Rottegoda - Rosso, 2008

Chapter 1
Preliminary Data Analysis

All natural processes, as well as those devised by humans, are subject to variability. Civil engineers are aware, for example, that crushing strengths of concrete, soil pressures, strengths of welds, traffic flow, floods, and pollution loads in streams have wide variations. These may arise on account of natural changes in properties, differences in interactions between the ingredients of a material, environmental factors, or other causes. To cope with uncertainty, the engineer must first obtain and investigate a *sample* of data, such as a set of flow data or triaxial test results. The sample is used in applying statistics and probability at the descriptive stage. For inferential purposes, however, one needs to make decisions regarding the *population* from which the sample is drawn. By this we mean the total or aggregate, which, for most physical processes, is the virtually unlimited universe of all possible measurements. The main interest of the statistician is in the aggregation: extrapolation → e.g. INFERENCE the individual items provide the hints, clues, and evidence.

A data set comprises a number of measurements of a phenomenon such as the failure load of a structural component. The quantities measured are termed *variables*, each of which may take any one of a specified set of values. Because of its inherent randomness and hence unpredictability, a phenomenon that an engineer or scientist usually encounters is referred to as a *random variable*, a name given to any quantity whose value depends on chance.[1] Random variables are usually denoted by capital letters. These are classified by the form that their values can possibly take (or are assumed to take). The pattern of variability is called a *distribution*. A *continuous* variable can have any value on a continuous scale between two limits, such as the volume of water flowing in a river per second or the amount of daily rainfall measured in some city. A *discrete* variable, on the contrary, can only assume countable isolated numbers like integers, such as the number of vehicles turning left at an intersection, or other distinct values.

Having obtained a sample of data, the first step is its presentation. Consider, for example, the modulus of rupture data for a certain type of timber shown in Table E.1.1, in Appendix E. The initial problem facing the civil engineer is that such an array of data by itself does not give a clear idea of the underlying characteristics of the stress values in this natural type of construction material. To extract the salient features and the particular types of information one needs, one must summarize the data and present them in some readily comprehensible forms. There are several methods of presentation, organization, and reduction of data. Graphical methods constitute the first approach.

1.1 GRAPHICAL REPRESENTATION

If "a picture is worth a thousand words," then graphical techniques provide an excellent method to visualize the variability and other properties of a set of data. To the powerful interactive system of one's brain and eyes, graphical displays provide insight into the form

[1] The term will be formally defined in Section 3.1.

ANALISI ESPLORATIVA
DI SERIE DI OSSERVAZIONI

Rappresentazione tabellare della serie storica

Sequenza cronologica

Anno	h [mm]
1918	150
1919	129.3
1920	140.8
1921	110
1922	72.5
1923	75.5
1924	195
1925	130
1926	293
1927	79
1928	190
1929	-
1930	209
1931	99
1932	110
1933	138
1934	137
1935	182
1936	155
1937	155
1938	152
1939	224
1940	125

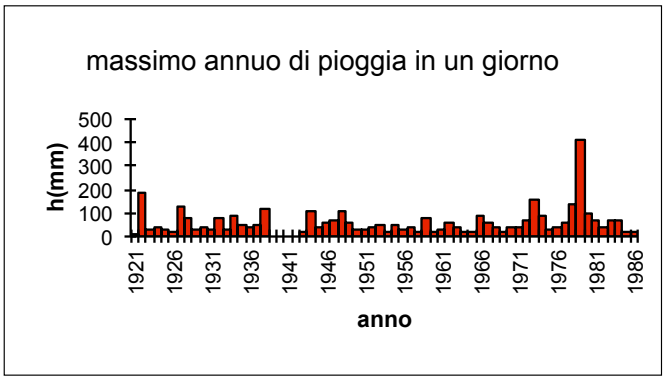
Sequenza ordinata

i	h_i
1	72
2	72.5
3	74
4	75.5
5	79
6	83
7	92
8	95
9	97
10	99
11	107.6
12	110
13	110
14	120
15	125
16	126
17	126.8
18	127
19	129
20	129.3
21	130
22	135
23	137
24	138
25	140.8
26	145
27	147

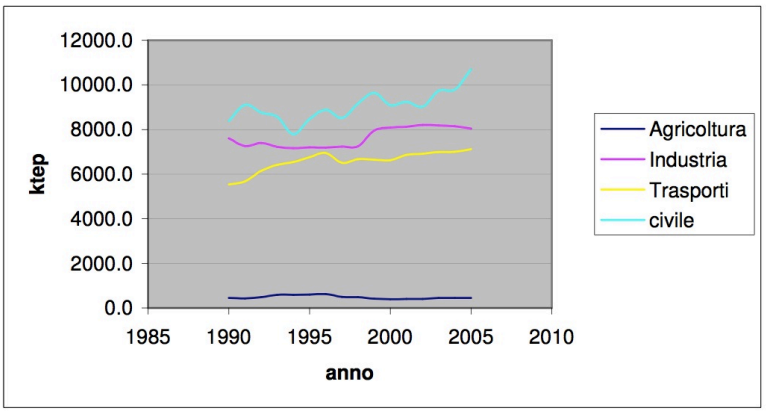
Osservazioni di massimo annuo di pioggia in un giorno

Rappresentazione grafica della serie storica (Sequenza cronologica)

Anno	h [mm]
1918	150
1919	129.3
1920	140.8
1921	110
1922	72.5
1923	75.5
1924	195
1925	130
1926	293
1927	79
1928	190
1929	-
1930	209
1931	99
1932	110
1933	138
1934	137
1935	182
1936	155
1937	155
1938	152
1939	224
1940	125



Altro esempio: regione Lombardia. Consumi energetici annui



Non stazionarietà (Civile, industria)
Bassa variabilità (Agricolt.)

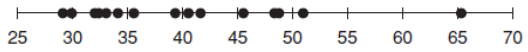
Figura 2 – Consumi finali totali per settore economico

i	h_i
1	72
2	72.5
3	74
4	75.5
5	79
6	83
7	92
8	95
9	97
10	99
11	107.6
12	110
13	110
14	120
15	125
16	126
17	126.8
18	127
19	129
20	129.3
21	130
22	135
23	137
24	138
25	140.8
26	145
27	147

Distribuzione del campione (Caratteristiche di variabilità)

Ampiezza del campione $A = (x_{max} - x_{min})$

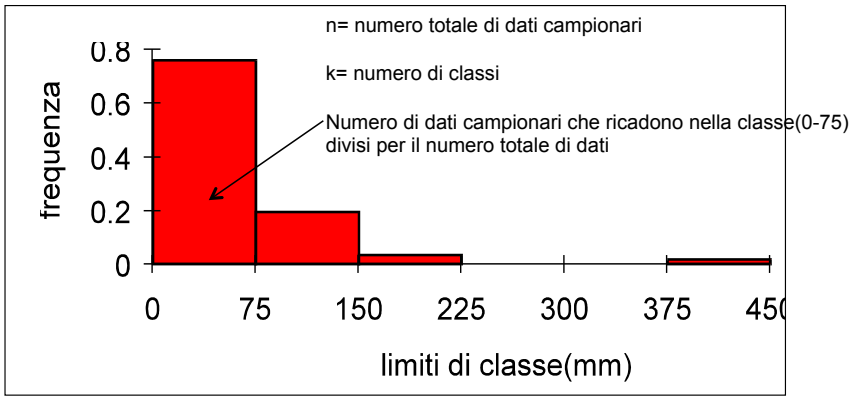
Diagramma a punti



Distribuzione del campione (Caratteristiche di variabilità)

- Rappresentazione ad **istogramma** delle **frequenze di classe**, sia
- Assolute (*n°elementi per classe*), che
- Relative: (*n° elementi per classe divisi per N=numero totale dati*)

frequenza relativa

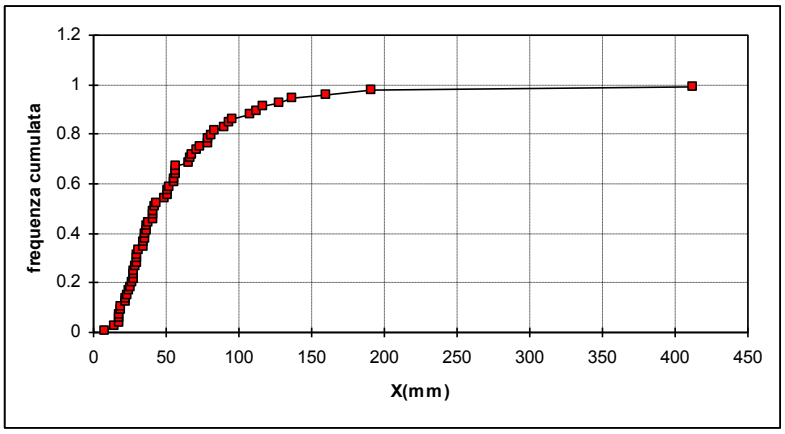


Per evitare arbitrarietà nella determinazione del numero di classi, si può utilizzare la relazione suggerita da Sturges che lega il numero delle classi, k, alla dimensione del campione, N, secondo la relazione : $k = \text{int}(1 + 3.3 \log N)$
(logaritmo in base 10)

Distribuzione del campione (Caratteristiche di variabilità)

i	h_i		*i/N*
1	72		0.0182
2	72.5		0.0364
3	74		0.0545
4	75.5		0.0727
5	79		0.0909
6	83		0.1091
7	92		0.1273
8	95		0.1455
9	97		0.1636
10	99		0.1818
11	107.6		0.2000
12	110		0.2182
13	110		0.2364
14	120		0.2545
15	125		0.2727
16	126		0.2909
17	126.8		0.3091
18	127		0.3273
19	129		0.3455
20	129.3		0.3636
21	130		0.3818
22	135		0.4000
23	137		0.4182
24	138		0.4364
25	140.8		0.4545
26	145		0.4727
27	147		0.4909

Curva di *Frequenza cumulata (campionaria)*



Frequenza cumulata campionaria: $\phi(x_i) = \frac{i}{n}$

MOMENTI CAMPIONARI

Media campionaria

singolo dato campionario

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Varianza

Media campionaria

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

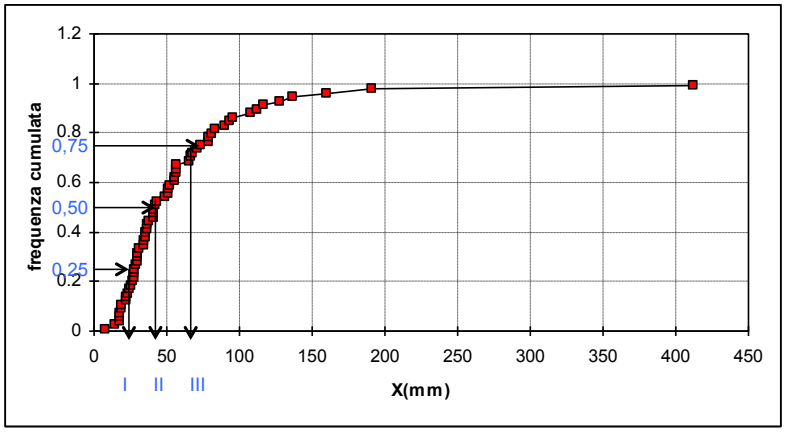
Coefficiente di asimmetria (skewness)

$$Ca = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

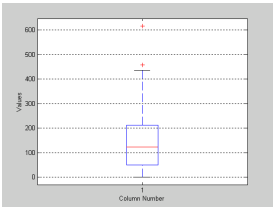
Coefficiente di appiattimento (kurtosi)

$$\kappa_s = \frac{1}{n}\frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4}$$

QUARTILI del Campione



Rappresentazione Box-Plot della serie



Limiti del *box*:

Inferiore:	I quartile del campione	x(Φ=0.25)
Superiore:	III quartile del campione	x(Φ=0.75)
Linea mediana:	II quartile del campione	x(Φ=0.50)

Si definisce *range interquartile* (**IQR**) la differenza:

IQR = $X(\Phi=0.75) - X(\Phi=0.25)$

Limiti dei *whiskers*:

Inferiore

Valor <u>minimo</u> della serie delle osservazioni (X_1)

oppure

I quartile - 1.5 volte ***IQR*** → $X(\Phi=0.25)$ - 1.5 ***IQR***

Se negativo può essere posto pari a zero quando le osservazioni sono definite positive

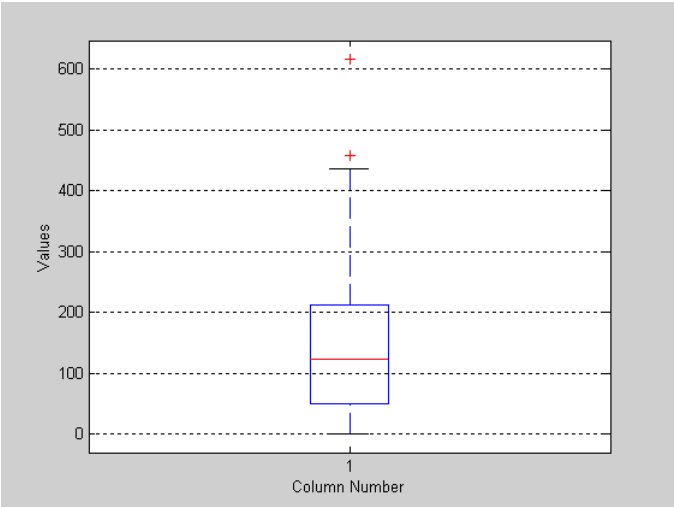
Superiore

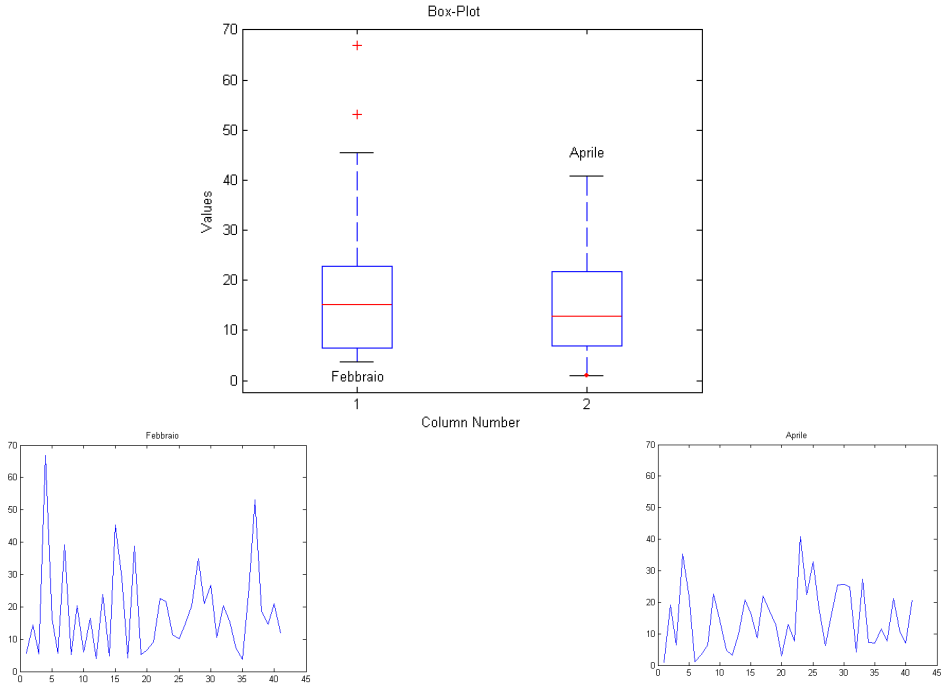
Valor <u>massimo</u> della serie delle osservazioni (X_n)

oppure

III quartile + 1.5 volte ***IQR*** → $X(\Phi=0.75)$ + 1.5 ***IQR***

Nella rappresentazione con i *whiskers* si possono indicare tutte le osservazioni di valore inferiore al *whisker* minimo e superiore al *whisker* massimo





SIMBOLOGIA

$P(.)$	probabilità associata ad una data condizione
$F(x)$	distribuzione di probabilità cumulata (teorica, della popolazione)
$\phi(x)$	frequenza cumulata (campionaria) associata al dato valore di x
$f(x)$	densità di probabilità
$p(x)$	massa di probabilità di funzione discreta
$x(F)$	quantile per fissata probabilità F)
θ_i, $\widehat{\theta}_i$	parametri e rispettivi valori stimati
μ, $E(x)$	media della distribuzione (teorica)
$\overline{x}$	media campionaria
σ^2, $var(x)$	varianza della distribuzione (teorica)
s_x^2, s_x	varianza e scarto quadratico medio del campione

CONCETTI FONDAMENTALI DELLA TEORIA DELLE PROBABILITA'.

Esperimento aleatorio.

Spazio campionario o popolazione.

Esempi:

Esperimento	Popolazione	Tipo
Numero di giorni piovosi in un anno	$\{0,1,2,.....365\}$	Finito, numero
Numero di giorni non piovosi consecutivi	$\{0,1,2.....\}$	Infinito, numero
Valori osservati della portata	$\{x; x \geq 0\}$	Infinito, non numero

- Evento aleatorio semplice A, B, C : ciascun elemento della popolazione (punto).
- Evento aleatorio composto A, B, C : insieme di due o più punti.
- Complemento dell'elemento A : $\overline{A}$: insieme dei punti che non appartengono ad A.
- Evento certo Ω : insieme di tutti i punti della popolazione.
- Evento nullo Φ : insieme vuoto.
- Unione di eventi A e B : $A \bigcup B$: insieme dei punti dei due eventi.
- Intersezione di $A \bigcap C$: insieme dei punti comuni ad A e a B

PROPRIETA' FONDAMENTALI DELLA PROBABILITA'.

Probabilità dell'evento A:

1. $0 \leq P[A] \leq 1$
2. $P[\Omega] = 1$
3. Se $B = A_1 \cup A_2 \cup A_3 \ldots$. e se $A_1, A_2, A_3 \ldots\ldots$ sono mutuamente escludentisi:

$$P[B] = P[A_1] + P[A_2] + P[A_3] + \ldots.$$

Esempi:

$P[\overline{A}] = 1 - P[A]$

$P[\Phi] = 1 - P[\Omega] = 0$

PROBABILITA' DELLE UNIONI DI EVENTI.

Esempio: In un istituto universitario vi sono 10 docenti (3 donne e 7 uomini) e 30 non docenti (10 donne e 20 uomini).

Qual è la probabilità che un membro dell'istituto scelto a caso sia un docente e/o una donna?

$$P[D \cup F] = P[D] + P[F] - P[D \cap F]$$

Eventi <u>mutuamente escludentisi:</u> $P[A \cap B] = 0$

PROBABILITA' CONDIZIONATA.

Esempio: Qual è la probabilità che un membro dell'istituto donna sia docente?

$$P[D|F] = \frac{P[D \cap F]}{P[F]}$$

Eventi statisticamente indipendenti: $P[A \cap B] = P[A] \cdot P[B]$

Eventi mutuamente escludentisi: $P[A|B] = 0$

TEOREMA DELLA PROBABILITA' TOTALE.

A evento qualsiasi.

$$B_1 B_2, B_3 \ldots . B_n \begin{cases} eventi\ mutuamente\ escludentisi. \\ B_1 \cup B_2 \cup \ldots . \cup B_n = \Omega \end{cases}$$

$(B_1 \cap A), (B_2 \cap A), (B_3 \cap A), \ldots ., (B_n \cap A)$ Altra serie di eventi mutuamente escludentisi.

$$(B_1 \cap A) \cup (B_2 \cap A) \cup (B_3 \cap A) \cup \ldots . \cup (B_n \cap A) = A$$

$$P[A] = P[A|B_1]P[B_1] + P[A|B_2]P[B_2] + \ldots \ldots P[A|B_n]P[B_n]$$

VARIABILI ALEATORIE E LORO DISTRIBUZIONE.

Variabili aleatoria o casuale.

Il valore assunto da una variabile aleatoria associata con un esperimento dipende dal risultato dell'esperimento.

Ad ogni punto dello spazio campionario si associa un valore della variabile.

Esempio:

"Testa e Croce": due monete (argento e oro) lanciate simultaneamente.

Variabile aleatoria (v.a.): X numero di teste ottenute.

Evento Semplice	**Descrizione**		**Valore di** X
	Dorata	**Argentata**	
A	Croce	Croce	$x = 0$
B	Testa	Croce	$x = 1$
C	Croce	Testa	$x = 1$
D	Testa	Testa	$x = 2$

VARIABILI DISCRETE.

V.a. che possono assumere solo valori interi un dato intervallo.

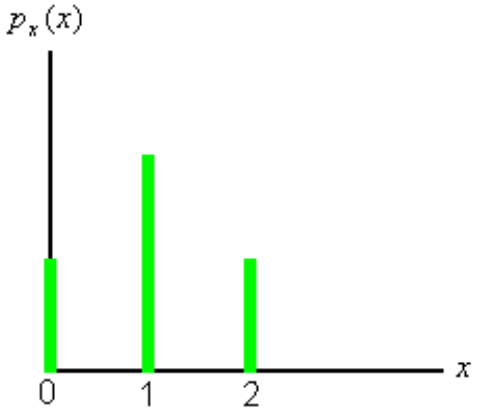
Funzione massa di probabilità (f.m.p.) associa una probabilità ad ogni valore della variabile.

$$P[X = x] = p_X(x)$$

Esempio:

$$p_X(0) = P[X = 0] = P[A] = \frac{1}{4} \qquad p_x(1) = P[X = 1] = P[B \cup C] = P[B] + P[C] = \frac{1}{2}$$

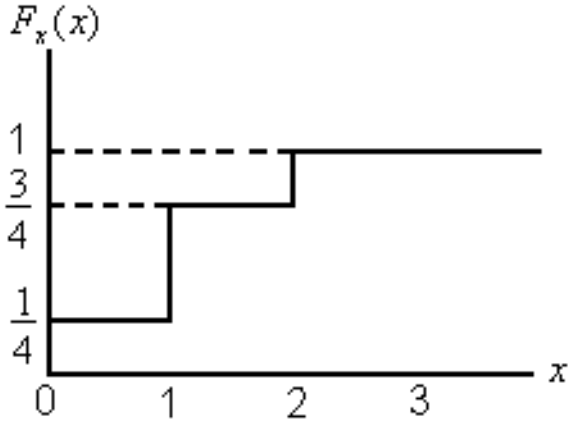
$$p_X(2) = P[X = 2] = P[D] = \frac{1}{4}$$



- $0 \le p_X(x) \le 1$
- $\sum p_X(x_i) = 1$
- $P[a \le X \le b] = \sum_{a \le x_i \le b} p_x(x_i)$

Funzione di distribuzione cumulata.

$$F_X(x) = P[X \le x] = \sum_{x_i \le x} p_X(x_i)$$



Esempio:

$$F_X(x) = P[X \le x] = \sum_{x_i \le x} p_X(x_i)$$

$$F_X(2) = 1 \quad F_X(1) = \frac{3}{4} \quad F_X(0) = \frac{1}{4} \quad F_X(-1) = 0$$

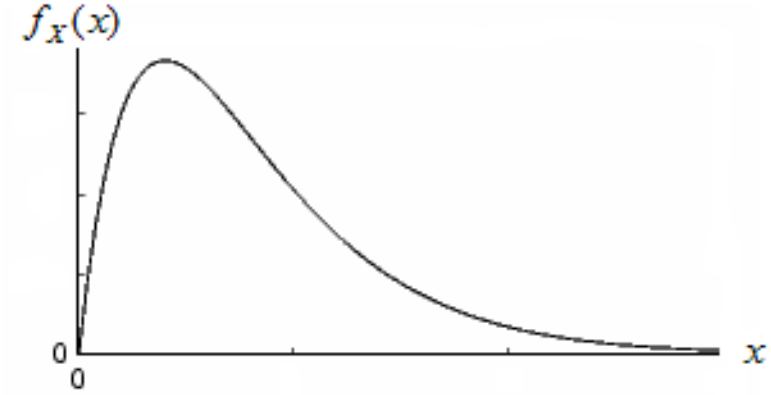
VARIABILI ALEATORIE CONTINUE.

Possono assumere qualsiasi valore numerico reale in un dato intervallo.

Funzione di densità di probabilità.

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P\left[x - \frac{\Delta x}{2} \le X \le x + \frac{\Delta x}{2}\right]}{\Delta x}$$

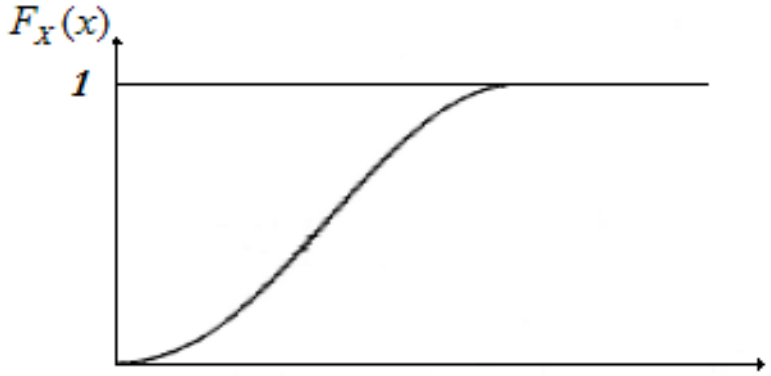
- $f_X(x) \ge 0$ $\qquad \int_{-\infty}^{+\infty} f_X(x)dx = 1$ $\qquad P[a \le X \le b] = \int_a^b f_X(x)dx$



Funzione di distribuzione cumulata:

$$F_X(x) = P[X \le x] = \int_{-\infty}^{+\infty} f_X(u)du$$

- $\frac{dF_X(x)}{dx} = f_x(x)$ solo per variabili assolutamente continue.
- $F_X(\infty) = 1;$ $F_X(-\infty) = 0$
- $F_X(x+\varepsilon) \ge F_x(x)$ per qualsiasi $\varepsilon > 0$; $F_X(x_2) - F_X(x_1) = P[x_1 \le X \le x_2]$



Per ogni tipo di variabile definita nell'intervallo $[a,b]$:

- $0 \le F_X(x) \le 1;$ $F_X(a) = 0;$ $F_X(b) = 1$

MOMENTI

MEDIA (VALORE SPERATO) di una variabile aleatoria discreta.

$$E[X] = \sum_{x_i} x_i P_x(x_i) = \mu$$

di una variabile aleatoria continua.

$$E[X] = \int_{-\infty}^{+\infty} x f_x(x) dx = \mu$$

di una funzione $g(x)$ di una v.a. continua o discreta:

$$E[g(x)] = \sum_{x_i} g(x_i) P_x(x_i)$$

$$E[g(x)] = \int_{-\infty}^{+\infty} g(x) f_x(x) dx$$

r-esimo momento di X :

$$\left.\begin{aligned} \mu_r &= E[X^r] = \sum_{x_i} x_i^r P_x(x_i) \\ \mu_r &= E[X^r] = \int_{-\infty}^{+\infty} x^r f_x(x) dx \end{aligned}\right\} \mu_1 = \mu = E[X]$$

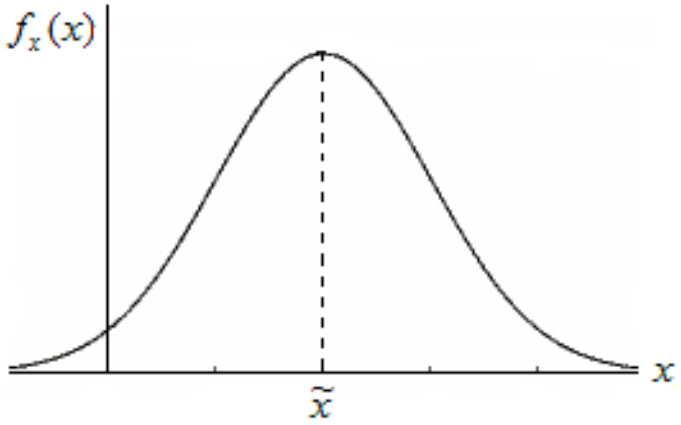
r-esimo momento di centrale di X :

$$\left.\begin{aligned}\mu_r^{'} &= E\left[(X-\mu)^r\right] = \sum_{x_i}(x_i-\mu)^r P_x(x_i)\\ \mu_r^{'} &= E\left[(X-\mu)^r\right] = \int_{-\infty}^{+\infty}(x-\mu)^r f_x^{'}(x)dx\end{aligned}\right\}\mu_1^{'}=0;\mu_2^{'}=\text{var}\left[X\right]=\sigma^2=E\left[X^2\right]-E^2\left[X\right]=\mu_2^{'}-\mu^2$$

MISURA DI LOCAZIONE.

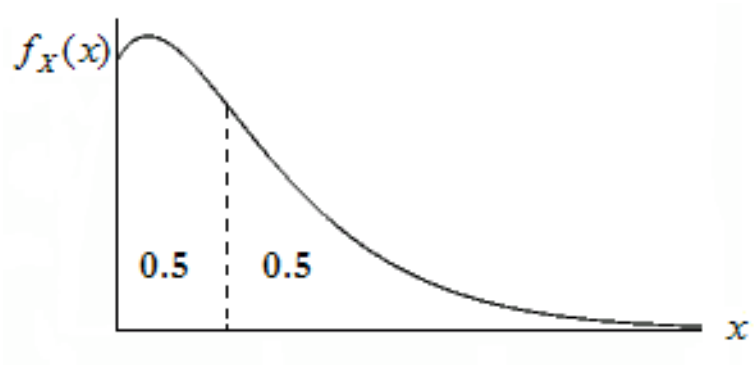
Moda: $\tilde{x}$

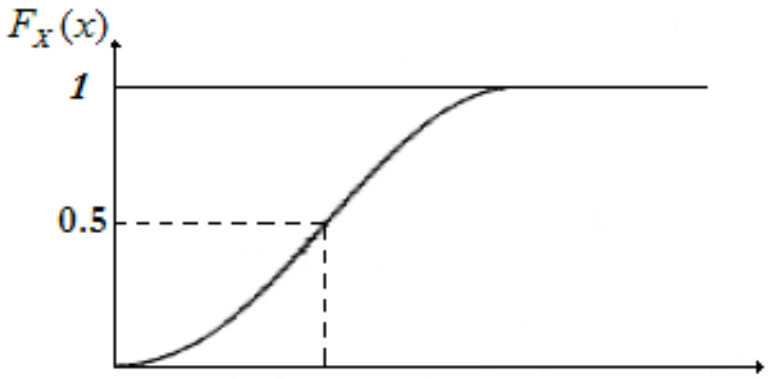
$f_x(\tilde{x}) = \max$



Mediana: $x_{0.5} = \breve{x}$

$F_x(\breve{x}) = 0.50$





Media: $\mu_x = E[x]$

Media geometrica: $M_g = \prod_k x_i k_i$

$\log M_g = E[\log x]$

MISURA DI DISPERSIONE.

Varianza: $\text{var}[x] = E[(x-\mu)^2] = \sigma^2$

Scarto quadratico medio: $\sigma = \sqrt{\text{var}[x]}$

Coefficiente di variazione: $Cv = \frac{\sigma}{\mu} = \gamma$

MISURA DI ASIMMETRIA.

Coefficiente di asimmetria: $\gamma_1 = Ca = \frac{\mu_3'}{\sigma^3}$

MISURA DI APPIATTIMENTO O CURTOSI.

Curtosi: $k = \frac{\mu_4'}{\sigma^4}$

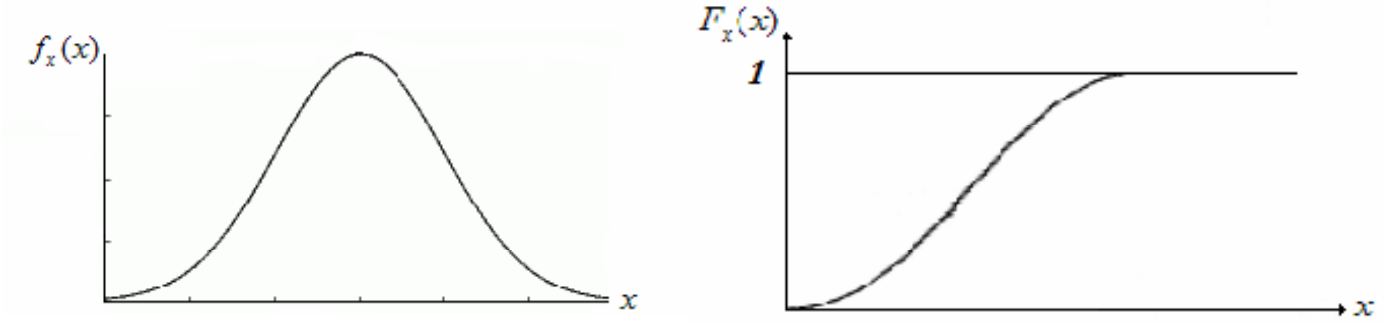
Coefficiente di eccesso o di Curtosi: $\gamma_2 = k - 3 = \frac{\mu_4'}{\sigma^4} - 3$

DISTRIBUZIONE NORMALE DEL CASO O DI GAUSS

Funzione densità di probabilità: $$f(x) = \frac{1}{\theta_2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2} \qquad -\infty < x < +\infty$$

Funzione di distribuzione cumulata: $$F(x) = \frac{1}{\theta_2\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2} dx$$

può essere calcolata numericamente per ogni θ_1 e θ_2.



I parametri θ_1 e θ_2 sono dati da:

$$E[x] = \theta_1 \qquad var[x] = \theta_2^2$$

DISTRIBUZIONE NORMALE IN FORMA CANONICA

Variabile normale standardizzata o ridotta $y = \dfrac{(x-\mu)}{\sigma} = \dfrac{(x-\theta_1)}{\theta_2}$

$E[y] = 0$ $\qquad$ $var[y] = 1$

$f(y) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$ $\qquad$ $F(y) = \displaystyle\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy$ $\qquad$ $F(-y) = 1 - F(y)$

Valori notevoli di *y*(F) e di F(*y*)

F(*y*)	*y*
0.025	-1.96
0.50	0.00
0.975	+1.96

y	F(*y*)
-2.0	0.0228
-1.0	0.1587
0.0	0.5000
1.0	0.8413
2.0	0.9772

Esempio: $X \cong N(\theta_1 = 10, \theta_2 = 3)$

Valore di *X* corrispondente a *F*(*x*) = 0.025 ovvero a $T = \dfrac{1}{F(x)} = 40$

$X_{0\,025} = \theta_1 - 1.96\theta_2 = 10 - 1.96 * 3 = 4.12$

DISTRIBUZIONI DERIVATE

Funzione $Y = g(x)$ strettamente monotona crescente e derivabile di una v.a. continua X

Esempio: $Y = log(x)$ oppure $x^{1/2}$ oppure $x^{1/3}$ $x \geq 0$

$$F_Y(y) = P[Y \leq y] = P[g(X) \leq g(x)] = P[X \leq x] = F_X(x)$$

$$dF_Y(y) = dF_X(x) \Rightarrow f_Y(y)dy = f_X(x)dx$$

$$f_Y(y) = f_X(x)\frac{dx}{dy} \qquad \text{ovvero:} \qquad f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$$

Quando si conosce la distribuzione della Y e si ricerca quella della X vale, ovviamente

$$f_X(x) = f_Y(y)\left(\frac{dy}{dx}\right) = f_Y(g(x))\left|\frac{d(g(x))}{dx}\right|$$

Esempio: variabile normale standard $y = \dfrac{(x-\mu)}{\sigma} = \dfrac{(x-\theta_1)}{\theta_2} \Rightarrow x = \theta_1 + y\theta_2$

$$f_Y(y) = \frac{dx}{dy}f_X(\theta_1 + y\theta_2) = \theta_2\frac{1}{\theta_2\sqrt{2\pi}}e^{-\frac{1}{2}\left[\frac{(\theta_1+y\theta_2-\theta_1)}{\theta_2}\right]^2} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}$$

Vale anche: $F_x(x) = P[X \leq x] = P\left[y \leq \left(\dfrac{x-\theta_{1x}}{\theta_{2x}}\right)\right] = F_y(y)$

MEDIA DI UNA VARIABILE FUNZIONE DI UN'ALTRA

Se c è una costante:

$$E[c] = \int_{-\infty}^{+\infty} c f_x(x) dx = c \int_{-\infty}^{+\infty} f_x(x) dx = c$$

Similmente

$$E[cX] = cE[X]$$

$$E[a + bX] = a + bE[X]$$

$$E[g_1(x) + g_2(x)] = E[g_1(x)] + E[g_2(x)]$$

In generale

$$E[g(x)] \neq g(E[g(x)])$$

Esempio:

$$E\left[\frac{1}{X}\right] \neq \frac{1}{E[X]} \qquad\qquad E[X^2] = [E[X]]^2 + \theta_{2x}^2$$

VARIANZA DI UNA VARIABILE FUNZIONE

$$var[x] = E[(x-\theta_1)^2] = E[x^2 - 2\theta_1 x + \theta_1^2] = E[x^2] - 2\theta_1 E[x] + \theta_1^2 = E[x^2] - \theta_1^2$$

$$var[c] = 0$$

$$var[cx] = c^2 var[x]$$

$$var[a+bx] = b^2 var[x]$$

VARIABILE STANDARDIZZATA

$$y = \frac{x-\theta_1}{\theta_2} \qquad E[y] = 0 \qquad var[y] = 1$$

Il confronto tra un campione e la popolazione si può effettuare attraverso la **comparazione delle forme delle curve di distribuzione** cumulata (campionaria e teorica).
Affinchè la comparazione sia coerente, per la distribuzione campionaria si deve usare una **Stima della probabilità cumulata della popolazione**, chiamata **Plotting position**

$$\phi(x_i) = \hat{P}(x_i)$$

Una possibilità valida se non si ha alcuna indicazione sulla distribuzione teorica da usare è:

$$\phi(x_i) = \frac{i}{n+1} \qquad \phi(x_i) = \mu(P(x_i))$$

detta *Weibull* Plotting position (è *distribution free*).

Corrisponde a porre α=0 nella relazione più generale:

$$\phi(x_i) = \frac{i-\alpha}{n+1-2\alpha}$$

- Distribution dependent $\phi(x_i) = P(\mu(x_i))$

$$\phi(x_i) = \frac{i-\alpha}{n+1-2\alpha}$$

Si hanno ad esempio:

- Distribuzioni debolmente asimmetriche (*Cunnane*)

$$\phi(x_i) = \frac{i-0.4}{n+0.2} \qquad (\alpha = 0.4)$$

- Distribuzioni debolmente asimmetriche (*Gringorten*)

$$\phi(x_i) = \frac{i-0.44}{n+0.12} \qquad (\alpha = 0.44)$$

- Distribuzioni fortemente asimmetriche (*Hazen*)

$$\phi(x_i) = \frac{i-0.5}{n} \qquad (\alpha = 0.5)$$

Analogamente, le stime dei momenti della popolazione richiedono alcune correzioni sulle espressioni dei momenti campionari:

per la varianza: $\widehat{var} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

per l'asimmetria $\widehat{Ca} = \frac{n}{(n-1)(n-2)}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\widehat{var}^{3/2}}$