da misure, o da formule –

- ogni data sets ha la sua distribuzione
- se genero due sets di numeri random, ne faccio due istogrammi, si vede che non sono identici, ma molto simili –
- una distribuzione analitica non viene creata raffrontando i dati misurati, ma valutando una formula matematica –
- Se normalizzo → probabilità → es. Gaussiana



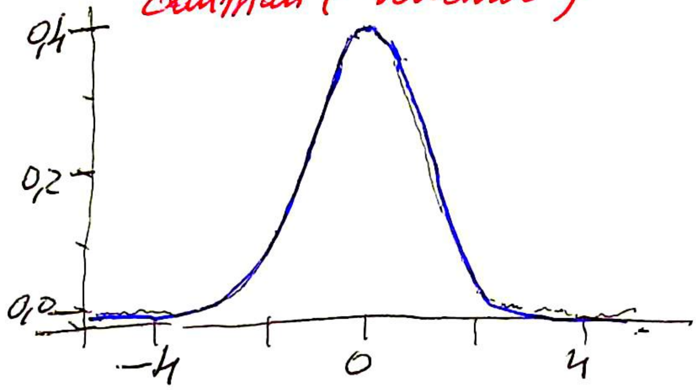
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/(2\sigma^2)} \qquad (4.1)$$

p. 107

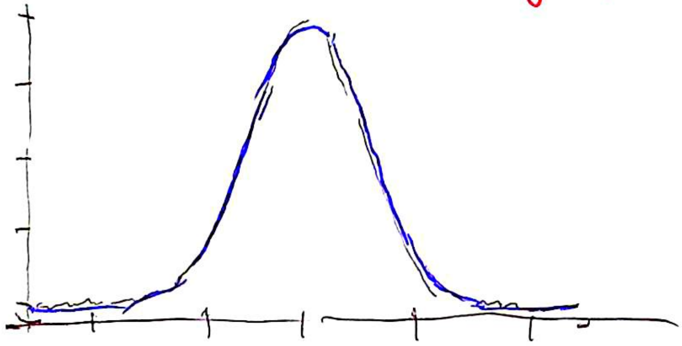
- ci sono miriadi di distribuzioni statistiche
  - danno una visione qualitativa delle caratteristiche dei dati
  - la maggior parte delle statistiche si basano su ipotesi relative alle distribuzioni sottostanti → è bene conoscere le distribuzioni
- inferenza statistica = ~~determinazione della probabilità che una data distribuzione statistica si possa riferire a statistiche descrittive di un campione~~ = interpretazione dei dati; che permette di trarre indicazioni su fenomeni non osservati direttamente — dalla osservazione di una parte dei dati "campione" – selezionato usualmente mediante un esperimento casuale (aleatorio), si deducono le caratteristiche della intera popolazione –
  - Esempi pratici in Biologia, Fisica, informatica – Es. AI tool
    molte volte da campionamenti casuali si determinano distribuzioni
- lo scarto tra le previsioni e i valori osservati "residuals" permette di valutare la qualità del modello
- legge dei grandi numeri, teorema del limite centrale – per comprendere i concetti statistici fondamentali –

Esempi di distribuzioni

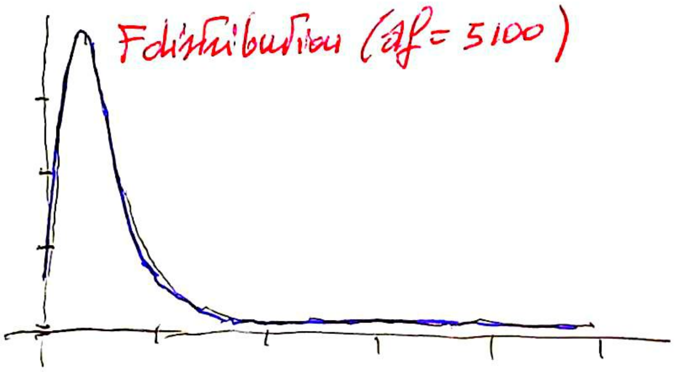
Gaussian ("bellcurve")



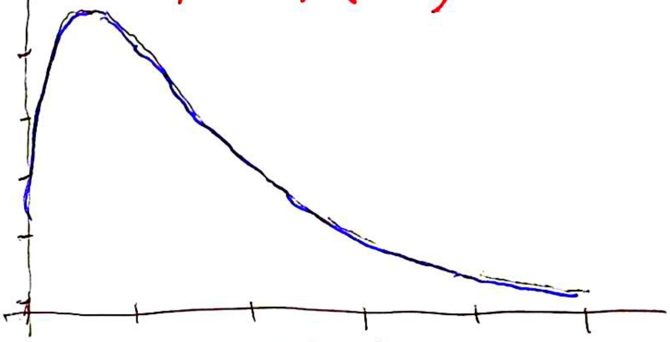
t distribution df=20



F distribution (df = 5,100)

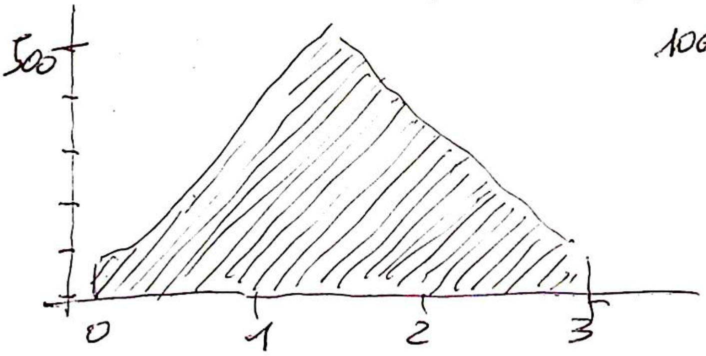


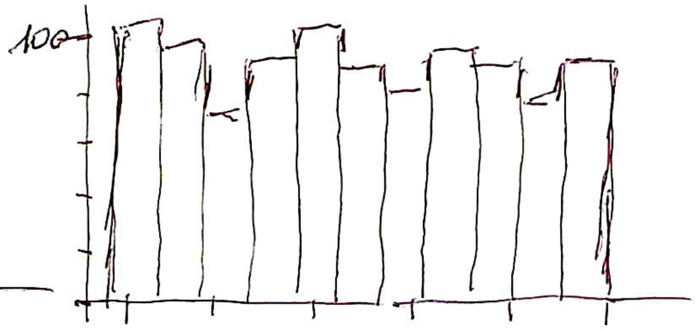
Chi-square (df=3)



Esempi di distribuzioni statistiche

```
x = np.linspace(-5, 5, 10001)
y = stats.norm.pdf(x)           Gaussiana
  = stats.t.pdf(x,20)           t-distribution
  = stats.f.pdf(x,5,100)        F-distribution
  = stats.chi2.pdf(x,3)         Chi-square distribution
```



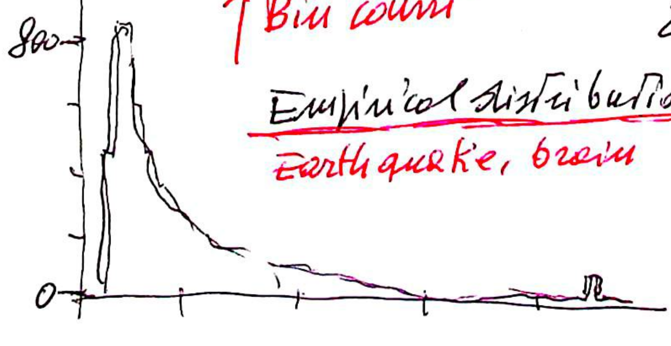


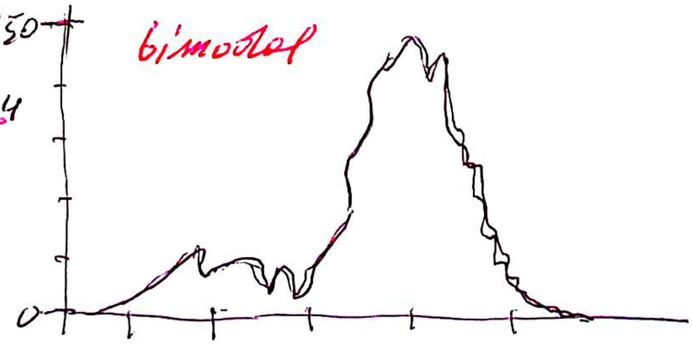
→ Data value

↑ Bin count

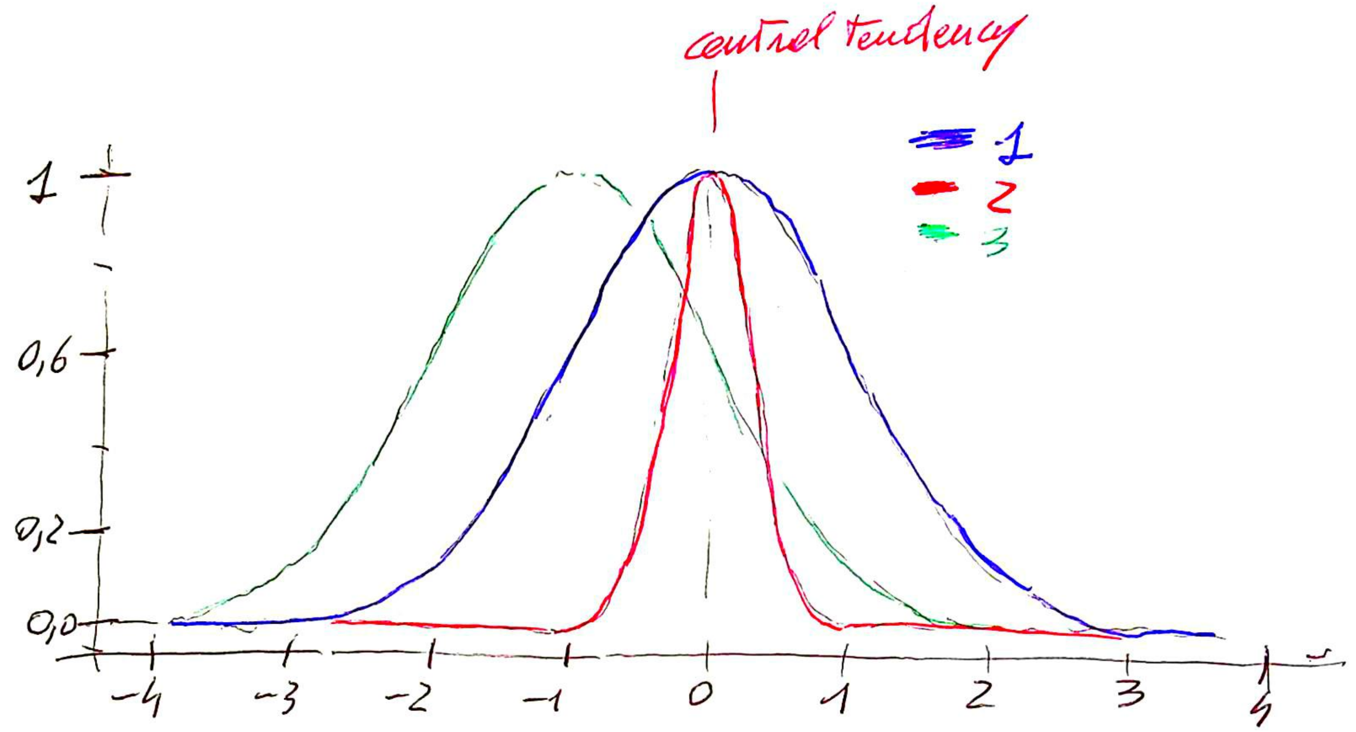
Empirical distribution

Earthquake, brain





%



Questi grafici danno un'idea qualitativa su "central tendency" e su "dispersion".

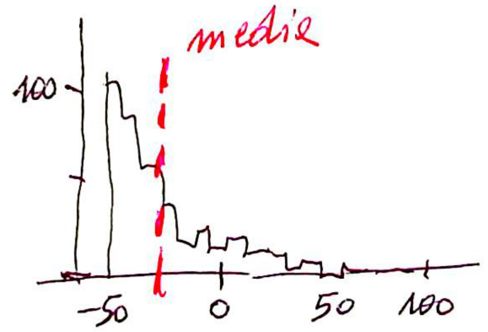
4.3 - Central tendency

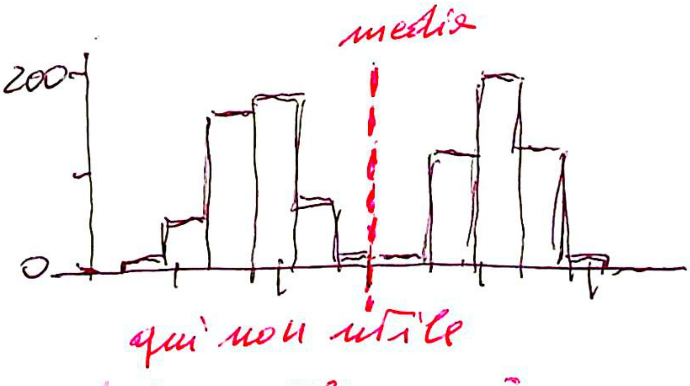
Ci sono molte modalità x quantificare "central tendency"

mean, median, mode

- <u>media</u> ——— $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$

  la possiamo visualizzare come un segmento verticale sovrapposto alla distribuzione. Si presta bene con distribuzioni circa simmetriche e unimodali





Media ≠ expected value - sensibile a outliers

Mediana divide in due metà ugualmente numerose i campioni

$$med(x) = \tilde{x}_i, \quad i = \frac{n+1}{2} \qquad (4,4)$$

sorted data
i = indice che corrisponde alla mediana

può essere un numero decimale
È facilmente interpretabile?
mediana ≠ valore atteso

Moda una misura della "central tendency"
L è il valore che compare più volte
la moda può essere calcolata se i dati sono discreti

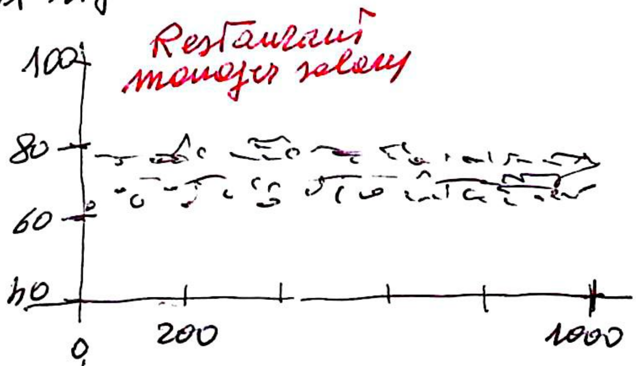
- ci può essere + di una moda

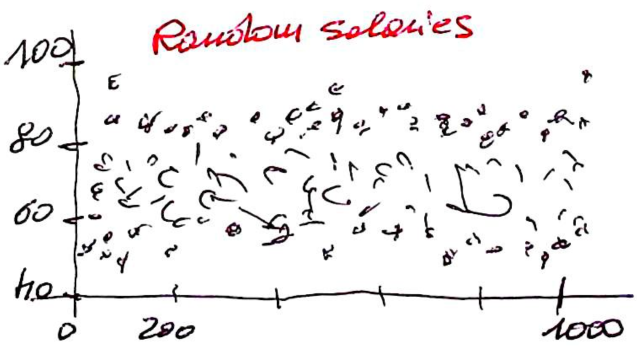
$$[\phi, \phi, 1, 1, 2, 7] \quad \begin{cases} moda\ 1 = \phi \\ moda\ 2 = 1 \end{cases}$$

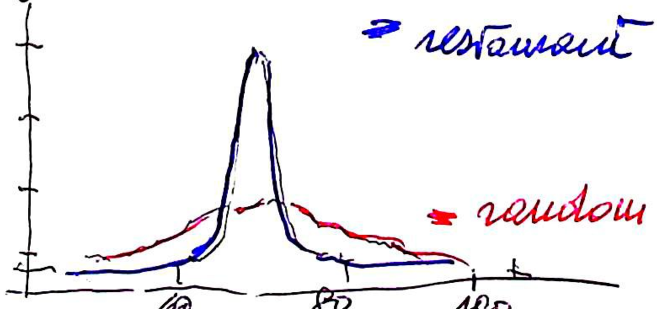
- in un istogramma la moda sta nella barra + alta

4,4 - Measures of dispersion

Si riferisce all'ampiezza della distribuzione







Istogramma

Varianza σ^2 – sample variance s^2

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (4,5)$$

MAD

$$MAD = \frac{1}{n-1} \sum_{i=1}^{n} |x_i - \bar{x}| \qquad (4,8)$$

absolute difference

1) s^2 enfatizza (-permette di individuare-) outliers
   L ~~x~~ usata x financial risk

2) s^2 e' continua e ha derivata morbida
   L ottimo x ottimizzazione

3) s^2 e' strettamente legata alla distanza Euclidea

4) s^2 e' momento di 2° ordine

5) s^2 e' legata a "least-squares-algorithm" for fitting regression model to data

a) MAD e' una buona misura della dispersione

b) MAD meno influenzata da outliers

c) MAD usata in machine-learning

d) MAD usata in "optimization methods"

$n-1$ = degrees of freedom

ha le stesse unità dei campioni

Standard deviation

$$STD = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (4,9)$$

Una comune normalizzazione:

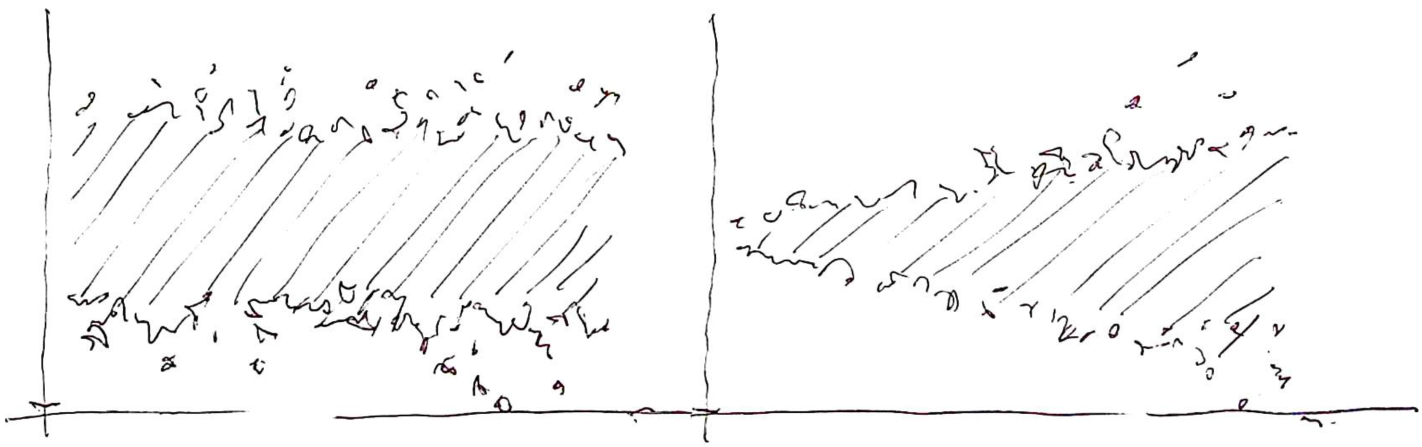
z-scoring → dati in unità di deviazione std

es. la regressione in queste unità è più facilmente interpretabile

Heteroscedasticity and Homoscedasticity

Homo— è una variabile che ha uguale varianza per tutti i valori

— in alcuni casi la varianza cresce secondo le x che aumentano



Homo—— Hetero——

rilevante x correlazione, regressione

Esempio di Hetero— : ricchezza e spese

se aumenta la ricchezza, la varietà degli acquisti aumenta

FWHM - Full width at half maximum

FM FWHM è una misura dell'ampiezza della Gaussiana
- L può essere calcolato analiticamente o empiricamente
- L ben interpretabile x distribuzioni ≃ gaussiane

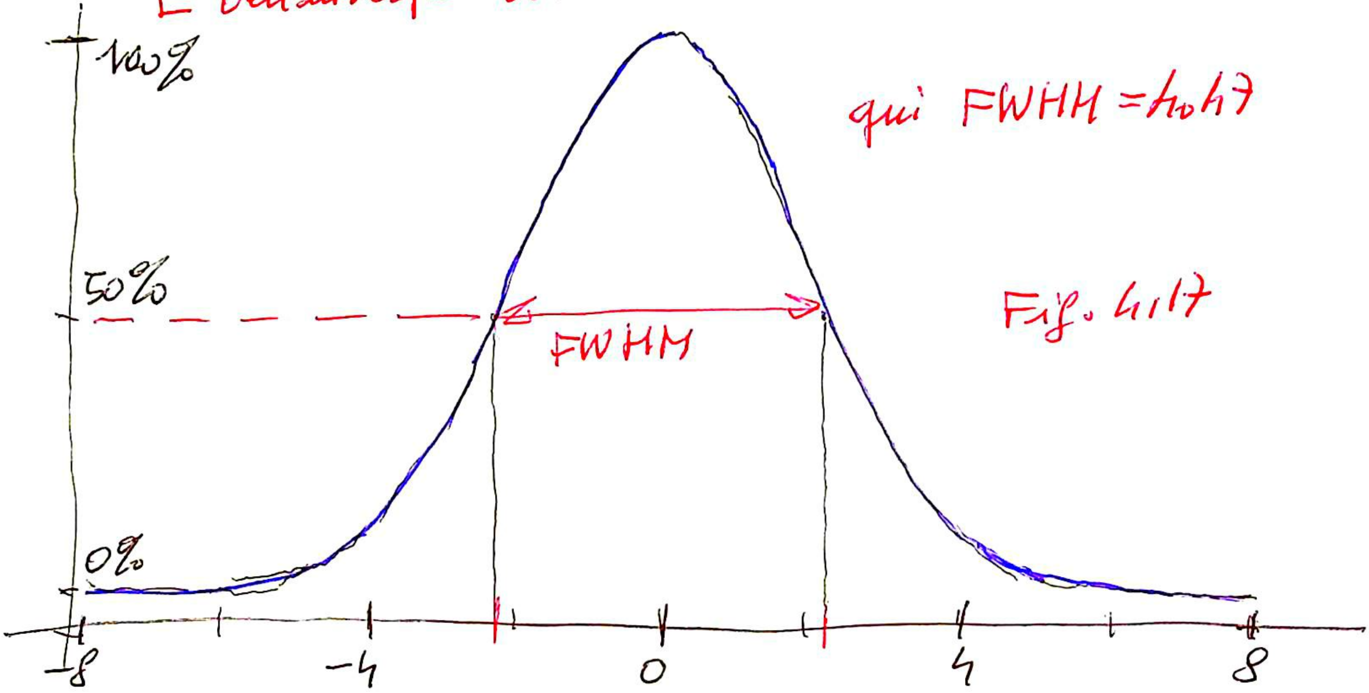


Fig. 4.17

nel caso della Gaussiana:

$$g(x) = \exp\left(\frac{-x^2}{2\sigma^2}\right) \qquad (4.10)$$

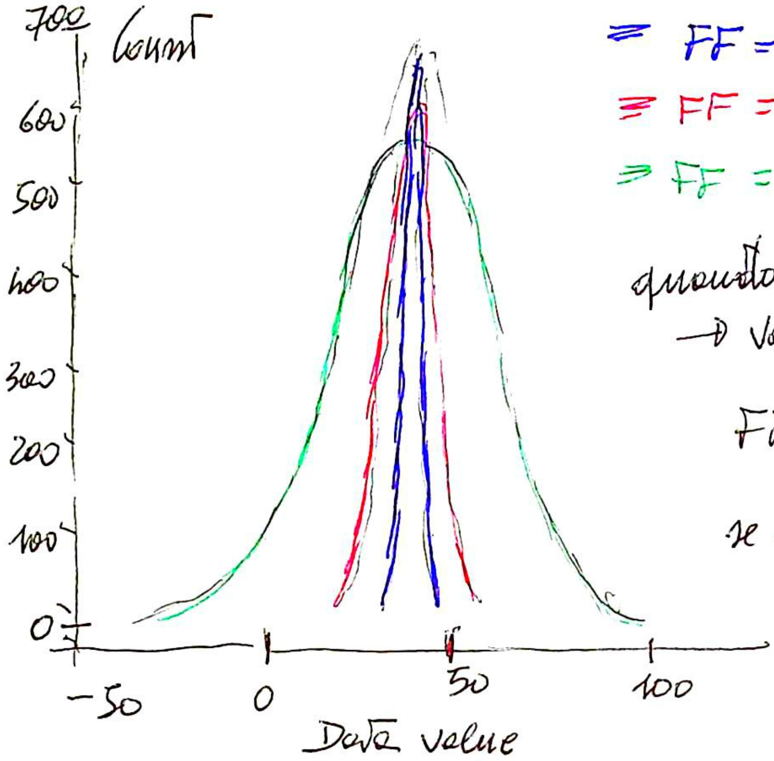
$$\mathrm{FWHM}(g(x)) = 2\sigma\sqrt{2\ln(2)} \qquad (4.11)$$

per una distribuzione empirica non è direttamente calcolabile - Si usano algoritmi per fare questo calcolo. p. 128
- L vedi Esercizio 10

Fano factor and CV   solo x datasets > ∅

$$FF = \frac{s^2}{\bar{x}}; \quad (4.12) \qquad CV = \frac{s}{\bar{x}} \quad (4.13)$$

vedi grafico seguente



= FF = 0,10

= FF = 1,02

= FF = 9,97

quando questo parametro è piccolo →
→ varianza piccola rispetto alla media

$FF = \frac{s^2}{\bar{x}}$; $CV = \frac{s}{\bar{x}}$;

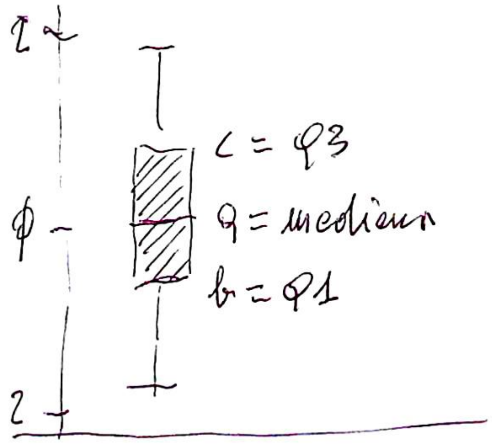
se aumentando → la varianza diventa maggiore della media

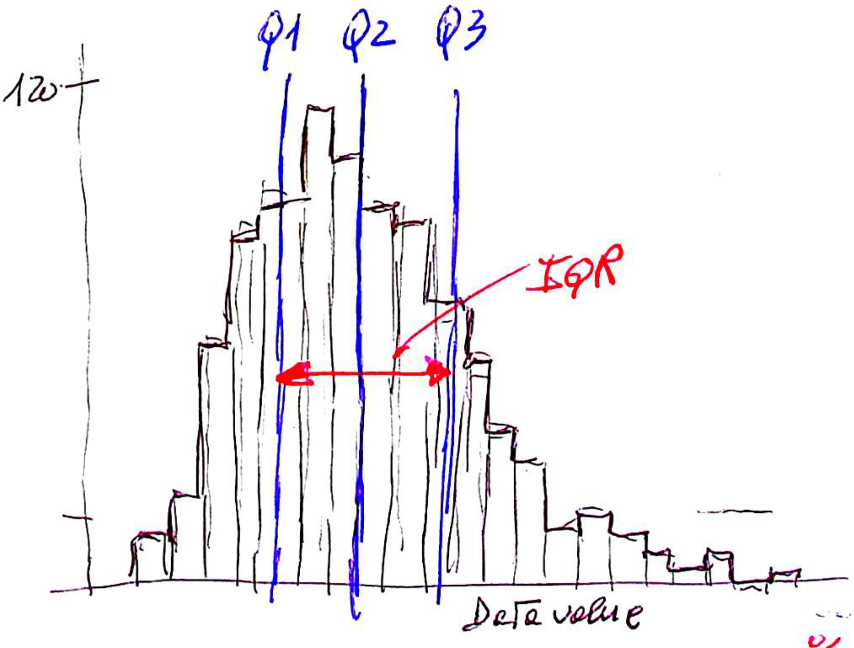
Fano è nelle unità dei campioni — CV è adimensionale

4,5 - Interquartile range (IQR)

IQR = distanza numerica tra 25% e 75% dei dati
↳ può essere calcolato con i seguenti passi:

1) calcolare la mediana "quartile 2" o "Q2"
2) ______ del subset < Q2 (a sinistra) "Q1"
3) ______ ______ > Q2 (a destra) "Q3"
4) calcolare Q3 - Q1 = IQR





questi confini sono anche detti "p25", "p75" dove p = percentile

Se IQR piccolo → distribuzione stretta

IQR è una misura non parametrica della variabilità (basata su mediana) → insensibile agli outliers

QQ plots

q = quantile

L es: IQR diviso in 4 quantili

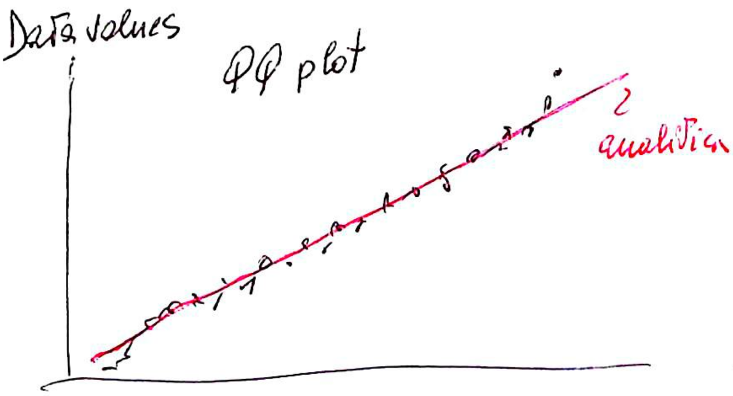
- QQ mostra la relazione tra i dati empirici e Gauss analitica
- QQ mostra se possibile ANOVA, se possibile regressione
- QQ aiuta a capire se ci sono problemi con i dataset

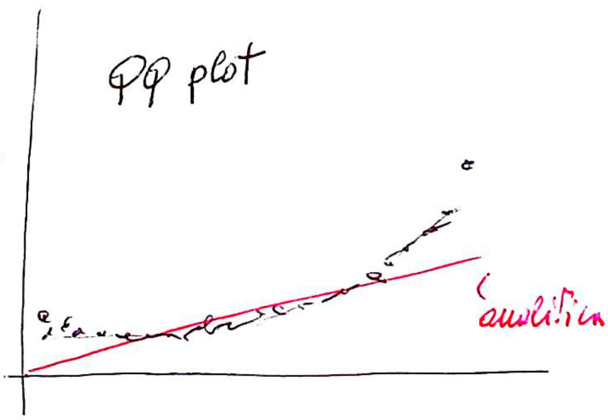
• abbiamo due datasets: provengono dal campionamento di una Gaussiana?

L posso fare istogramma sovrapponendo Gaussiana

L ma QQ ce lo dice meglio

- su x metto Gaussiana analitica
- su y — dati empirici





4.7 - Statistical "moments"

momenti = numeri che definiscono lo "shape"

∃ 1°, 2°, ecc.

Momento non standardizzato $m_k = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^k$ (4.14) il valore del momento non standardizzato dipende dalle unità di misura → standardizzo:

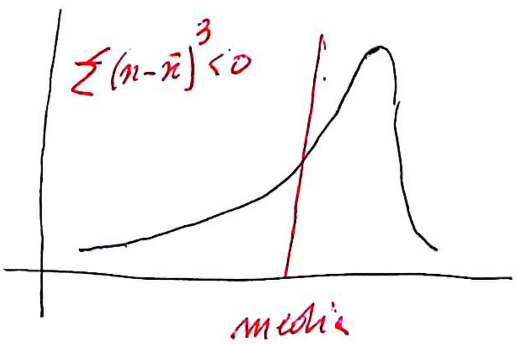
$$m_k = \frac{1}{N\sigma^k} \sum_{i=1}^{N} (x_i - \bar{x})^k \qquad (4,15)$$

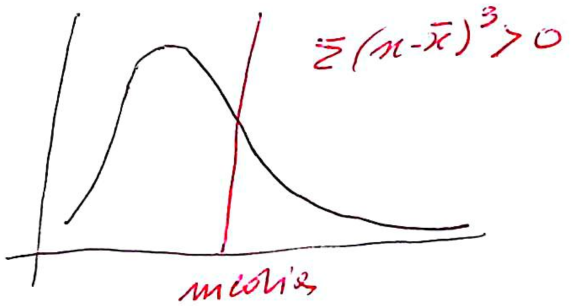
N° momento	Nome	Descrizione	Formula
Primo	Mean	Average	
Secondo	Variance	Dispersion	
terzo	Skew	Asymmetry	
quarto	Kurtosis	Tail fatness	

First moment = media = central tendency of the distribution = centro di massa dei dati

Second moment = varianza = dispersione attorno alla media

Third moment = skew – normalmente si calcola in modo standardizzato = asimmetria della varianza rispetto alla media



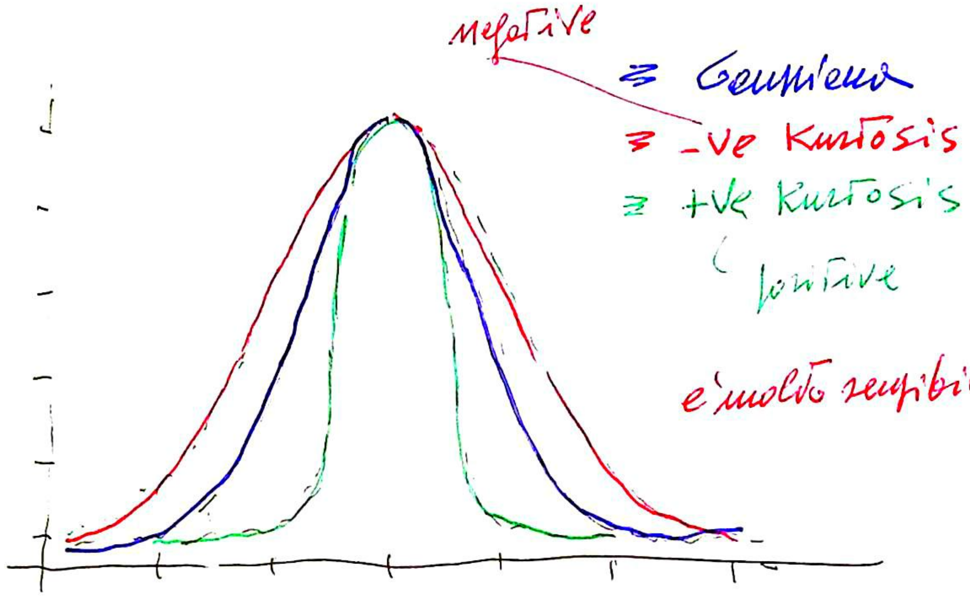


Fourth moment = Kurtosis stats.kurtosis(x)

$$\text{Kurtosis} = \frac{\mu_4}{\sigma^4}$$ – momento di 4° ordine

Gaussiana Kurtosis = 3

L = allungamento di una distribuzione attorno al suo valore modale



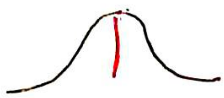
/ potenza 4

è molto sensibile a outliers

usata x stimare il rischio nel mondo finanziario

— isolare il rumore, mediante una tecnica denominata

— "independent components analysis"

media  ; varianza  ;

skew  ; kurtosis 

4.8 – Istograms part 2: Number of bins

p. 139

Relazione tra "bin count" (variabile K) e larghezza del bin (w) –

$$K = \left\lceil \frac{\max(x) - \min(x)}{w} \right\rceil \qquad (4.17)$$

rounded up

ci sono vari metodi x determinare n° bins

Method	Formula	Advantage
Arbitrario	$K = 40$	simple
Sturges	$K = \lceil \log_2(N) \rceil + 1$	depends on count
Friedman-Diaconis	$W = 2 \cdot (IQR / \sqrt[3]{N})$	" on count & spread

= FD = F-D

Inter-quartile range

plt.hist(data, bins='fd')

si può avere anche una W variabile, ma è di difficile interpretazione.

Other descriptive stats

∃ molte altre distribuzioni (es. Hurst exponent).

Le serie temporali hanno "spettro" e "autocorrelazione"

I set di dati multivariati → covarianza

Le matrici → rango, "condition number", "singular value spectrum"...

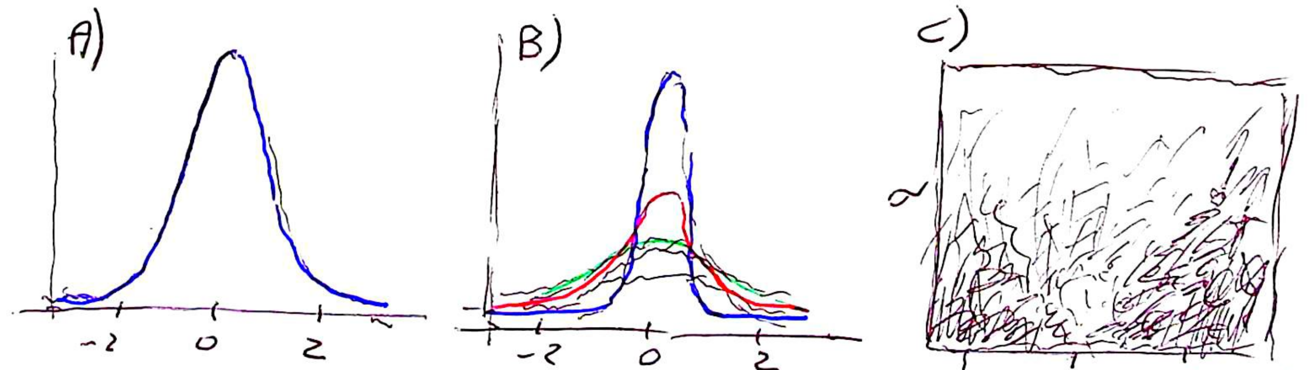
— in ogni caso — anche a queste — si applicano i concetti esposti in questo

4.9 - Esercizi

La Gaussiana (a.k.a. distribuzione normale, curva a campana) è la più importante.

- crea una famiglia di gaussiane in una matrice (differiscono x σ), es. 50 gaussiane con $.1 \le \sigma \le 3$.
- crea una immagine come in fig. 4,29C

Fig. 4,29




```
plt.imshow(G, extent=[x[0], x[-1], sigmas[0], sigmas[-1]],
        cmap='grey', aspect='auto', origin='lower', vmin=0, vmax=.8)
N=50
x = np.linspace(-3, 3, 111)
sigmas = np.linspace(.1, 3, N)
G = np.zeros((N, len(x)))
for i in range(N):
    a = 1/(sigmas[i]*np.sqrt(2*np.pi))
    eTerm = x**2/(2*sigmas[i]**2)
    G[i,:] = np.exp(eTerm)
```

L'integrale delle gaussiane (qui sono norm.) = 1, noi facciamo integrale solo in un sottoinsieme di x → integrale < 1

– per un risultato migliore → estendere x

– se rimuoviamo il fattore moltiplicativo → il picco = 1, non più l'integrale

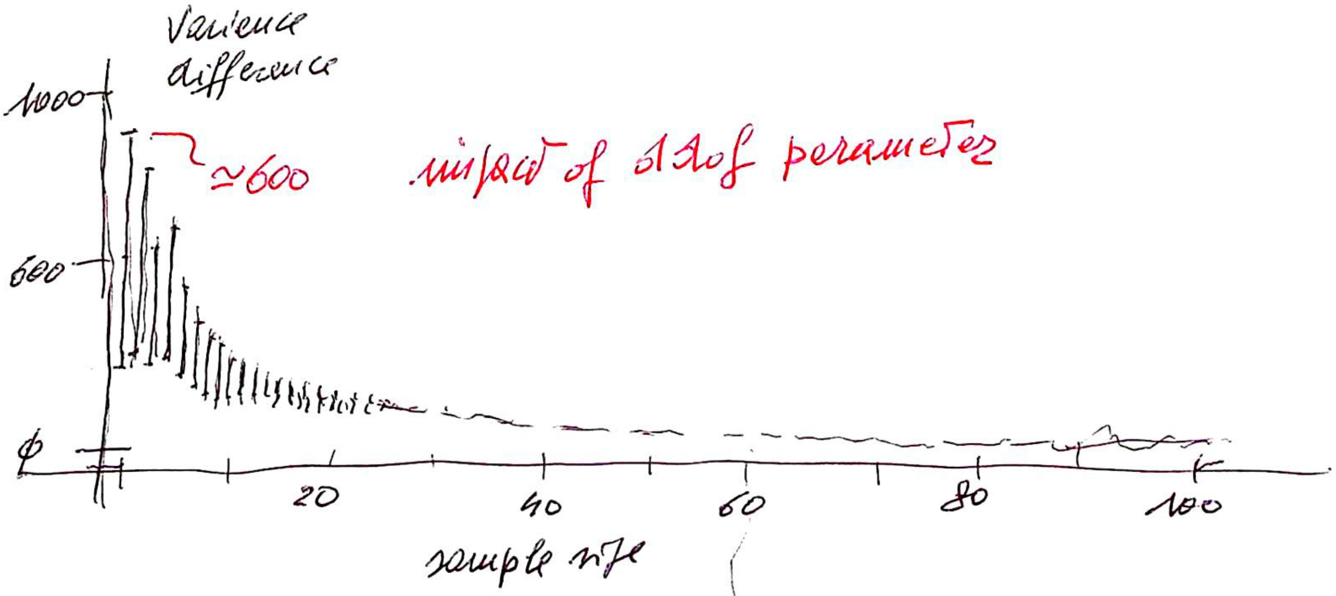
Scriviamo la media senza usare "numpy" – calcolare 3 statistiche

"media", "mediana", "varianza"

Perché dividere per "N-1"? che differenza fa?

– numeri casuali tra -100 e 100, calcoliamo varianza due volte (una con ddof = 1, l'altra con ddof = 1 – ne calcoliamo la differenza in valore assoluto – ripetiamo al variare del n° di campioni –

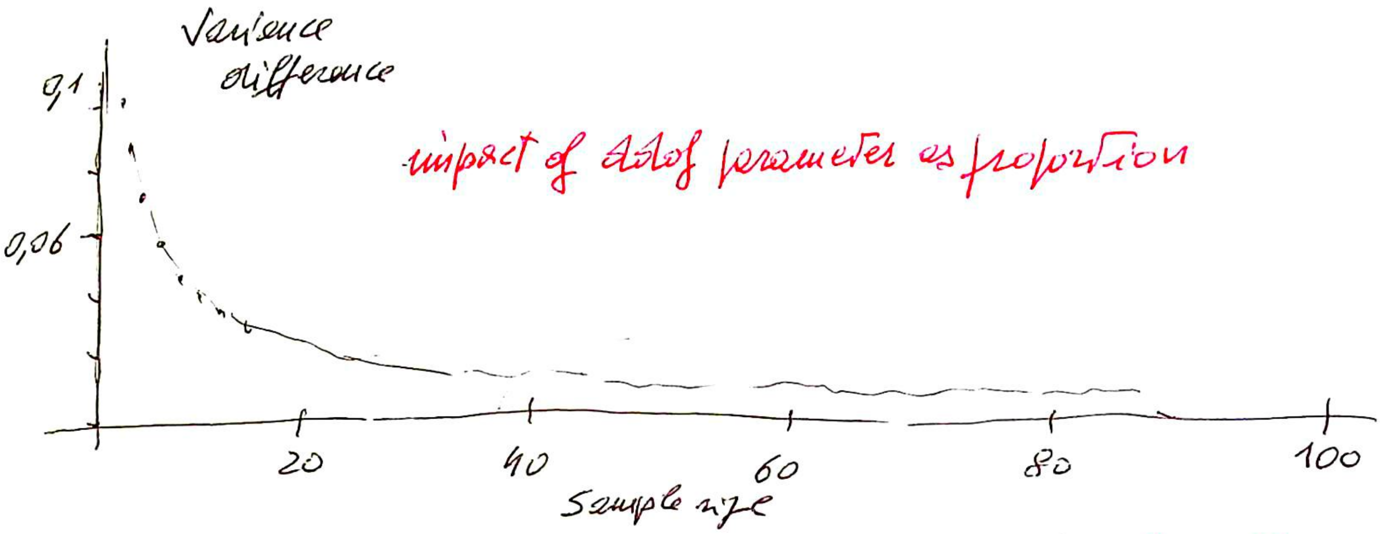
– quando numeri casuali → possiamo ripetere esperimento e fare la media – vediamo una barra errore – quando 25 runs



600 è un errore molto grande? → provo con valori -10 ÷ 10 e noto che questo valore ↓ → difficile valutare

↳ questo è funzione del valore dei dati → normalizziamo e ricalcoliamo.

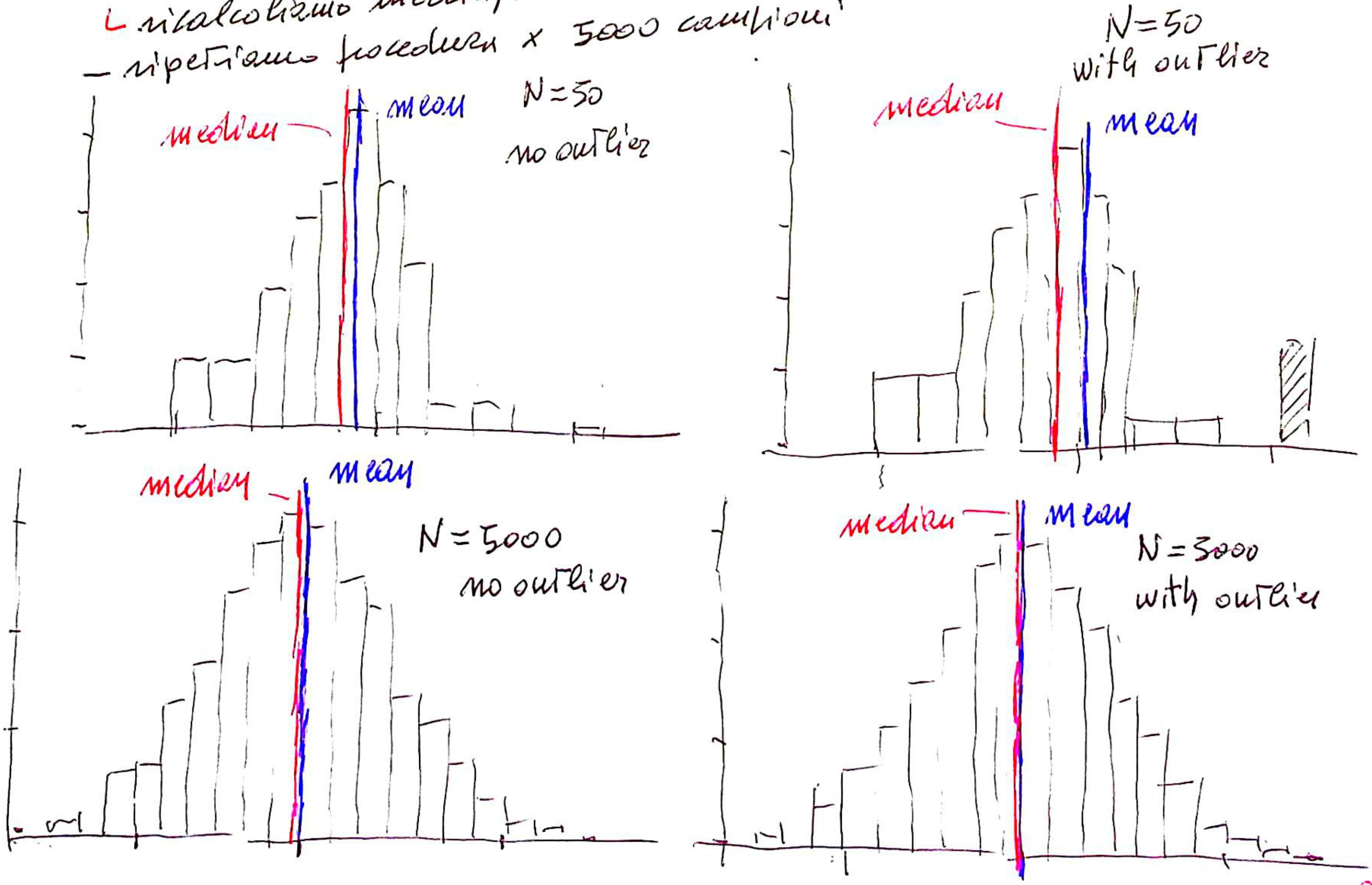
Calcoliamo nuovo valore di scala $\frac{V_1 - V_0}{V_1 + V_0}$ varianza con ddof = 1 con normalizzazione:



il miglioramento è dato da un fattore $1/(2N-1)$ → non dipende dal valore dei dati

Esploriamo l'impatto di un singolo outlayer sulla media, rispetto alla mediana (x campioni tenti e pochi) mediana

- creiamo 50 numeri random da Gaussiana, calcoliamo media,
  - visualizziamo su istogramma.
- creiamo un set anomalo sostituendo il valore maggiore × 4
  - ricalcoliamo media, mediana
- ripetiamo procedura × 5000 campioni



N=50 mean increased of 0.53 | N=5000 ——— 0,02

median ——— 0,00 | ——— 0,00

Calcolare i momenti statistici

mean, variance, skew, kurtosis = stats.norm.stats(loc=1, scale=2, moments='mvsk')

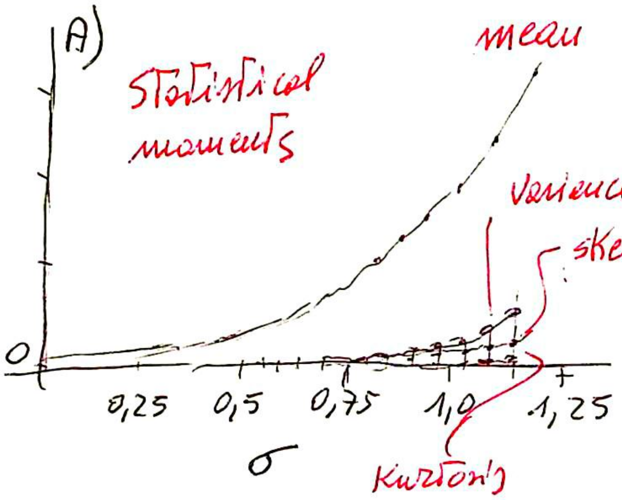
Average = 1
Variance = 4
Skewness = 0
Kurtosis = 0

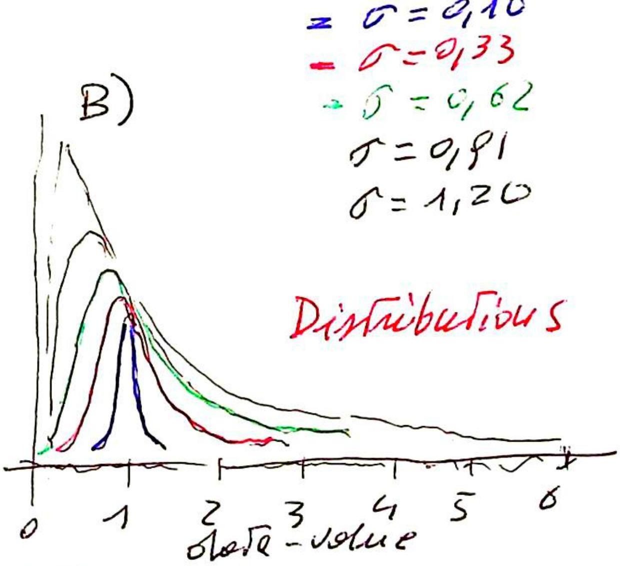
uniform
lognorm
expnorm → Vedi Python x i diversi parametri

Come variano i momenti cambiando la distribuzione?

scipy.stats.skew()

— N = 13'524 casuali, gaussiana → trasformare in 20 log-normal, dovranno definito come $\exp(X\sigma)$ con $.1 \leq \sigma < 20$ in 20 steps





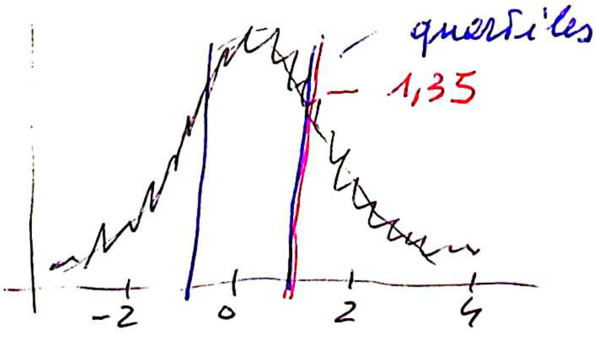
IQR simile a deviazione std — attorno a media

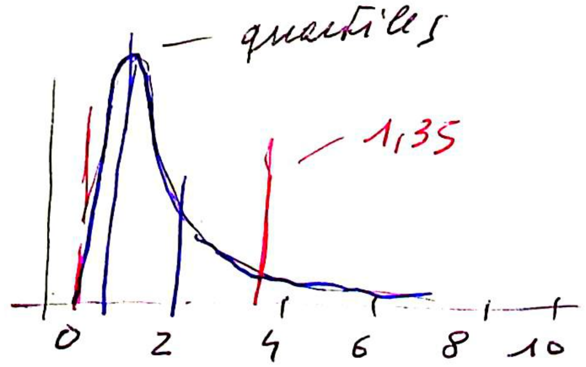
attorno a mediana

x Gauss: IQR = $1{,}35\sigma$

N=10'000 — verificare // cambiare e^X, quanto vale IQR?

└ fare da PC: IQR = 1,360 ; IQR = 1,481 x e^X





Calcolare FWHM empirico x Gauss

- definisco la funzione emp FWHM(x,y)
- normalizzo i dati $\tilde{y} = (y - y\min(y))/(\max(y) - \min(y))$
- trovo il picco come indice

L essendo normalizzata devo cercare attorno a 0,5 — pre-peak — post-peak

L

- calcolo FWHM — distanza tra pre e post = 4,46
- calcolare FWHM x Gauss formula $\sigma = 1,9$, $-8 < x < +8$, $N = 1001$

L FWHM = 4,47

- variamo $.1 \leq \sigma \leq 5$

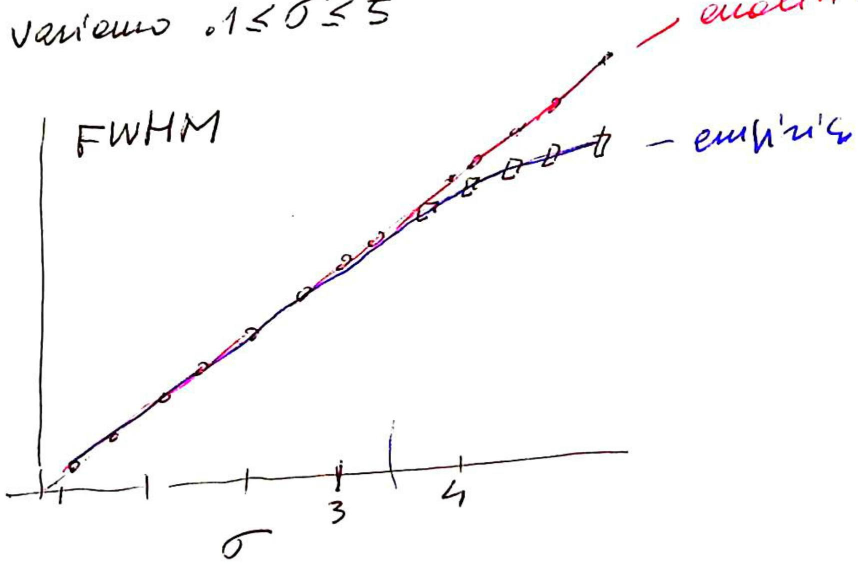
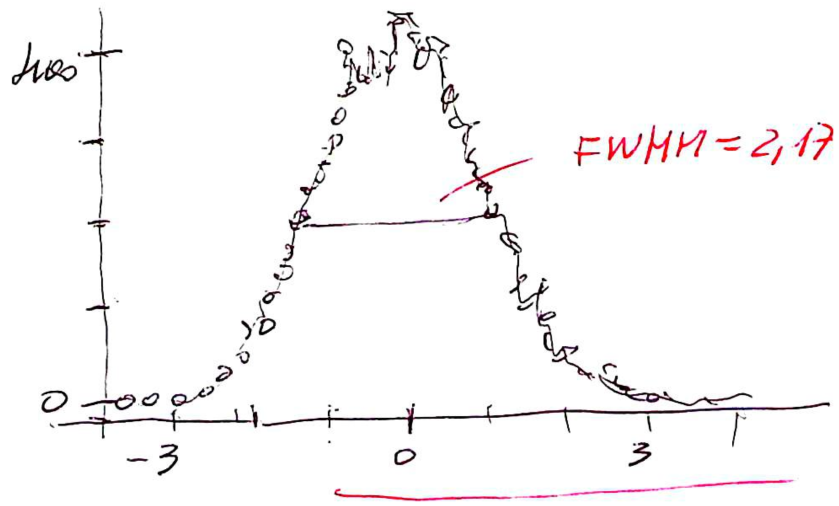


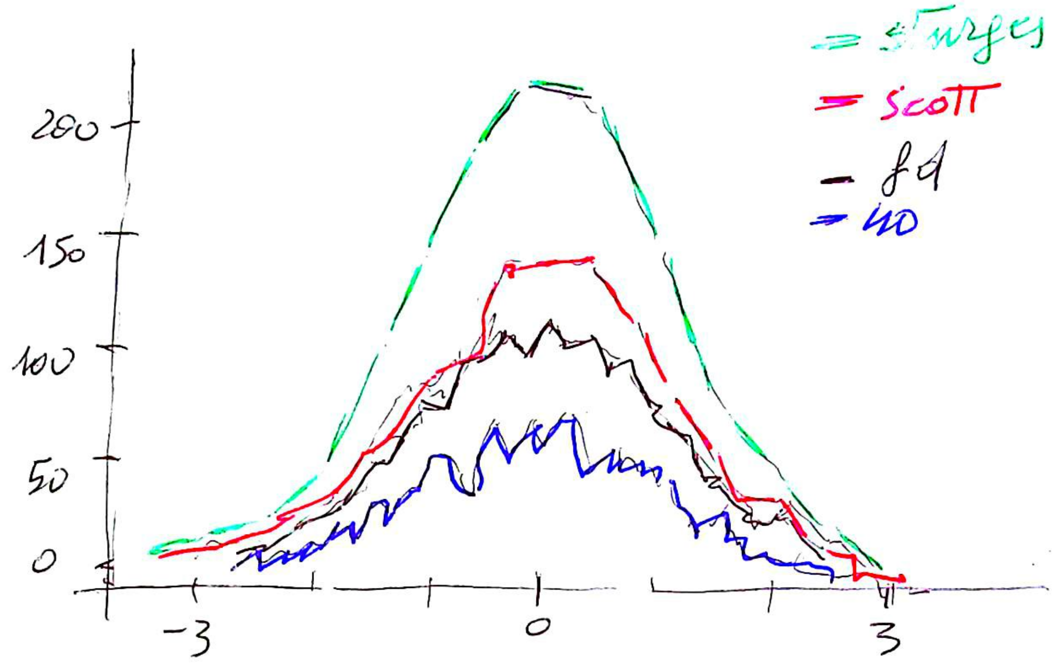
Fig. 4,35

N = 12345 Gaussian, bins = 100, disegnare FWHM



Influenza delle varie regole di calcolo di bin su istogramma

N=1000 Gauss



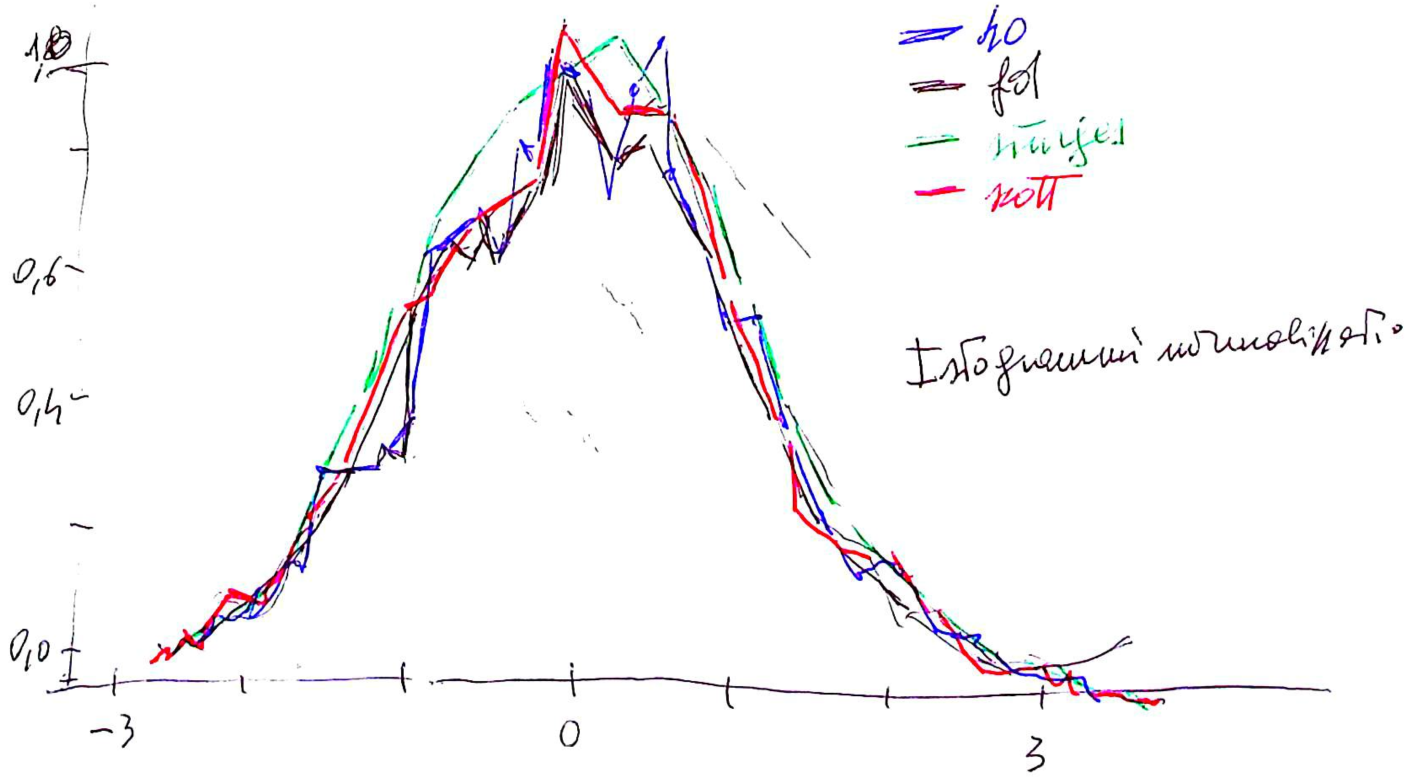
i valori sono alquanto diversi, ma gli andamenti sono simili

→ non influenzano qualitativamente sui risultati! p. 156

variando il n° dei bins, varia la quantità di dati inclusi

↳ altezze differenti

↳ si può normalizzare plottando y/np.max(y)



— Fine Esercizio 4 —

5 - Simulating data

5.1 - Why simulate data?

- Validare i metodi di analisi
  - ∃ una gamma di metodologie, ciascuna con più parametri che possono influenzare il risultato. Simulando si possono confrontare metodologie diverse. Questo è particolarmente utile quando i dati simulati hanno caratteristiche simili a quelli reali
- Capire vantaggi e limiti dei vari metodi. Consente di gestire l'influenza del rumore. Cosa non fattibile con i dati reali.
- Consente di capire in modo più approfondito come funzionano i metodi dei segnali
- Capire meglio i dati di cui si dispone
- Pensare in modo + critico e attento ai dati. Come selezionare il metodo di analisi.
- Usare un approccio "pratico"
- Statistica computazionale (o empirica). Ricavare significatività statistica, intervalli di confidenza. Questo diventa necessario se i dati violano le assunzioni richieste dalla statistica parametrica
- Molte diverse simulazioni richiedono pensiero critico
- ______________ migliorano le capacità di programmazione e codifica.
- Consente molte prove senza necessità di esperimenti

5.2 - Random data from distributions

– meglio partire da istogramma

5.2.1 - Normally distributed random data

$$X \sim \mathcal{N}(\mu, \sigma^2) \qquad (5.1)$$

notare tilde

$\mathcal{N}(\phi, 1)$ $\mu = \phi;\ \sigma^2 = 1$ = normal std distribution

- σ^2 allarga o restringe la distribuzione
- μ = media

5.2.2 - Uniformly distributed data

$$X \sim \mathcal{U}(a, b) \qquad (5.2)$$

$\mathcal{U}(\phi, 1)$ = std uniform distribution

se scrivo $\quad Y = 2\pi X - \pi \qquad (5.4)$

└ quale sarà il suo intervallo? $[-\pi, \pi]$

$$\forall a, b \text{ con } a < b \Rightarrow Y = a + (b - a)U \qquad (5.5)$$

dove $\quad U \sim \mathcal{U}(\phi, 1) \qquad (5.6)$

in questa distribuzione

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(a-b)^2}{12} \qquad \begin{matrix}(5.7)\\(5.8)\end{matrix}$$

vale

$$Y = \mu + \sqrt{3}\,\sigma\,(2U - 1) \qquad (5.9)$$

np.random.uniform (a, b, size = N)

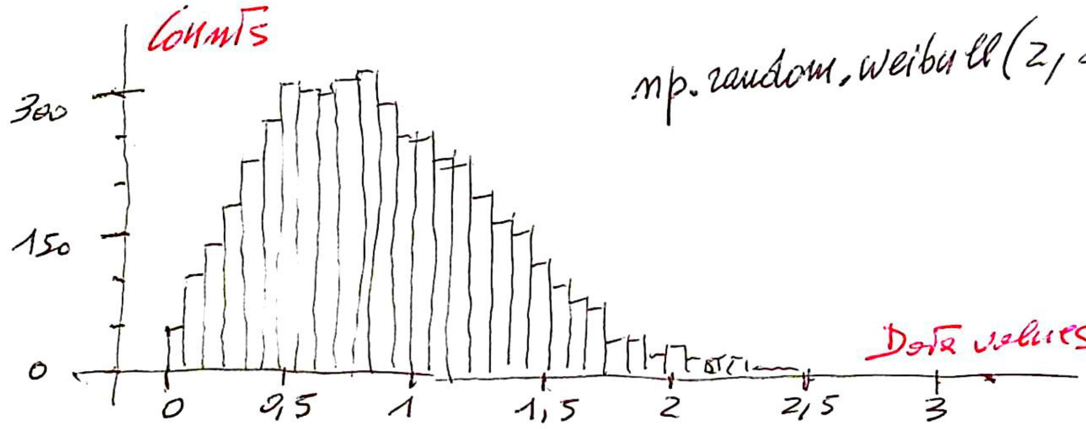
5,2,3 - Random data from other distributions

- dati casuali da una distribuzione di Weibull

np.random.weibull in formule

$$Y = \exp(X\sigma + \mu) \qquad (5,11)$$

np.random.weibull(2, 2500)



5,2,4 - Random integers

np.random.randint(1, 5, size=10000) esempio

np.random.lognormal(1, 0,5, 5000)

5,3 - Random elements of a set

es. (1, 2, 3, 6, 7, 8)

S = [1, 2, np.pi, 10]

np.choiche(S, 1) = 2

t = ["a", "b", "hello"]

np.random.choice(t, 1)

es. np.random.choice(S, 4)

↳ array([10, 1, 1, 3.14])

A) Sampling without replacement

np.random.choice(S, 4)

B) Sampling with replacement

(S, 4, replace=False)

NOTA con replacement = creiamo un nuovo dataset più grande, perchè gli elementi di partenza possono essere selezionati più volte
↳ la statistica del nuovo insieme sarà diversa

senza replacement = produco un dataset identico, tranne l'ordine

5.4 - Random permutations

↳ è un modo di distribuire casualmente gli elementi.

```
l = np.arange(5)
print(l)
print(np.random.permutation(l))
```

⟹ [0 1 2 3 4]
[2 0 1 3 4]

posso ordinare (sort) gli elementi

```
theData = np.arange(-3,4)**3
newIdx = np.random.permutation(len(theData))
shufData = theData[newIdx]
```

⟹

[-27 -8 -1 0 1 8 27] theData
[3 4 1 6 2 0 5] newIdx
[0 1 -8 27 -1 -27 8] shufData

- randomizzazione di dati accoppiati es. altezze e pesi
  - rimescolo altezze, ma non pesi ⟶ non più correlate
- calcolare la correlazione tra due set di dati ci può dire quanto siano indipendenti

5.5 - Reproducing randomness

- genero due volte dati casuali : mi serve che siano uguali? sì quando sto provando, in modo da avere risultati confrontabili

np.random.randn(3,3) — qui le matrici sempre differenti ma se scrivo

rs = np.random.RandomState(17)

rs.randn(3,3) — qui ottengo lo stesso risultato

perché Python ha settato il valore di "seed"

usare con cautela

numeri casuali ci danno informazioni sulla "robustezza" di alcune ipotesi o algoritmi.

5.6 - Running experiments with random numbers

- genero i valori di una popolazione casuale, Gauss
- faccio la media
- mi pongo la domanda: in che modo σ^2 influisce sulla correlazione con un'altra variabile?
- decidiamo come varia σ^2 e il grado di correlazione

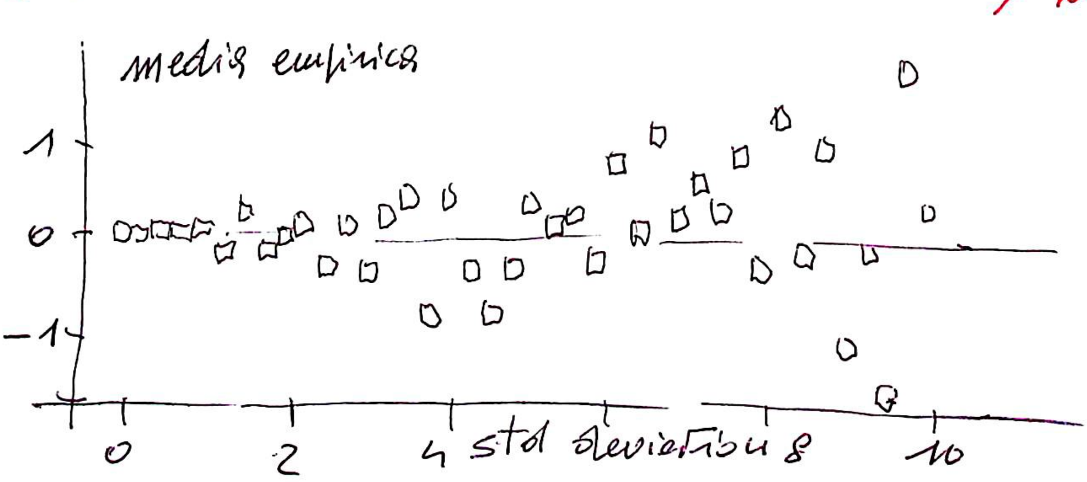
es. ANOVA sono robusti rispetto al variare di σ^2

es. impatto di variazione σ^2 sulla media x dati Gauss
- in teoria non ha nessun impatto

In concreto: $N=100$

in un ciclo for faccio variare σ, calcolo i campioni Gauss, ne faccio la media results

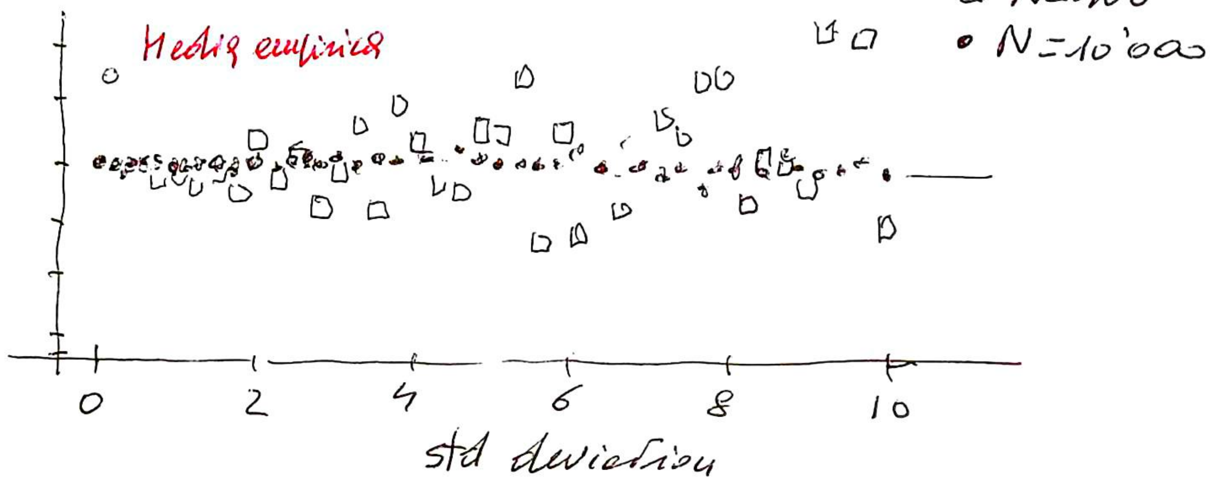
grafico σ vs results ; grafico media ; perché / radq campioni



questa variabilità empirica → statistiche inferenziali

L poiché le variazioni della media sono simmetriche (circa) rispetto allo zero → la deviazione non è sistematica - questo effetto apparirà in una distribuzione log-norm-

Altro esercizio: impatto di N sulla media



5.7 - The emerging world of data-simulations

Trellico

- serie temporali, dati finanziari, meteo, processi bio, fisica,
- "Generative deep learning" = ritratti inventati, video fasulli

L esamineremo poco di questo, ma questi sono fondati sugli argomenti del libro

5.8 - Finding publicly available real datasets

- sono sparsi come tipologia, sono specifici
- alcuni archivi sono free, alcuni free + registrazione

L su questo ∄ standardizzazione

- due siti popolari specifici
  - UCI machine learning repository[7] // Kaggle[8]

https://archive.ics.uci.edu/ml/datasets.php

https://www.kaggle.com/datasets

Esercizio 1 – relazione tra N e accuratezza della media empirica e sulla varianza – $N = 10$ $\mathcal{N}(0,2)$

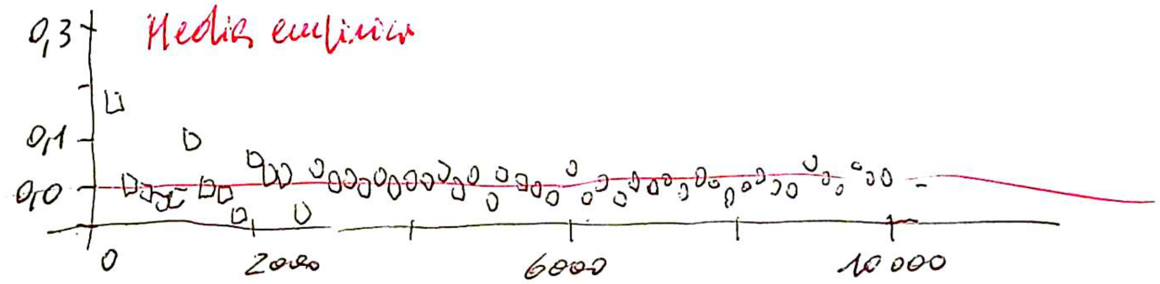
media empirica $= -0,184$
varianza " $= 1,688$

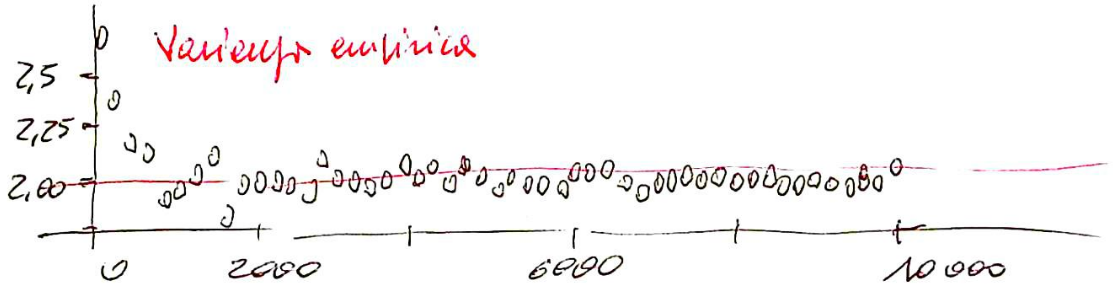
| in teoria 0
2

in un loop for $N = 10 \div 10'010$ step 200

L in ogni loop nuovo dataset (stessa media, varianza) memorizzo

faccio grafico





– le discrepanze sono sia + che –
– si intravede la legge dei grandi numeri p. 184
– ______ Teorema del limite centrale

grandi numeri: se $N \to \infty \Rightarrow$ media empirica $\to \mu$

limite centrale: $\sum_{i=1}^{n} x_i$ ⟶ è una distribuzione Gaussiana
stessa distribuzione arbitraria

L es.
ripeto la più volte su campioni diversi presi da una popolazione di distribuzione arbitraria – ripetendo molte volte $\sum x_i \Rightarrow$ Gauss

in formule:

dato X_j (famiglia di n) aleatorie indipendenti, identicamente distribuite $\Rightarrow$

n; X_j; siano $E[X_j]=\mu$, $Var[X_j]=\sigma^2$

LIMITE CENTRALE

con $j=1,\dots,n$; $0<\sigma^2<+\infty$ poniamo

$$Y_n = \frac{\frac{1}{n}\sum_{j=1}^{n} X_j - \mu}{\sigma/\sqrt{n}} ; \Rightarrow Y_n \xrightarrow{D} Y \sim N(0,1)$$

con altra descrizione

dati n X_i con media $=\mu_x$, varianza $=\sigma_x^2$ $\Rightarrow$ consideriamo

la variabile $\underbrace{(\bar{X}-\mu_x)/\sigma_{\bar{x}}}$ dove $\sigma_{\bar{x}}^2 = \sigma_x^2/n$

dove $\bar{X} = \frac{\sum X_i}{n}$

se $n \to \infty$ $= \bigcirc \sim N(0,1)$

Esercizio 2 confermiamo empiricamente

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(a-b)^2}{12}$$

Y = np.random.uniform (a, b, size = 1324)

meanDiff = np.mean(Y) $-$ $\frac{(a+b)}{2}$ $\longrightarrow$ $-0{,}040$

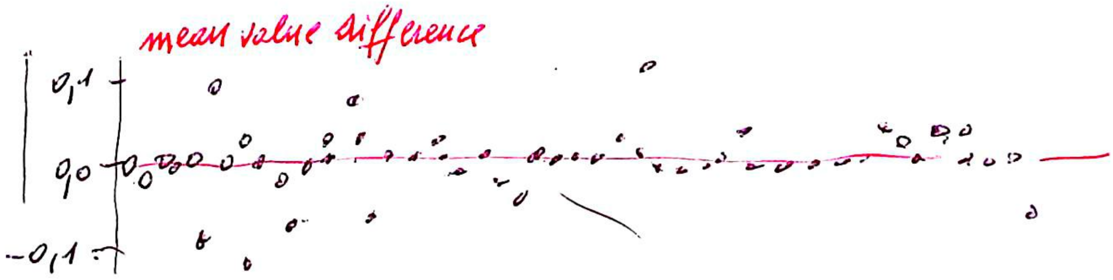
varDiff = (np.var (Y, ddof=1) $-$ (b-a)**2/12)**2 $\longrightarrow$ $0{,}000$

ora invece Ns = np.arange (10, 10200, step=200)

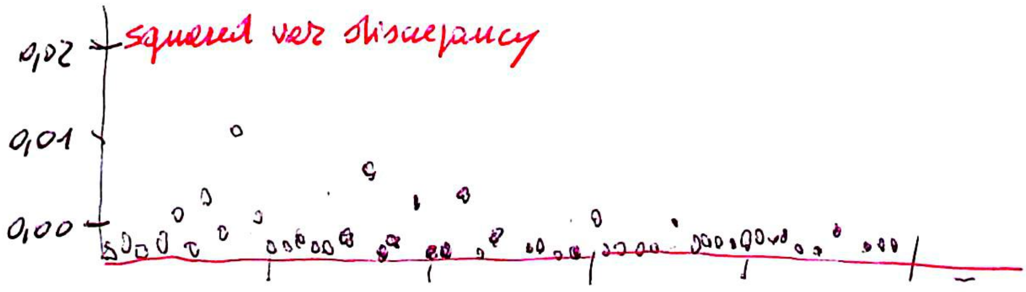
in un ciclo "for" rendiamo i, N in Ns

genero ad ogni passo una serie di dati, ne calcolo media e varianza disegno la discrepanza

A



B



il pannello B mostra valori solo $> \emptyset$ → perché σ^2 ⟹
non si hanno bias sistematici
- le regressioni e le ANOVA si basano su discrepanze quadrate.

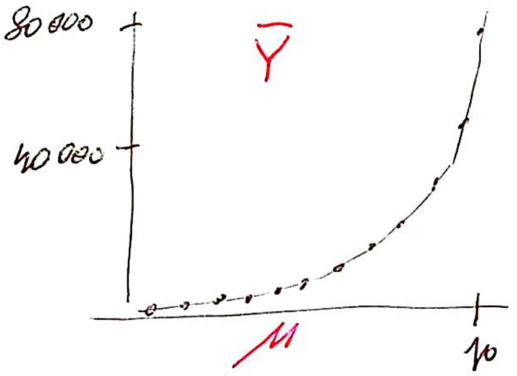
Esercizio 3

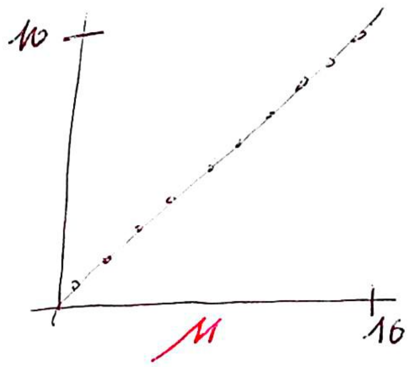
abbiamo visto come da dati non gaussiani si possa passare a una Gaussiana. partiamo da $Y = \exp(X\sigma + \mu)$ con $\mu = 2$, $\sigma = 1,5$
- calcoliamo la media

media attesa $\bar{Y} = e^{\mu + \sigma^2/2}$ (5,13)

- confermiamo numericamente che $\mu \simeq 2$
- ciclo "for" - ricalcoliamo μ ad ogni passo (13 valori di μ tra 1 e 10)
- compariamo μ e $\bar{Y}$

$\ln(\bar{Y}) - \frac{1}{2}\sigma^2$





Esercizio 4 - relazione tra media e deviazione std di una distribuzione uniforme nell'intervallo (a, b)

$Y =$ np.random.uniform $(a, b, \text{size} = N)$, $N = 1001$

calcolo le medie con due formule empiriche → 5,0161 / 4,9839

calcolo le medie formali → 4,9895 / 5,0000 media delle medie

calcolo la deviazione std con due formule

→ $\sigma\_a = 0,5713$
$\sigma\_b = 0,5713$
std(Y) $= 0,5867$

$Y = \mu + \sqrt{3}\,\sigma(2U - 1)$
$U \sim \mathcal{U}(\emptyset, 1)$

creiamo un dataset con data media e deviazione:

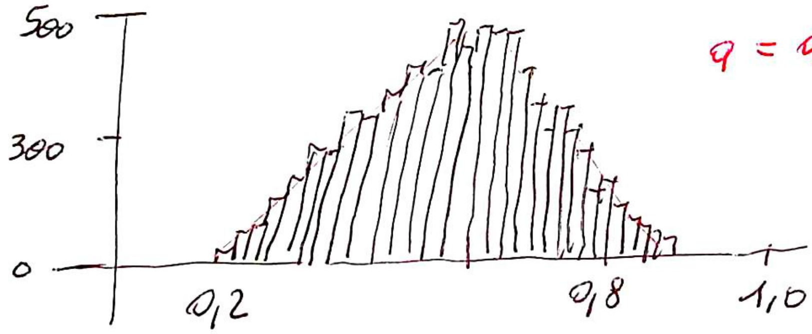
$\mu = 3,5$, $\sigma = 1,8$, $N = 100000$, calcoliamo media deviazione teorica → con mean $= 3,5$, std $= 1,798$

Esercizio 5 – quale mediana e moda x distrib. uniforme (a, b) ?

Esercizio 6 – distribuzione triangolare $[a, b] \subset \mathbb{R}$, $c \in [a, b]$

$N = 10\,000$

$$f(x) = \begin{cases} \frac{2}{b-a} \cdot \frac{x-a}{c-a} & \text{se } a \leq x < c \\ \frac{2}{b-a} & \text{se } x = c \\ \frac{2}{b-a} \frac{b-x}{b-c} & \text{se } c < x \leq b \end{cases}$$



$a = 0{,}2$, $b = 0{,}9$, $c = 0{,}6$

```
Fc = (c-a)/(b-a)
U = np.random(N)
Y[U<Fc] = ______
Y[U>Fc] = ______
```

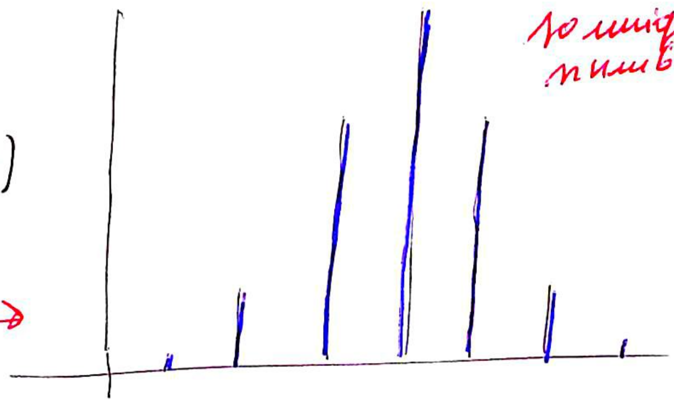
```
plt.hist(Y, 'fd', ___)
___.xlim([a-.2, b+.2])
```

facciamo la stessa cosa con valori diversi –

Esercizio 7 X=np.random.normal(loc=0, scale=1, size=100000)

```
x = np.round(X)
print(np.unique(X))
        plt.hist(X, bins='fd')
```

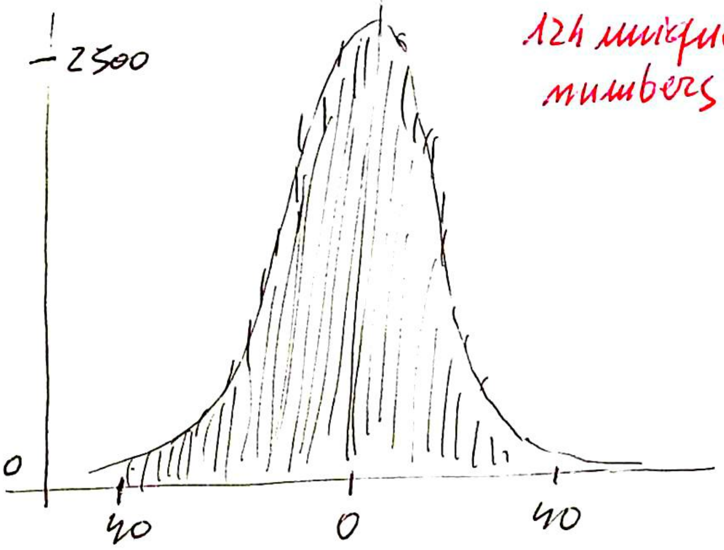
e' corretto, ma lascia a bocca asciutta



10 unique numbers

↳ vario i valori

```
X= np.random.normal(loc=0, scale=15, size=100000)
X = np.round(X)
plt.hist(X, bins='fd')
```



124 unique numbers

Esercizio 8 – combiniamo diverse tecniche, e usiamo un nuovo metodo (correlazione, permutazione)

– matrice 100 x 2 con numeri presi da Gaussiana std

– la seconda colonna = 2ª colonna + la prima

N=100
M = np.random.randn(N,2)
M[:,1]+= M[:,0]

calcoliamo il coefficiente di correlazione

r_real = np.corrcoef(M.T)[1,0] — matrice con diagonale nulla

trasposta

tra -1 e 1 – se r=1 correlazione perfetta
se r=0 " nulla → r = .6 nel nostro caso

ora poniamo a "shuffle", metto in memoria di lavoro

utilizzo numeri casuali interi (posizione) x riorganizzare le righe della prima colonna senza modificare l'ordine della seconda –

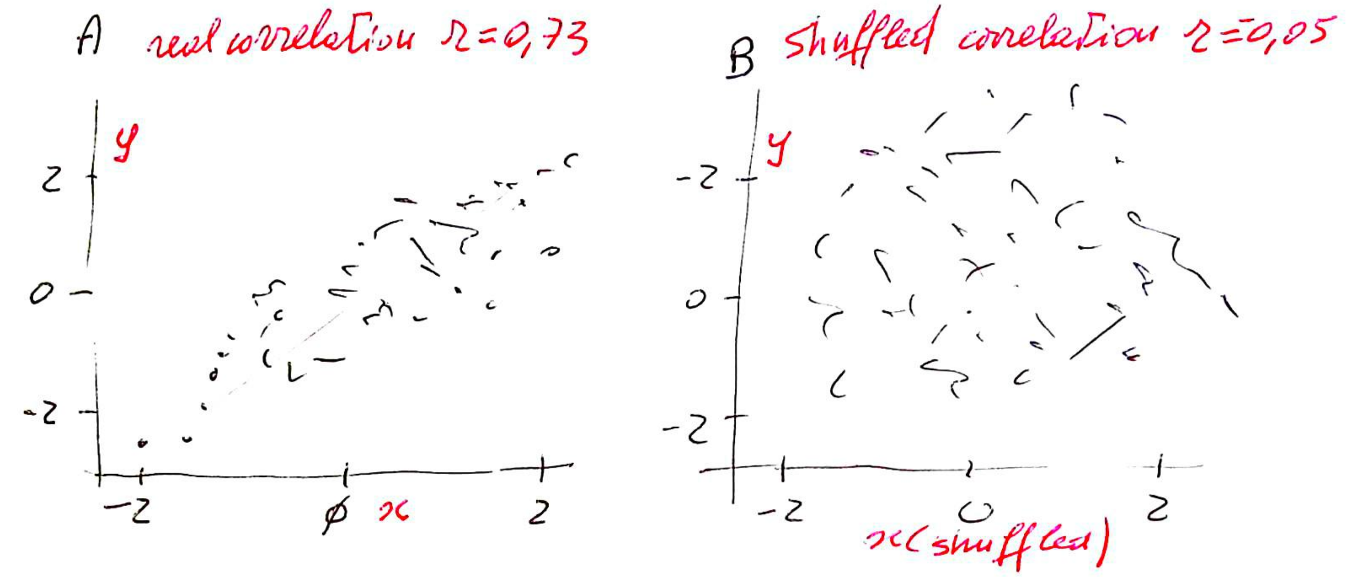
I dati non sono cambiati, è variato l'ordine → dev. std, media non sono cambiate –

Ricalcoliamo il coefficiente di correlazione – Grafichiamo –

→ r_real = 0,666
shuffled correlation = -0,076

Vediamo meglio con un grafico

%



Esercizio 9

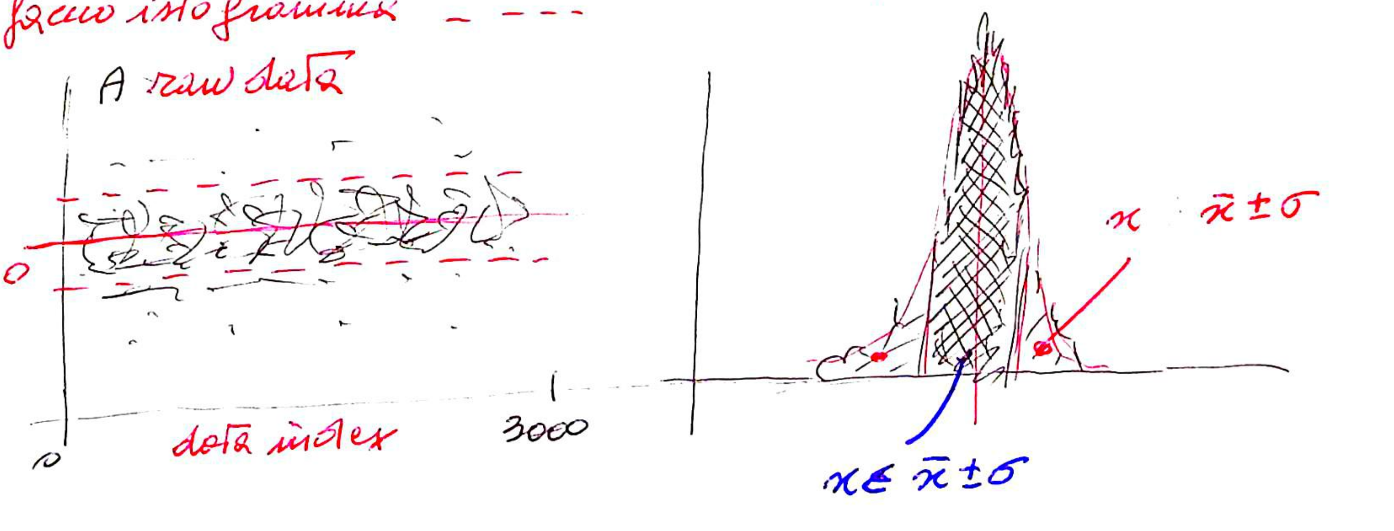
$N = 3000$ andiamo con una di queste

```
data = stats.exponnorm(3, size=N)
 —   = —.norm.rvs(loc=3, scale=1, size=N)
data = stats.laplace.rvs(size=N)
 —   = —.gumbel_r.rvs(size=N)
```

produrre x le varie distribuzioni

calcolo media teorica, dev. std teorica

faccio istogramma - - -



Esercizio 10

$\mu = 0$, $N = 10\,000$, σ in 10 valori $.01 \div 10$

log-normal — la media e la dev. std sono una funzione non lineare di μ e σ — $X = e^N$ con $N = \log X \sim N(\mu, \sigma^2) \Rightarrow$

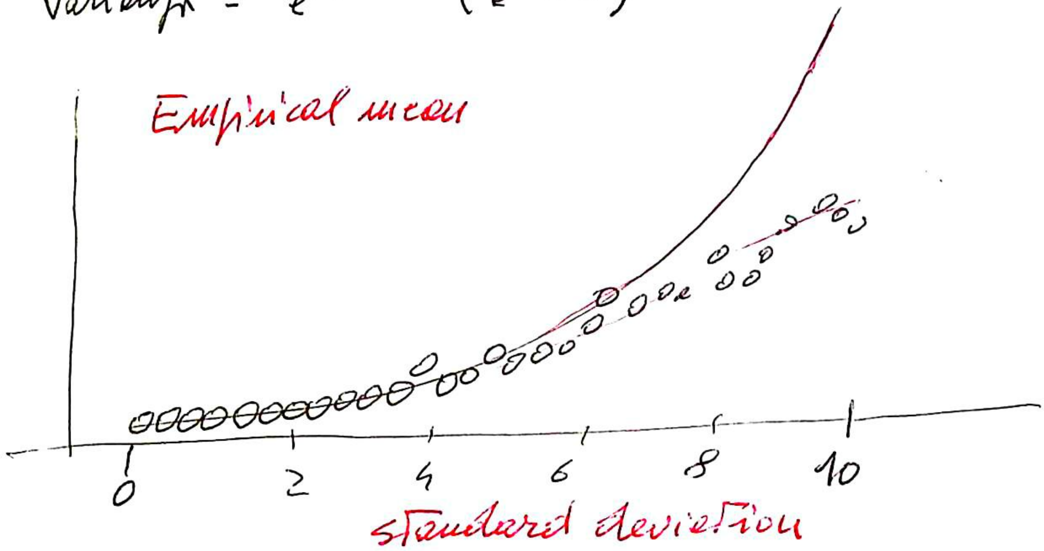
$\Rightarrow X = e^N$ segue una lognormale $\log N(\mu, \sigma^2)$

$$f(x) = \frac{e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}}{x\sqrt{2\pi}\,\sigma} \quad \text{con} \quad x > 0$$

in modo formale $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+_\phi$

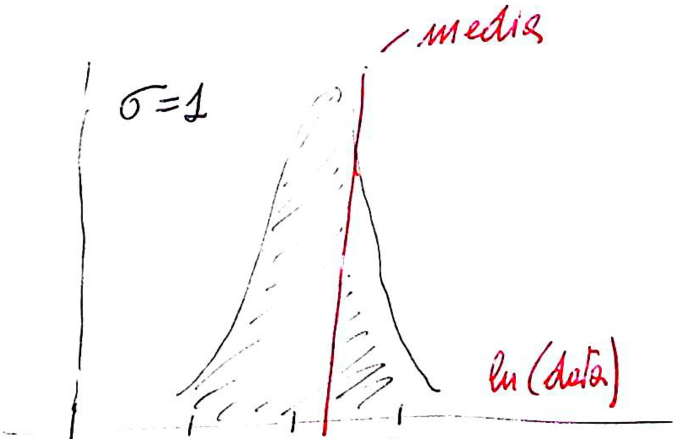
valore atteso $= e^{\mu + \frac{\sigma^2}{2}}$

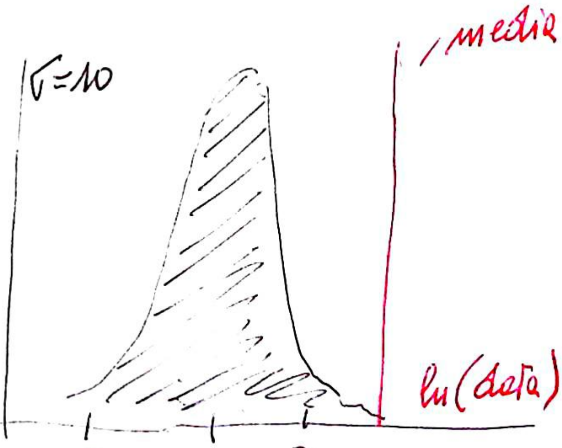
varianza $= e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$



la media attesa varia in funzione della deviazione std.

- si riscontra una sottostima a partire da ca. $\sigma = 5{,}11$
  perché? e^x sale molto rapidamente. Inoltre σ controlla direttamente la ripidezza. Questa è una situazione critica da comprendere.
  Inoltre è inclusa anche la media in questi valori così grandi. Questi valori in concreto mancano nelle distribuzioni empiriche → discrepanza.
- per illustrare il concetto vedere fig. 5.17 che mostra istogrammi di log-transformed data x due valori di σ. Entrambi sembrano apparire corretti, ma anche la media è log-trasformata.





In realtà (concreta), empirica, la probabilità di dati così grandi è piccolissima.

- questa difficoltà si presenta anche nelle statistiche computazionali (es. quando si calcolano i p-values da shuffled data).

– per queste problematiche, riflessione, ragionevolezza.

— FINE CAPITOLO 5 —

6 - TRANSFORMATIONS

6.1 - What, why, and how of data transformation

- Cosa sono le trasformazioni di dati?
  - sequenze di operazioni matematiche sui dati
- Perche' trasformare i dati?
  - moltiplicare x rendere confrontabili le scale
  - passare da una distribuzione → Gauss
  - normalizzare il range di variazione
  - eliminare influenza dei valori troppo estremi (fuori dei 3σ)
  - z-scoring
  - x memorizzare + facilmente
- Come trasformare i dati?
  - scipy.stats
  - usando funzioni scritte da altri
  - funzioni di nostra implementazione
- Quali tipi di trasformazioni?
  - lineari, non lineari
  - lossless, con perdita di dati
  - iterative, non iterative

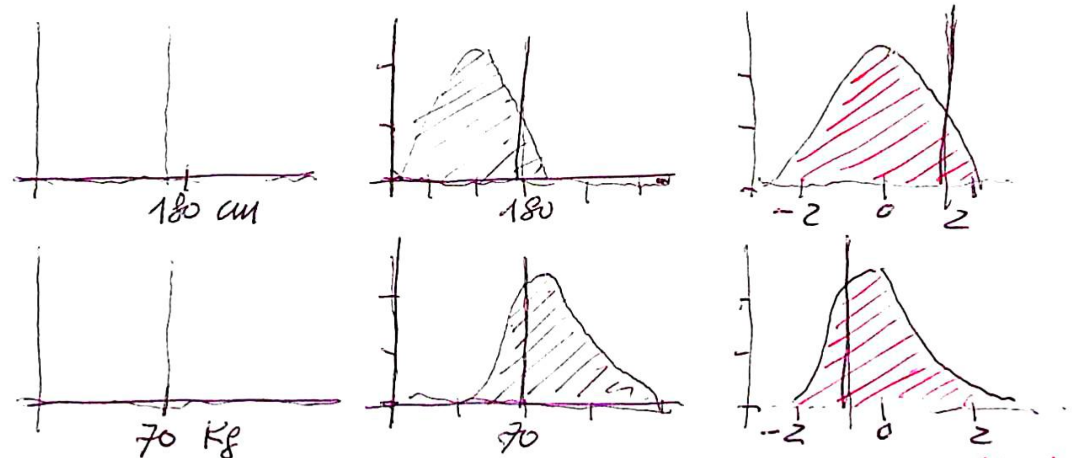
A_{1-2} altezza, peso di una persona

B_{1-2}

C_{1-2} distribuzioni normalizzate (misura in std deviation)

6.2 - Z-score standardization

es. correlare peso e altezza delle persone



raw values — normalized

• z-score math

– Hard & Soft assumptions –

1) shift the data $\longrightarrow \bar{x} = \phi$

2) scale the data to the std deviation

$$z_i = \frac{x_i - \bar{x}}{s} \qquad (6,1)$$

Vediamo i risultati applicati a un caso numerico:

	mean	std
ERA	0,60	3,36
DIVENTA	0,00	1,00

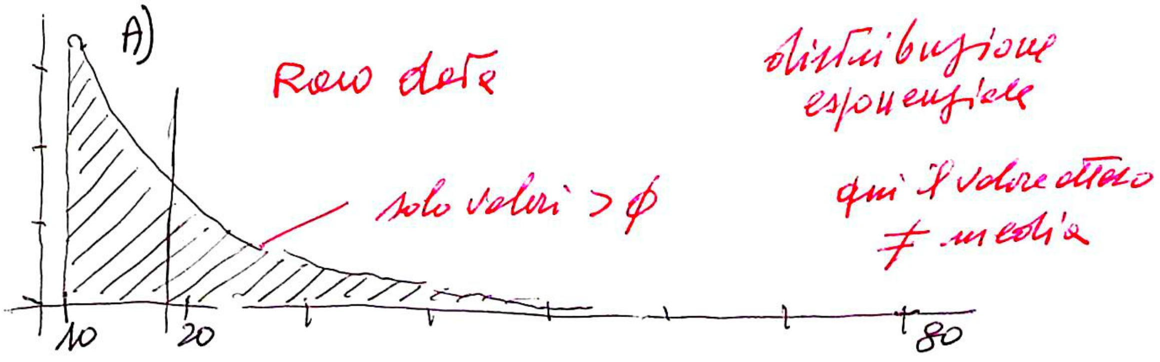
• Interpretazione

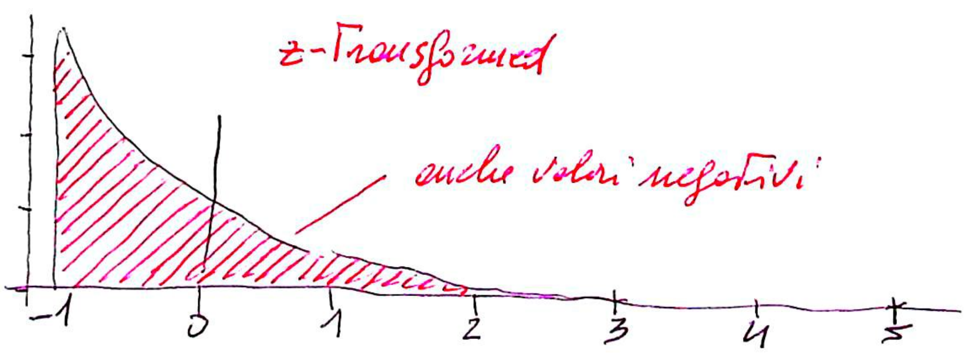
– la z-trasformation sposta e modifica i dati, ma → non varia la distribuzione

• Hard & Soft assumptions

– la std deviation $\neq \phi$ hard

– media e std hanno significato reale in questa distribuzione (può anche non essere vero → soft)
  └ bene applicarla a Gauss





p. 210

• The modified z-score method

- si usa quando la distribuzione è "fortemente non gaussiana" →
  - ↗ regular z-score
  - ↘ modified z-score → sottrae la mediana, divide per la differenza assoluta della mediana

↳ in concreto sottrae una misura di tendenza centrale

$$M_i = \frac{x_i - \tilde{x}}{1{,}4826 \cdot MAD}$$ dove $\tilde{x}$ = mediana, $MAD = \text{median}(|x_i - \tilde{x}|)$

Terzo quartile della distrib. gauss

Vediamo una comparazione:

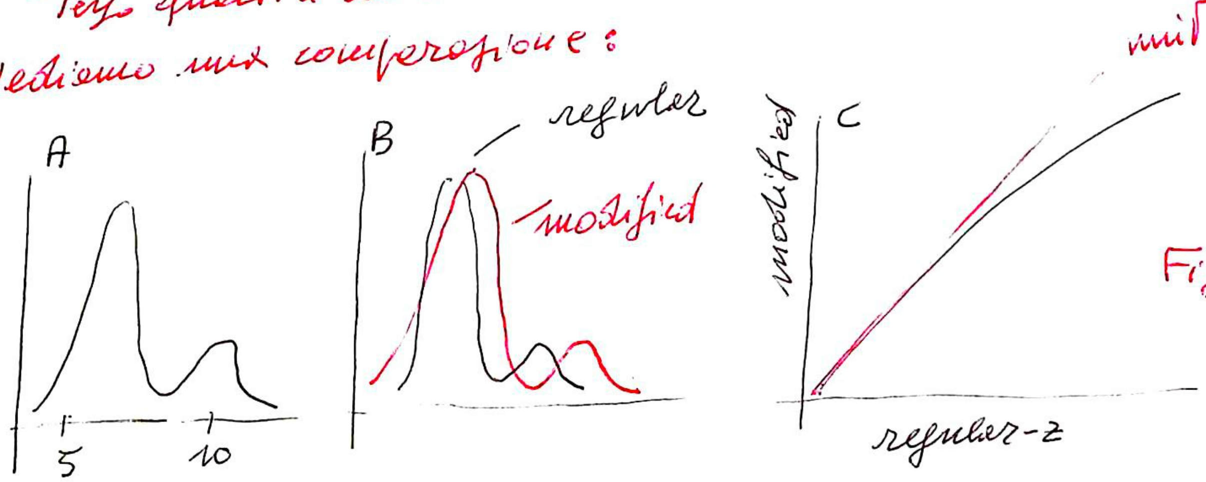


Fig 6.4

- sono tra loro abbastanza simili in generale
- ma allora, quando diventa utile?
- intanto nella modified gli outliers contano poco →
- modified è utile x testare se ci sono valori anomali durante la pulizia dei dati

p. 213

6.3 - Min-max normalization

Tipicamente tra $[0,1]$, $[-1,1]$, $[0, 2\pi]$ range univoco

$$\tilde{x}_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \qquad (6,9)$$

tra i valori reali e quelli riscalati, correlazione = 1

per un range arbitrario

$$x^* = a + (b-a)\,\tilde{x}_i \qquad (6,10)$$

6.4 - Z-scoring vs. min-max scaling

Z-scoring produce una media e una deviazione std nota

la scala min-max non preserva le caratteristiche descrittive, tranne i limiti sup e inf - Min-max si presta meglio se i dati non hanno distribuzioni con lunghe code -

Min-max si presta bene x machine-learning - In questo ambito si fa distinzione tra "normalizzazione" e "standardizzazione"

min-max ┘ └ z-scoring

Trasformazione in %

$$pctchgn = 100 \cdot \frac{ref - new}{ref} \qquad (6,11)$$

la % permette di paragonare dati molto diversi tra loro - La si usa con valori > ∅, con i negativi l'interpretazione è problematica

6.6 - Nonlinear data transformations

Qui l'intento è cambiare la forma della distribuzione - Necessita molta cautela nell'utilizzo - In quelle che presenteremo la relazione monotona tra i punti è conservata -

cioè se $x_1 < x_2 \longrightarrow \tilde{x}_1 < \tilde{x}_2$

questo è molto utile x interpretare i risultati di analisi statistiche come i "Test t", correlazioni, regressioni -

esempio di non monotona: $\tilde{x} = \sin(x)$

06 904

– Rank-transform

modifica i dati numerici con una scala significativa (pollici, euro) in posizioni ordinali. Es.

$$X = [1, 2, 3, 9348753945, 2.01] \qquad (6.12)$$

$$\tilde{X} = [1, 2, 4, 5, 3] \qquad (6.13)$$

prima ordino, assegno un progressivo, creo $\tilde{X}$

qui abbiamo una perdita di informazione

esempio

$$X = [10, 2, 4, 5, 5] \qquad (6.14)$$

$$\tilde{X} = [4, 1, 2, 3, 3]$$

fractional ranking $\tilde{X} = [5, 1, 2, 3.5, 3.5]$ "tied rank"

questa procedura è usata in molte statistiche inferenziali non parametriche (es. test di Wilcoxon, che è una alternativa non parametrica al "t-test") (es. la correlazione di Spearman, che è una correlazione non parametrica).

Logarithm and square root transformation

log-transform si prende il log dei dati

Raw data

Raw data histogram

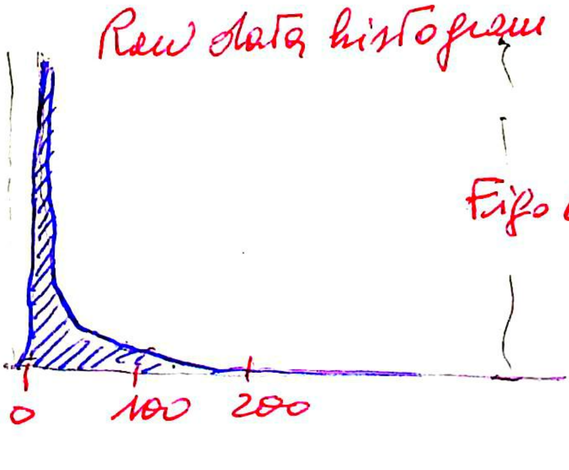
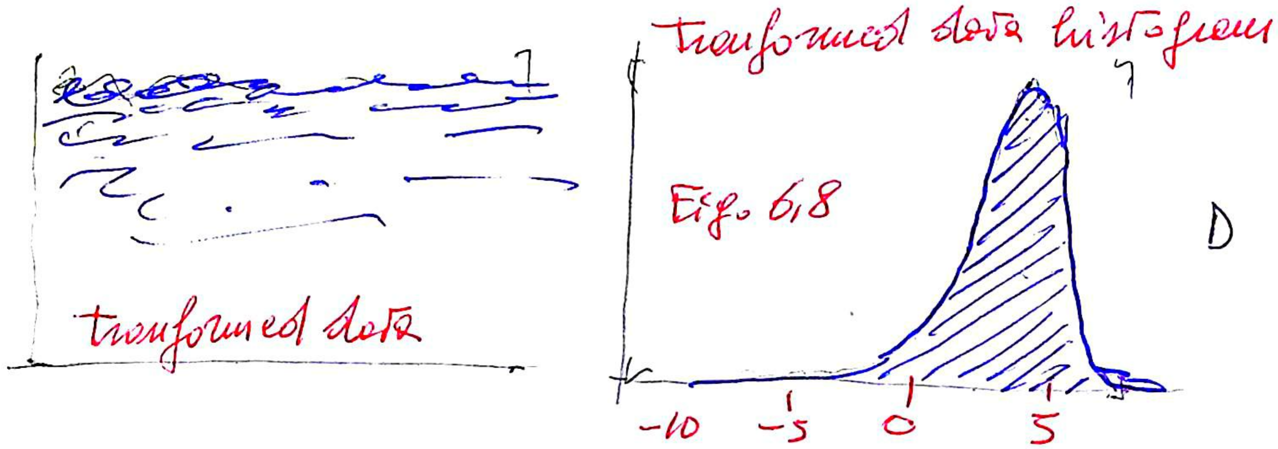


Fig. 6.8



notare la distorsione negativa nel grafico D, però la distribuzione assomiglia abbastanza a una gaussiana.

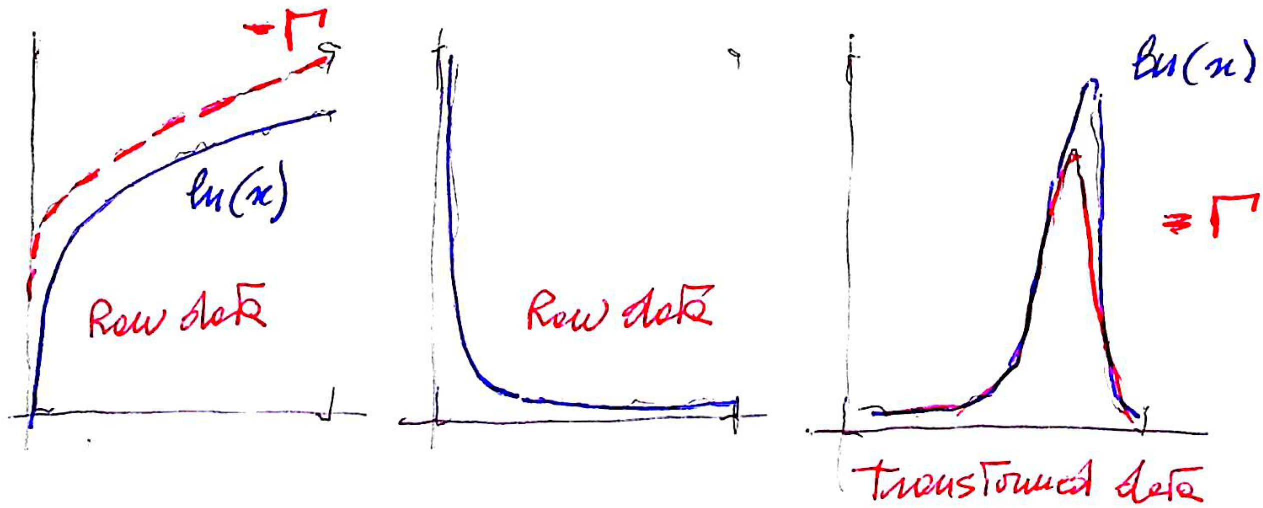
Si presta bene se i dati seguono una legge di potenza → la variabilità è $\propto$ ai valori (eteroschedasticità, vedi ch. 4)

Si può applicare solo per valori $> \phi$.

Se si hanno valori $< \phi$, si può procedere preventivamente a una traslazione.

– Radice quadrata

valida x valori $\geq \phi$. Questa trasformazione può ridurre eteroschedasticità



Fisher-Z

x deformare una distribuzione uniforme $(-1,1)$ in una approssimativamente Gauss. Vengono "allungati" i valori prossimi agli estremi dell'intervallo.

$$x_z = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right) \text{ con } -1 < x < 1 \qquad (6,17)$$

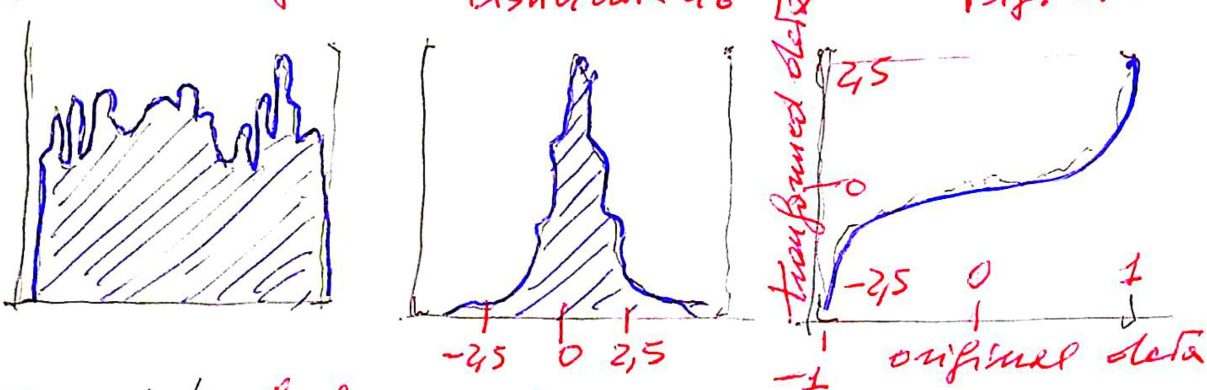


Fig. 6,10

se $x \to \phi$ la frazione $\to 1$

$x \to 1$ $\to$ molto grande > 0

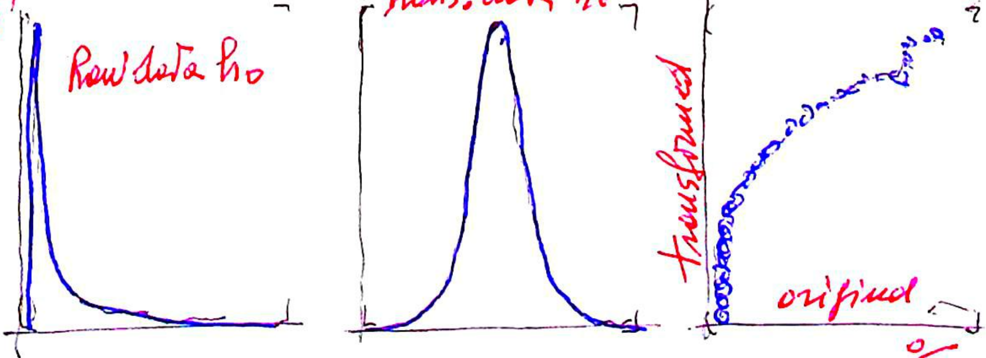
$x \to -1$ $\to \phi \to$ log() diventa $< \phi$

ma la 6,17 equivale alla tangente iperbolica $\to$ la si può ottenere con np.arctanh()

Trasformare ∀ distribuzione in Gaussiana

1) trasformare il rango
2) min-max in [-.999, .999]
3) Fisher-Z $\longrightarrow$ Gaussiana (e monotona)

e' conservativa dei dati



Interpreting transformed data

Z-transform è facile da interpretare. Ogni aumento della deviazione std in A → incremento di .4 nella deviazione std di B

es. trasformo log, quindi eseguo analisi di regressione. Potrei avere un effetto lineare di una variabile su un'altra, ma tutto questo si basa su trasformazioni non lineari → in realtà l'effetto è non lineare. TENERE PRESENTE

- min-max offre una interpretazione + sicura, anche se non necessariamente la + precisa.

When to transform your data

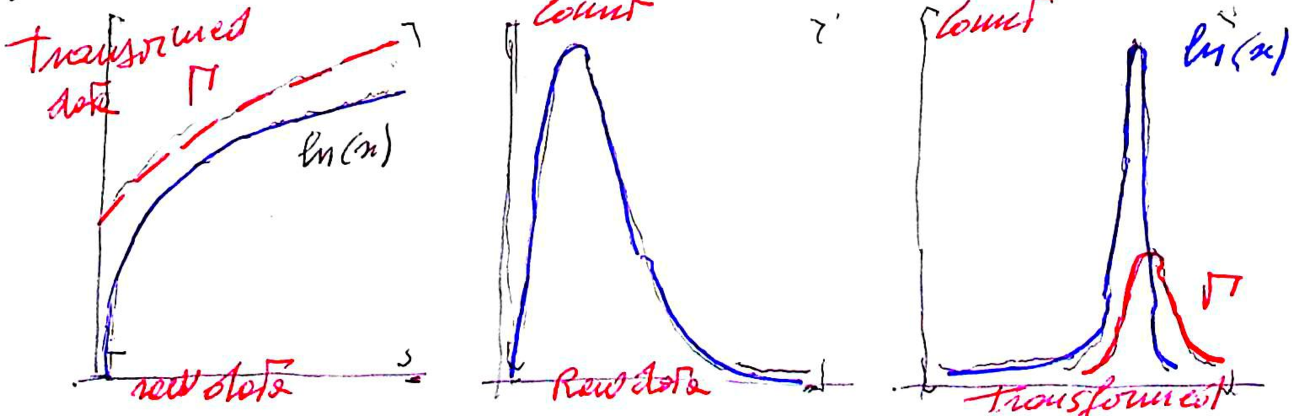
lo z-scoring può essere utile x confrontare variabili su scale diverse.

Usare le trasformazioni solo se necessario.

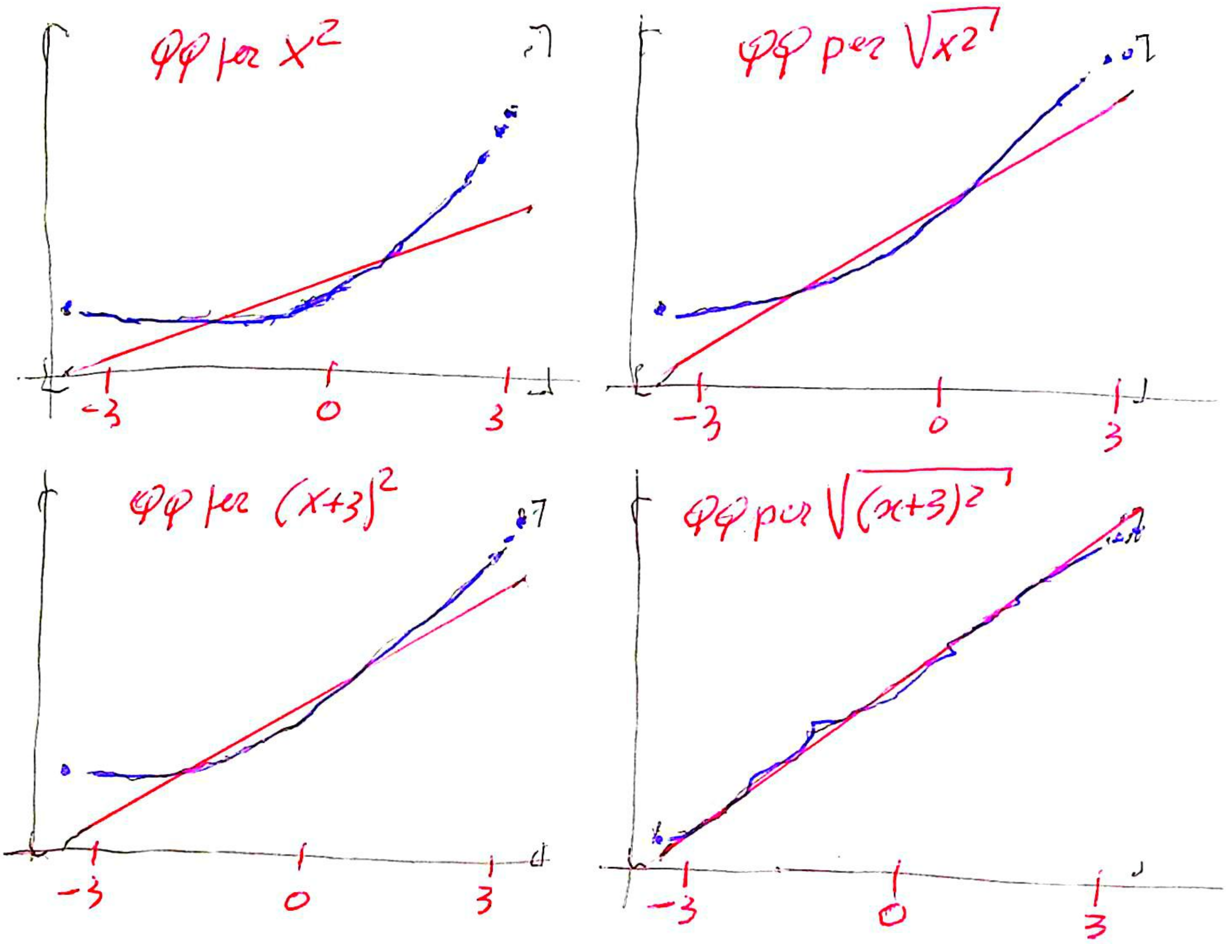
6.8 - Exercises

1) combinare insieme 6.9 e 6.10 - tradurre in Python - testare su 10 numeri Gauss (14,3, 34) scalare →
   min value = 14,3 ; max value = 34
   su Web ∃ funzione che lo fa, trovarla

2) Fig. 6.8 - Modifica usando $(X+3)^2$ invece che X^2



• ricreiamo ora i QQ plots



infine sposta i dati trasformati in T in modo che abbiano media $= \phi$

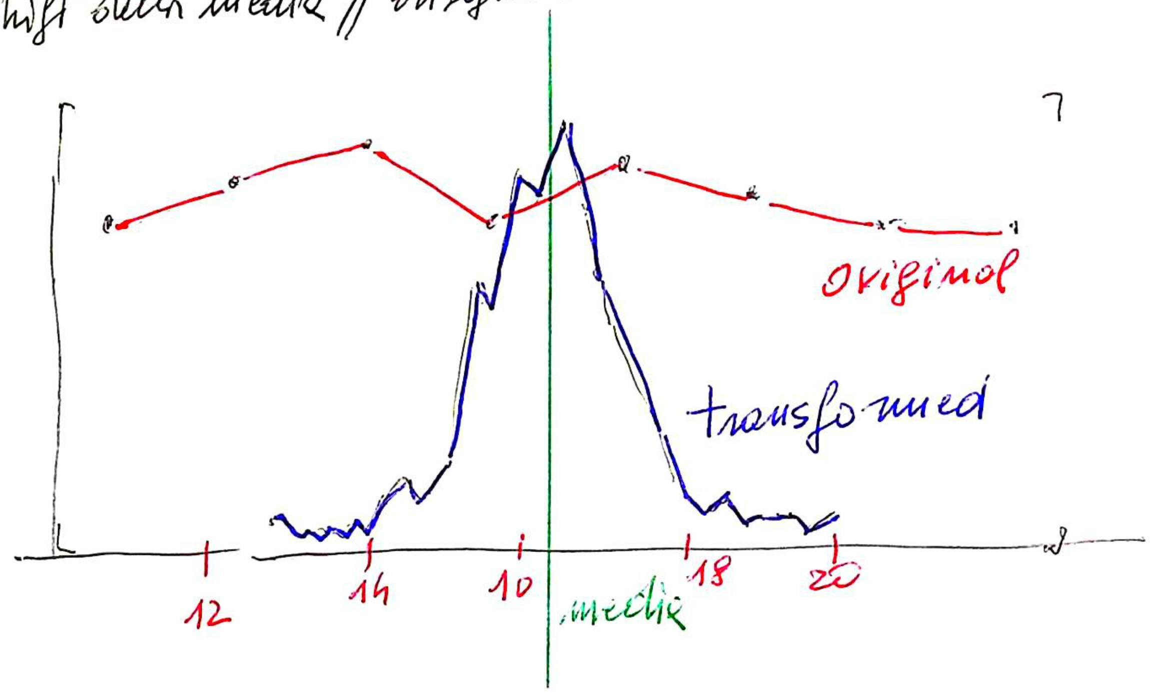
3) la normalizzazione perde dati? Risolvere la 6,1 per x_i, se questo è possibile → la trasformazione è invertibile. Fare questo anche con codice, mediante 25 interi casuali tra 1 e 14. z-trasformiamo, antitrasformiamo. Verificare che il risultato coincida con originale.

4) applichiamo molte trasformazioni / riconduci a Gauss preservando la media dei dati.

313 numeri casuali da distribuzione uniforme tra 3π e π
assicurati che tutti i campioni siano reali:
hist di raw data, hist transformed data

– descrizione procedura:

1) genero i dati // 2) noto tra (−999, 999) // Fisher-Z // shift della media // disegno i due hist diventano Gauss



5) ∄ np function x z-score ⟶ scipy.stats − video 3 e 9

scipy.stats.zscore oppure scale

calcolare entrambi e confrontare che siano uguali

− ora proviamo con una matrice_ Calcolare $\bar{X}$ e s da ciascuna colonna separatamente, o da intera matrice?

− creare una matrice con valori di x^2 con $0 \leq x < 11$ applicare scipy.stats.zscore

$$\begin{bmatrix} 0 & 1 & 4 \\ 9 & 16 & 25 \\ 36 & 49 & 64 \\ 81 & 100 & 121 \end{bmatrix}$$

calcolo

Column-wise z-scoring

Matrix-wise z-scoring

posso calcolare std, media x ogni colonna, ma è sufficiente ispezione visiva

z-scoring x Matrice può essere appropriato quando tutte le colonne contengono caratteristiche nella stessa scala e intervallo numerico_

6) trovare modo di usare Fisher-Z x generare una distribuzione di numeri casuali (A) nessuna simmetria (B) forte asimmetria positiva (C) leggera asimmetria negativa

– calcolare l'asimmetria empirica, graficare

X1 np.arctanh(np.random.uniform(-.999, .999, size=5000)) (A)

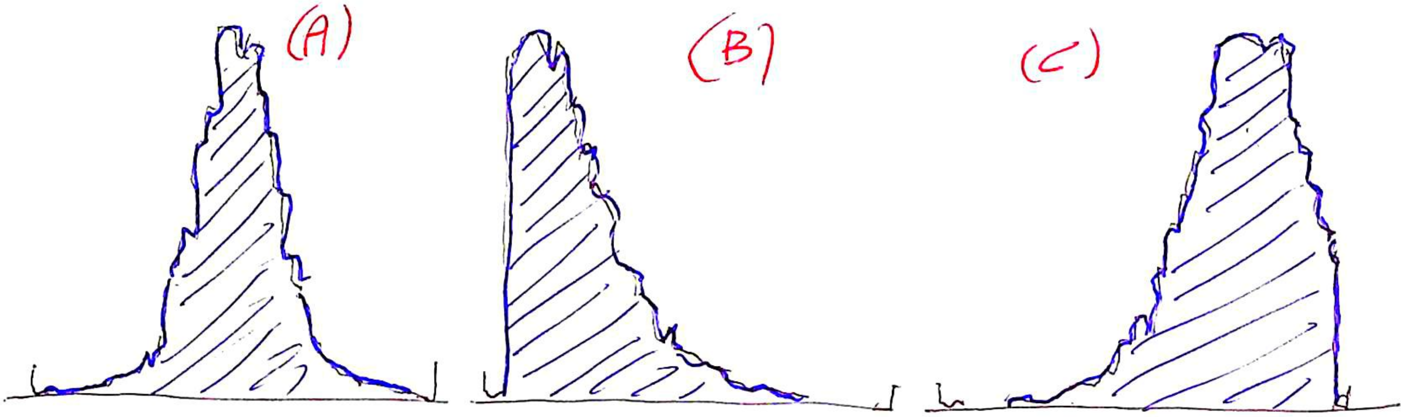
X2 ———————— (-.2, .999 ————) (B)

X3 ———————— (-.999, .8 ————) (C)

skew[0] = stats.skew(X1)

——— 1 ——————— X2

2 X3



skew = 0,02 skew = 1,25 skew = -0,74

7) esplorare relazione tra std e MAD in distribuzioni che si allontanano da Gauss

– dati casuali esponenziali tra 0,2 e 1 – Calcolare tre statistiche descrittive (1) std (2) MAD (3) Scaled MAD – Graficare

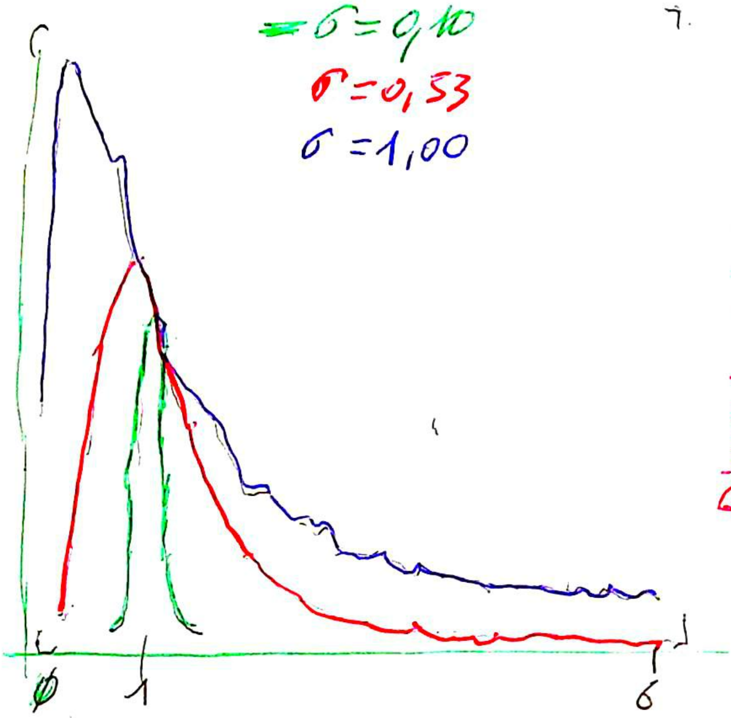
data[i] = np.exp(X*s) casuali n dati

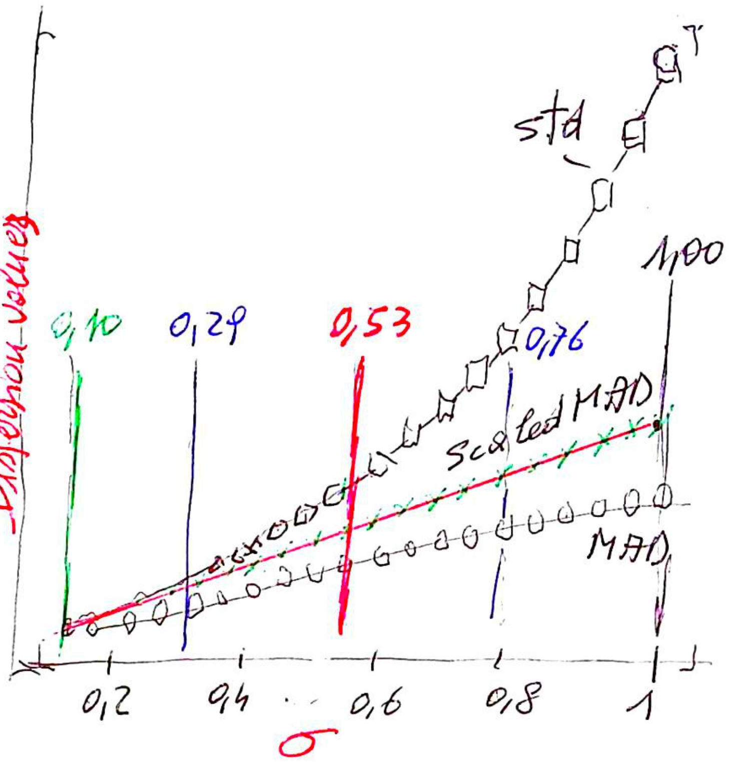
M[i,0] = np.std(data[i], ddof=1)

M[i,1] = np.median(np.abs(data[i] - np.median(data[i])))

sigmas = np.linspace(.1, 1, 20) vari valori di σ

%

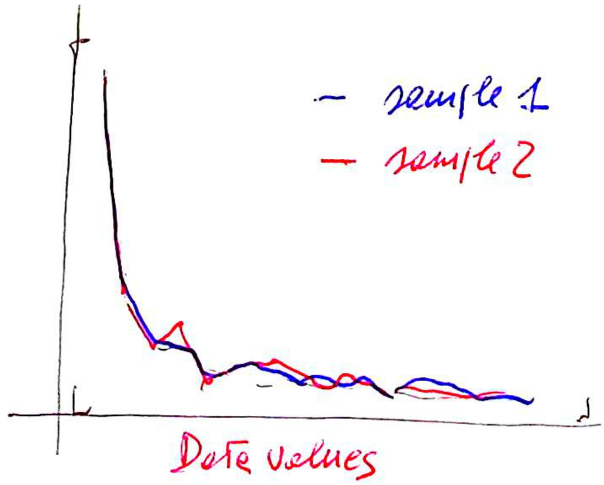


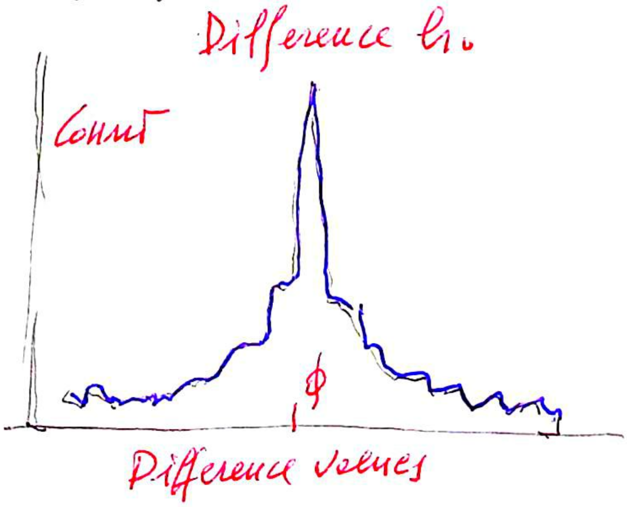


8) La differenza tra due variabili distribuite in modo simile, non normale → può essere ~ Gauss

2 set di dati $N = 300$ casuali da legge di potenza

loro istogrammi, e istogramma differenza.





— FINE CAPITOLO 6 —