

Frattili di una distribuzione

Una distribuzione può essere descritta per mezzo dei suoi **frattili**.

Si dice frattile (sinonimi: **centile, percentile e quantile**) *p-esimo* di una distribuzione quel valore x_p tale che la frequenza relativa cumulata $F(x_p) = p$.

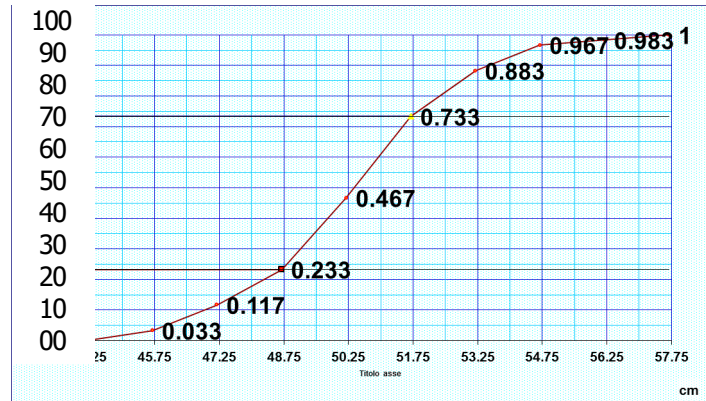
Ad esempio, il 50° centile di una distribuzione è il valore che, sull'asse dei numeri reali, ha alla sua sinistra il 50% dei valori della distribuzione, e **coincide con la mediana**.

Il 10° centile è il valore che ha alla sinistra il 10% della distribuzione.

ALTEZZA(cm)
di un campione
di 60 neonati.

limiti di classe	x_i	$f(x_j)$	$x_i f(x_j)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
Σ		60	3022.5

Nei **grafici cumulati**, i valori riportati sull'asse verticale indicano la **frequenza** delle rilevazioni con **valore pari o minore** ai valori in corrispondenza sull'asse orizzontale



La Moda

Più di rado si incontra una terza misura di posizione, la moda; è il *valore che si verifica più spesso (frequenza assoluta più elevata)*; la modalità della variabile in cui si registra il maggior numero di casi.

Quanto sono usualmente lunghi i bimbi alla nascita?

Guardando i dati a nostra disposizione, è subito evidente maggior numero (16) di bimbi è lungo tra i 50.3 cm e i 51.7 cm.

la classe modale è dunque 50.25-51.75.

Se la distribuzione ha più di due valori massimi o se la frequenza più alta riscontrata nell'insieme considerato non supera di molto le altre la moda non è un buon indicatore di tendenza centrale.

La moda

Lunghezza supina (cm) in un campione di 60 neonati.
Valori ottenuti con l'infantometro Harpenden.

Estremi di classe	Valore Centrale	Freq Semplici		Freq cumulate	
		n	%	n	%
44.3-45.7	45.0	2	0.033333	2	0.033333
45.8-47.2	46.5	5	0.083333	7	0.116667
47.3-48.7	48.0	7	0.116667	14	0.233333
48.8-50.2	49.5	14	0.233333	28	0.466667
50.3-51.7	51.0	16	0.266667	44	0.733333
51.8-53.2	52.5	9	0.15	53	0.883333
53.3-54.7	54.0	5	0.083333	58	0.966667
54.8-56.2	55.5	1	0.016667	59	0.983333
56.3-57.7	57.0	1	0.016667	60	1

Nella classe 50.3-51.7 , piu' vicino alla casse con freq=14

$$50,25 + \frac{1,5 \times |16 - 14|}{|16 - 14| + |16 - 9|} = 50,583$$

quale misura di posizione usare?

A quale misura di tendenza centrale ci riferiamo?

- Il proprietario di una ditta afferma "Lo stipendio medio nella nostra ditta è 2.700 euro"
 - Il sindacato dei lavoratori dice che "lo stipendio mensile è di 1.700 euro".
 - L'agente delle tasse dice che "lo stipendio è stato quasi sempre di 2.200 euro".
- Queste risposte diverse sono state ottenute tutte dai dati della seguente tabella.

Media aritmetica= lire 2.700
Mediana = lire 2.200
Moda = lire 1.700

Stipendio mensile	N° di lavoratori
1.300	2
1.700	22
2.200	19
2.600	3
6.500	2
9.400	1
23.000	1

interpretazione delle misure di posizione

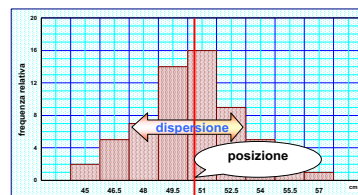
- La **media aritmetica** indica che, se il denaro fosse distribuito in modo che ciascuno ricevesse la stessa somma, ciascun dipendente avrebbe avuto 2.700 euro
- La **moda** ci dice che la paga mensile più comune è di 1.700.euro
- La moda si considera spesso come il valore tipico dell'insieme di dati poiché è quello che si presenta più spesso. **Non tiene però conto degli altri valori** e spesso in un insieme di dati vi è **più di un valore** che corrisponde alla definizione di moda.
- La **mediana** indica che circa metà degli addetti percepiscono meno di 2.200.euro, e metà di più.
- La mediana **non è influenzata dai valori estremi** eventualmente presenti ma solo dal fatto che essi siano sotto o sopra il centro dell'insieme dei dati.



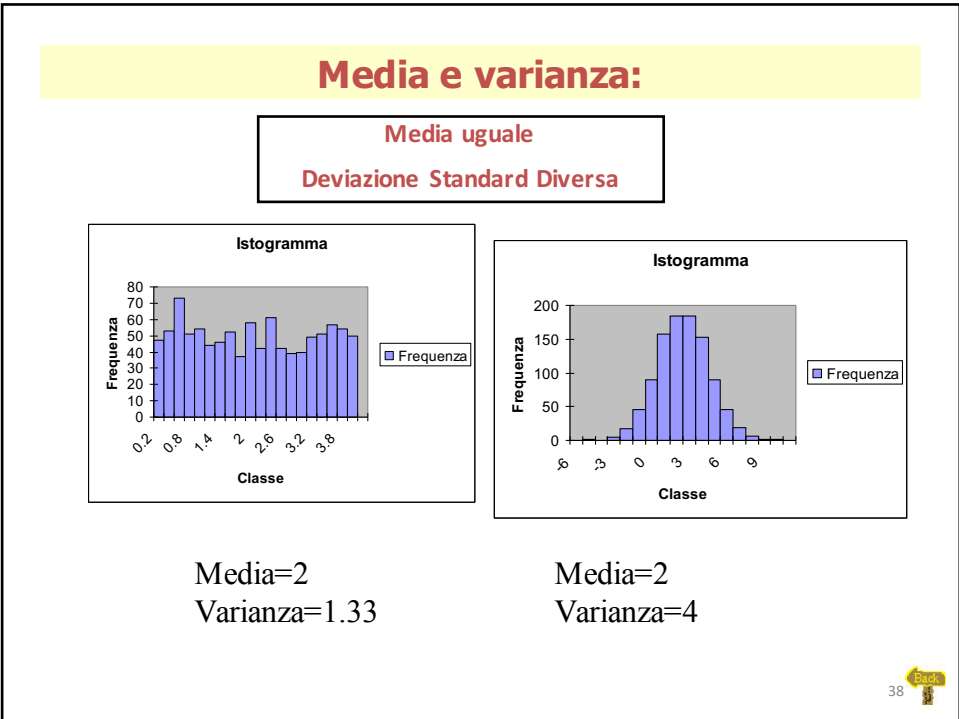
Statistica Descrittiva

- Intervallo di variazione
- Devianza
- Varianza
- Deviazione Standard
- Intervallo interquartile

dispersione di una distribuzione




37



dispersione di una distribuzione

Numero di ore di sonno	frequenza	
	Maschi	Femmine
1	1	3
2	3	6
3	3	7
4	7	8
5	11	5
6	8	3
7	4	1
8	2	1
9	1	1
10	-	-
11	-	1
12	-	1
13	-	1
14	-	1
15	-	1

← Diamo un'occhiata alla distribuzione di frequenza delle ORE DI SONNO indotte da un sonnifero, dormite da **40 maschi** e **40 femmine**.

39 

dispersione di una distribuzione

- ⊕ La **misura della variabilità**, permette di descrivere in modo più completo la distribuzione di una variabile.
- ⊕ Le misure di tendenza centrale: **media, mediana e moda** individuano l'elemento "centrale" della distribuzione.
- ⊕ Diamo, di nuovo, un'occhiata alla distribuzione di frequenza delle **ORE DI SONNO** dei 40 soggetti.
 - ✓ La **media** è di **5 ore** ma uno sguardo alla tabella mostra che **un buon numero di pazienti sono molto diversi tra loro**.
 - ✓ Alcuni presentano un periodo di sonno **più breve** ed altri **più lungo della media**.
- ⊕ La media **non dice** in che misura i dati siano dispersi attorno al valore centrale.

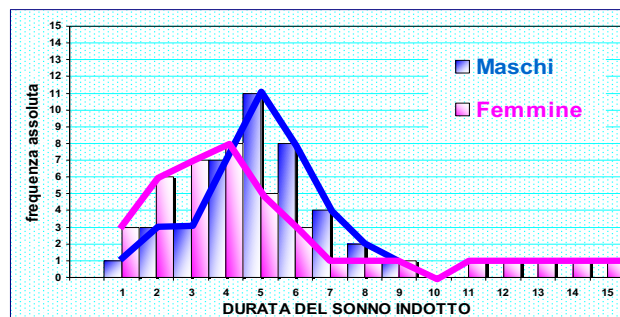
40

dispersione di una distribuzione

Il numero medio di "letture" risulta di 5 ore in entrambe i sessi

Uguale durata del sonno indotto ?

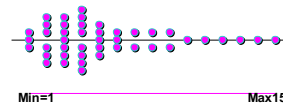
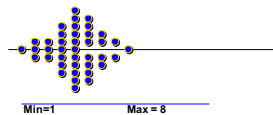
Per facilitare i confronti riportiamo i dati in grafico.



41

L'intervallo di variazione

- ⊕ Mentre **in media** le femmine presentano un durata del sonno uguale ai maschi, alcune di loro hanno un durata del sonno ancora superiore ai tempi più elevati dei maschi.
- ⊕ Quindi le medie non sono insufficienti: per completare il quadro occorrono alcune misure di variabilità.
- ⊕ L'**intervallo di variazione** o range consiste semplicemente nella differenza tra il valore massimo e il valore minimo della distribuzione.



42

L'intervallo di variazione

Esempio:

Gli insiemi di valori di VES {A}: { 8, 5, 7, 6, 35, 5, 4} hanno la stessa
 {B}: { 11, 8, 10, 9, 17, 8, 7} media ($\bar{x}=10$),

ma in {A} i valori sono più dispersi che in {B}:

in {A} i valori sono inclusi tra 4 e 35

in {B} i valori sono inclusi tra 7 e 17

La differenza tra il massimo e il minimo valore di un insieme di dati è detto **intervallo di variazione** (o **range**).

il **range** di {A} è $R_A = 35 - 4 = 31$

il **range** di {B} è $R_B = 17 - 7 = 10$

Il **range** è il più **intuitivo** fra gli indici di dispersione, ha però il difetto di basarsi solo sui due valori estremi, nei quali si manifesta maggiormente la variabilità di campionamento e l'errore di misura.

43

La devianza

Gli indici di dispersione di più largo uso sono basati sugli **scarti dalla media**: per un campione di dimensione n , $\{x_1, x_2, \dots, x_n\}$, sono così definiti

Devianza: $D = \sum (x_i - \bar{x})^2$

Varianza campionaria: $s^2 = \frac{D}{n-1}$

Deviazione standard: $s = \sqrt{s^2}$

Coefficiente di variazione: $CV\% = 100 \times \frac{s}{\bar{x}}$

La **devianza** è la somma dei quadrati degli scarti tra ogni elemento del campione (x_i) e la media campionaria (\bar{x}).

44

formule di calcolo della devianza

devianza per dati singoli

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



$$= \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

devianza per dati raggruppati in classi

$$D = \sum (x_i - \bar{x})^2 f(x_i)$$



$$= \sum x_i^2 f(x_i) - \frac{(\sum x_i f(x_i))^2}{\sum f(x_i)}$$

45

calcolo degli indici di dispersione

- Nell'**esempio** dei due insiemi di valori di VES si ha:

{A}: { 8, 5, 7, 6, 35, 5, 4}

$$D = 8^2+5^2+\dots 4^2 - (8+5+\dots 4)^2/7 = 1440-700=740$$

$$s^2 = 740/6 = 123.33 \quad s = \sqrt{123.3} = 11.1 \quad i = \{-1.1, 21.1\}$$

$$CV\% = 100 \times (11.1/10) = 111\%$$

{B}: { 11, 8, 10, 9, 17, 8, 7}

$$D = 11^2+8^2+\dots 7^2 - (11+8+\dots 7)^2/7 = 768-700 = 68$$

$$s^2 = 68 / 6 = 11.33 \quad s = \sqrt{11.33} = 3.4 \quad i = \{6.6, 13.4\}$$

$$CV\% = 100 \times (3.4/10) = 34\%$$

In {A} l'intervallo $\pm s$ include anche valori negativi di VES, che ovviamente non sono possibili. L'uso di s per esprimere la dispersione dovrebbe essere quindi limitato alle **distribuzioni simmetriche** (o quasi).

46

calcolo della devianza (dati in classi)

limiti di classe	x_i	$f(x_i)$	$x_i f(x_i)$
44.25- 45.75	45.0	2	90.0
45.75- 47.25	46.5	5	232.5
47.25- 48.75	48.0	7	336.0
48.75- 50.25	49.5	14	693.0
50.25- 51.75	51.0	16	816.0
51.75- 53.25	52.5	9	472.5
53.25- 54.75	54.0	5	270.0
54.75- 56.25	55.5	1	55.5
56.25- 57.75	57.0	1	57.0
Σ		60	3022.5

47

calcolo della devianza (dati in classi) _5di5

Nell'esempio della lunghezza dei neonati:

x_i	$f(x_i)$	$x_i \cdot f(x_i)$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot f(x_i)$	x_i^2	$x_i^2 \cdot f(x_i)$
45.0	2	90.0	-5.375	28.891	57.781	2025.00	4050.00
46.5	5	232.5	-3.875	15.016	75.078	2162.25	10811.25
48.0	7	336.0	-2.375	5.641	39.484	2304.00	16128.00
49.5	14	693.0	-0.875	0.766	10.719	2450.25	34303.50
51.0	16	816.0	0.625	0.391	6.250	2601.00	41616.00
52.5	9	472.5	2.125	4.516	40.641	2756.25	24806.25
54.0	5	270.0	3.625	13.141	65.703	2916.00	14580.00
55.5	1	55.5	5.125	26.266	26.266	3080.25	3080.25
57.0	1	57.0	6.625	43.890	43.890	3249.00	3249.00
	60	3022.5			365.812		152624.25

$$\text{media} = 3022.5 / 60 = 50.375$$

$$D = (45.0 - 50.375)^2 \cdot 2 + (46.5 - 50.375)^2 \cdot 5 + \dots + (57.0 - 50.375)^2 \cdot 1 = 365.812$$

$$D = 152624.25 - (3022.5)^2 / 60 = 152624.25 - 152258.44 = 365.813$$

$$\text{Var} = 365.812 / 59 = 6.2$$

$$\text{Deviazione standard} = 2.49$$



calcolo della varianza (dati in classi)

x_i	$f(x_i)$	x_i^2	$x_i \cdot f(x_i)$	$x_i^2 \cdot f(x_i)$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2 \cdot f(x_i)$
1	4	1	4	4	-4	64
2	9	4	18	36	-3	81
3	10	9	30	90	-2	40
4	15	16	60	240	-1	15
5	16	25	80	400	0	0
6	11	36	66	396	1	11
7	5	49	35	245	2	20
8	3	64	24	192	3	27
9	2	81	18	162	4	32
10	0	100	0	0	5	0
11	1	121	11	121	6	36
12	1	144	12	144	7	49
13	1	169	13	169	8	64
14	1	196	14	196	9	81
15	1	225	15	225	10	100
Σ	80		400	2620		620

$$\text{Devianza} = 620; \text{Varianza} = \text{Devianza} / (N-1) = 620 / 79 = 41.33$$

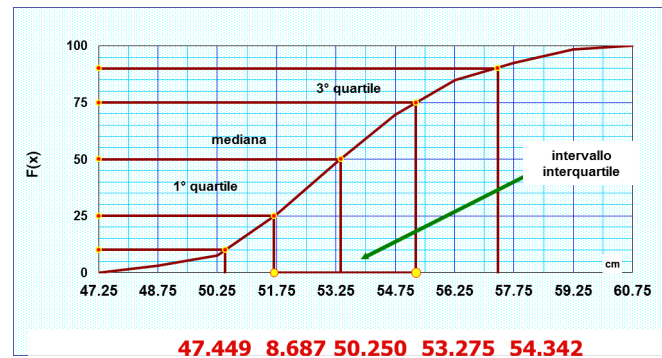
Torniamo all'esempio delle ORE
DI SONNO

$$\text{Deviazione standard} = 6.429$$



l'intervallo interquartile

Un indice di dispersione di uso comune è l'**intervallo interquartile**, dato dalla **differenza tra 3° e 1° quartile** (cioè tra 75° e 25° centile): tale intervallo contiene la metà dei valori inclusi nel campione, indipendentemente dalla forma della distribuzione della variabile.



Indici di dispersione:

$x_{max} - x_{min}$ Range (intervallo di variazione)

$\frac{1}{n} \sum_{i=1}^n |x_i - \mu|$ Scarto medio assoluto

$\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$ Media dei quadrati degli scarti

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

$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ Deviazione standard campionaria

p-esimo quantile: si considera np per $[0 \leq p \leq 1]$
 Se np non è intero, considero k l'intero successivo e il p -esimo quantile è x_k
 Se $np = k$ è intero, il p -esimo quantile è $(x_k + x_{k+1})/2$

Q_1 = primo quartile = 25° percentile
 Q_2 = secondo quartile = 50° percentile = mediana
 Q_3 = terzo quartile = 75° percentile

Principali indici statistici

I grafici finora analizzati ci danno informazioni qualitative; possiamo quantificarle ricorrendo ai seguenti indici.

Siano x_1, x_2, \dots, x_n n osservazioni numeriche

52 

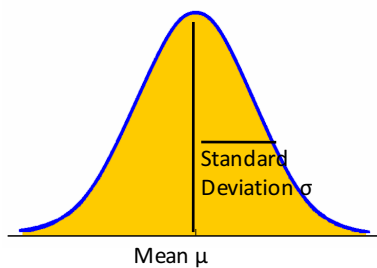
La distribuzione normale



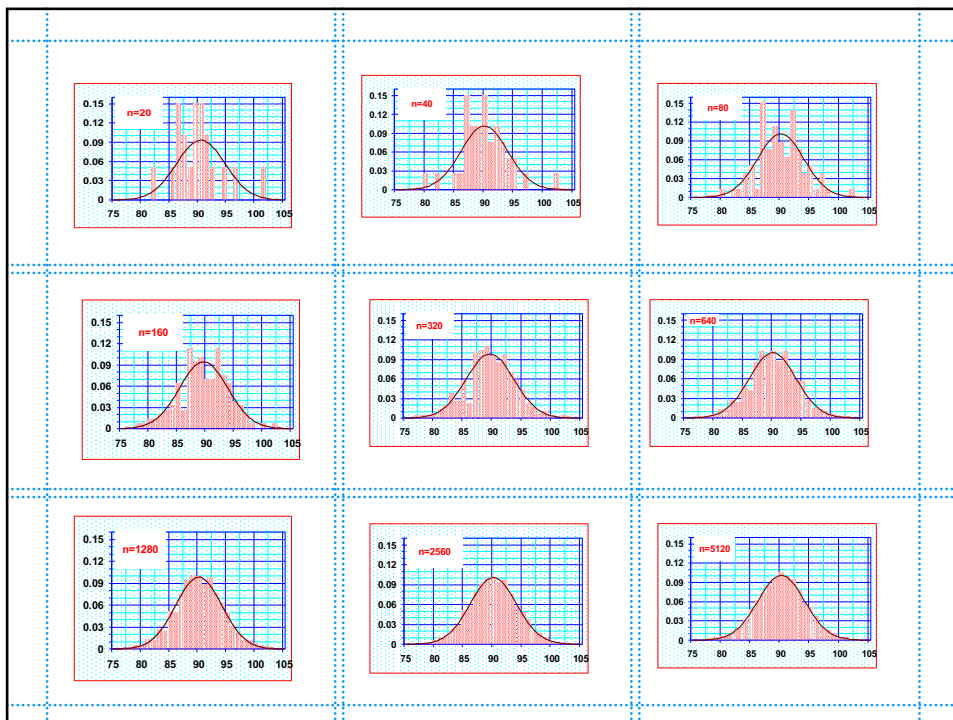
Johann Carl Friedrich Gauss
(1777-1855)

LA FORMA DELLA DISTRIBUZIONE DEGLI ERRORI DI MISURA

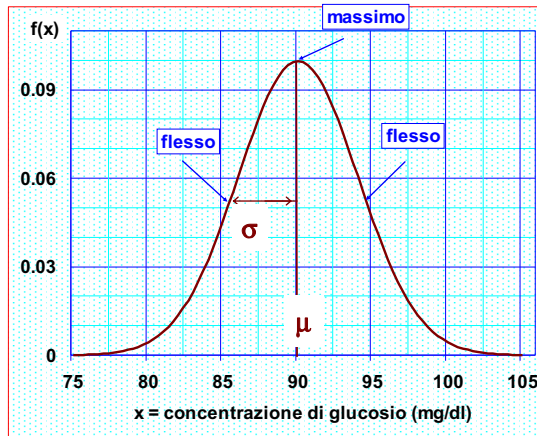
All'aumentare del numero di misure, i valori tendono ad accentrarsi attorno alla loro media e l'istogramma assume una forma **a campana** sempre più regolare, che può essere approssimata con una funzione reale nota come **funzione di gauss/ funzione normale**.



Johann Carl Friedrich Gauss
(1777-1855)



La funzione di Gauss



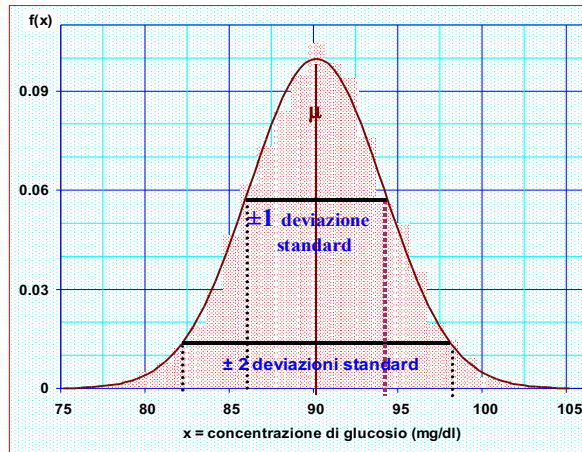
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

dove: σ è la deviazione standard della totalità delle misure;
 μ è la media della totalità delle misure;

La funzione di Gauss

- Gli errori casuali di misura, considerati nel loro complesso, mostrano un comportamento tipico che può essere così descritto:
- Gli **errori piccoli** sono più frequenti di quelli **grandi**;
- Gli errori di **segno negativo** tendono a manifestarsi con la stessa frequenza di quelli con segno positivo;
- All'aumentare del numero delle misure si ha che circa **2/3** dei valori tendono ad essere inclusi nell'intervallo **media +/- 1 deviazione standard**
- Il **95%** dei valori tende ad essere incluso nell'intervallo **media +/- 2 deviazioni standard**

La funzione di Gauss



Rilevanza della distribuzione Normale

Può essere utile per descrivere molti fenomeni

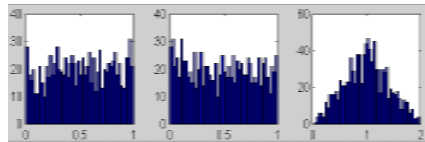
Molte distribuzioni discrete possono essere approssimate con una distribuzione normale al crescere del numero di elementi

Molte distribuzioni continue possono essere trasformate in distribuzioni normali

Gli errori di una misura si distribuiscono attorno ad un valore medio seguendo una legge di questo tipo

Teorema del limite centrale

- TLC: la distribuzione della somma di variabili aleatorie indipendenti e identicamente distribuite (iid) tende ad una gaussiana.

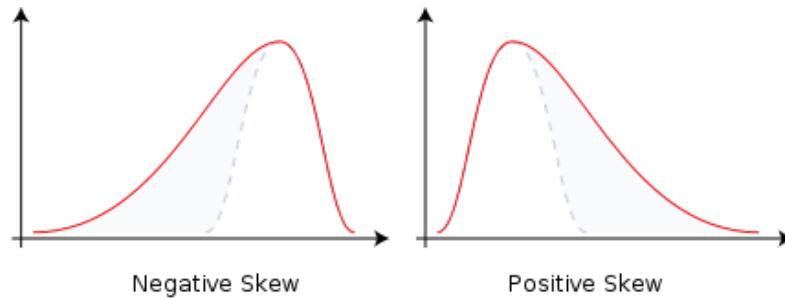


- i.i.d.= se ogni variabile ha la stessa distribuzione di probabilità delle altre variabili, e sono tutte statisticamente indipendenti.
- Tale ipotesi può essere rilassata se le varianze delle singole variabili sono diverse da zero, e se i valori delle variabili sono superiormente limitati

Quali sono I migliori descrittori statistici per un campione?

**Dati estratti da pdf Gaussiana:
Media +/- Deviazione Std**

Skewness

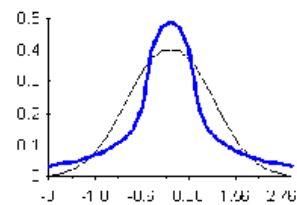


62

CURTOSI: leptocurtica

Curtosi: distribuzione leptocurtica

Distribuzione **leptocurtica**:
una frequenza maggiore
dei valori estremi e dei
valori centrali.



CURTOSI:

distribuzione platicurtica

Curtosi

Distribuzione **platicurtica**:
 frequenza minore dei
 valori estremi e dei valori
 centrali.

Introduzione alla teoria della probabilità JTP 6-133 64

Indici di forma

Skewness

INDICE DI ASIMMETRIA

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} = \frac{\sum (x_i - \mu)^3}{n\sigma^3}$$

>0 coda a destra
 <0 coda a sinistra
 =0 simmetrica

Per la distribuzione gaussiana $\gamma=0$

Kurtosis

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}$$

Misura quanto la distribuzione è appuntita
 >3 poco appuntita
 =3 caso della distribuzione normale
 <3 molto appuntita

Per la distribuzione
 gaussiana $g_2=3$

65

Coefficienti di skewness di Pearson

Karl Pearson ha suggerito i calcoli più semplici come una misura di asimmetria:

La modalità di asimmetria di Pearson, definito da

- $(\text{media} - \text{Moda}) / \text{deviazione standard}$,

Asimmetria primo coefficiente di Pearson, definita da

- $3(\text{media} - \text{moda}) / \text{deviazione standard}$,

Asimmetria secondo coefficiente di Pearson, definito da

- $3(\text{media} - \text{mediana}) / \text{deviazione standard}$.

66

Quali sono i migliori descrittori statistici per un campione?

**Dati estratti da pdf Gaussiana:
Media +/- Deviazione Std**

**Dati estratti da pdf Non-Gaussiana:
[Mediana +/- Int. Interq, 3° misura]
(range, skewness, kurtosis, etc)**

Applicazioni

- La simmetria ha benefici in molti settori. In molti modelli è semplicistico supporre che i dati abbiano una distribuzione [normale] simmetrica intorno alla media.
- La distribuzione normale ha una asimmetria di zero. Ma in realtà, spesso i punti dati non sono perfettamente simmetrici.
- La comprensione dell'asimmetria della serie di dati reali indica che le deviazioni dalla media stanno più nel verso positivo o più nel verso negativo.
- Il test K^2 (D'Agostino) è un Goodness-of-fit test di normalità basato sulla asimmetria e curtosi campionaria.

68

Indici: Schema riassuntivo

di posizione	<ul style="list-style-type: none"> • media: $\bar{x} = \frac{\sum_i x_i}{N}$ • moda: punto di max della distribuzione • mediana: valore sotto al quale cadono la metà dei valori campionari. Si dispongono i dati in ordine crescente e si prende quello che occupa la posizione centrale (N dispari) o la media dei 2 valori in posizione centrale (N pari)
di dispersione	<ul style="list-style-type: none"> • varianza $s^2 = \frac{\sum_i (x_i - \bar{x})^2}{N - 1}$ • deviazione standard S • range $R = x_{\max} - x_{\min}$
di di forma	<ul style="list-style-type: none"> • skewness (coeff. di asimmetria) $\frac{\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^3}{N}$ • curtosi: misura quanto la distribuzione è appuntita $\frac{\sum_i \left(\frac{x_i - \bar{x}}{\sigma}\right)^4}{N}$ <p>>3 poco appuntita <3 molto appuntita</p>

>0 coda a ds
<0 coda a sin
=0 simmetrica

69