Statistica3 – 28/09/2015

Che cosa interessa realmente al biologo quando ad esempio determina la glicemia in un gruppo di 6 animali?

~~La glicemia di questi 6 animali~~

La glicemia degli animali sani

Le tappe sono essenzialmente 2

1 Raccogliere il campione

2 conoscere le probabili caratteristiche della popolazione a partire dai dati raccolti

Fino a che punto i dati raccolti su un campione permettono di stimare le caratteristiche della popolazione di origine?

Le popolazione dei dati

Lo statistico intende per popolazione una serie di numeri. Quando si preleva un campione di sangue a 6 animali, si centrifuga, si aggiunge al plasma gli opportuni reagenti e si ottiene allo spettrofotometro una lettura di assorbimento che corrisponde ad un valore di glicemia diverso per ciascun animale, si ottengono 6 valori e *statisticamente* si parla di popolazione di numeri.

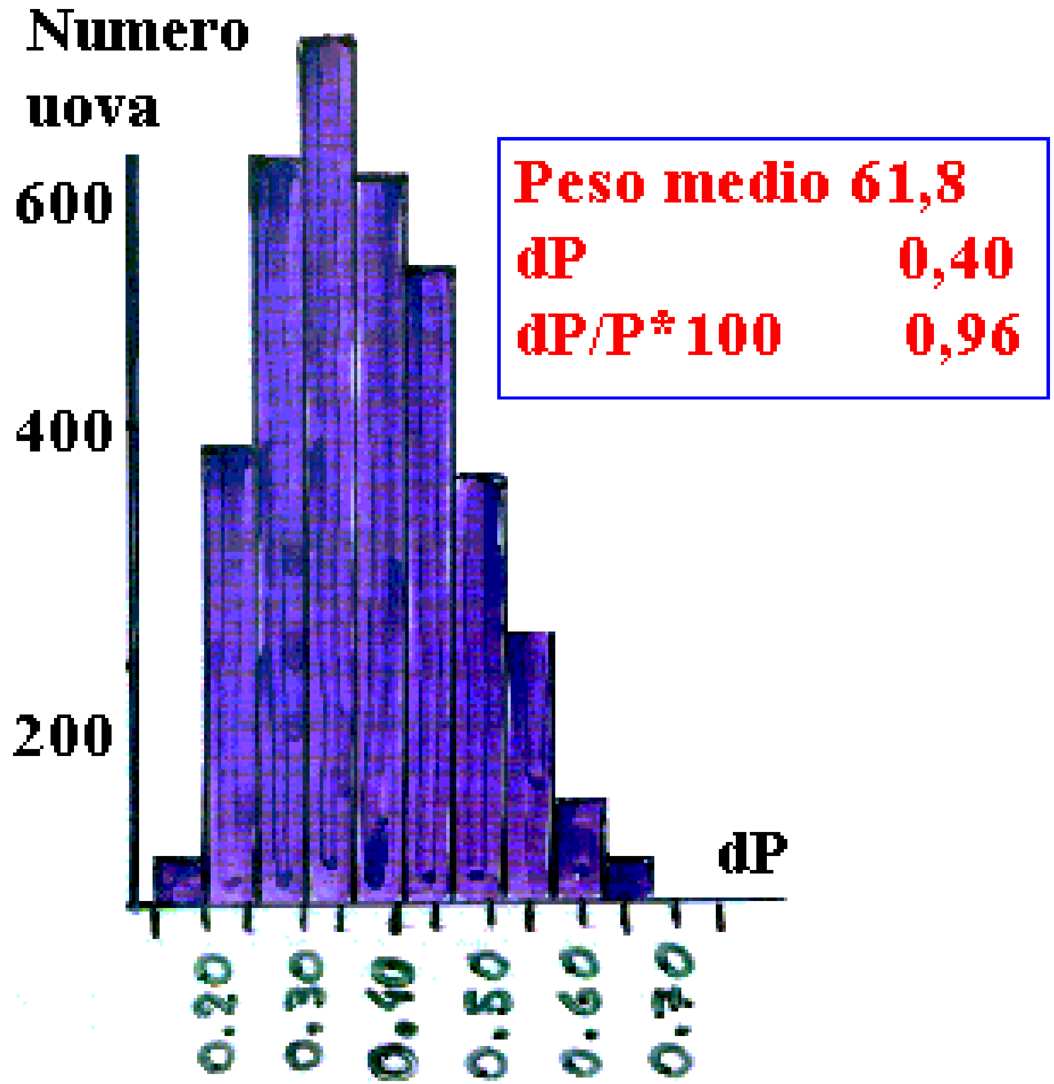
partiamo dalla conoscenza diretta di una popolazione.......

Come può essere rappresentata?

Si dispone in ascissa i valori ed in ordinata il numero delle volte che tali valori vengono rilevati

Istogramma della soluzione di antibiotico che “penetra” nelle uova attraverso i pori del guscio

Trattamenti di disinfezione in Avicoltura



da: Lucidi lezioni Avicoltura prof. Bagliacca unipi.it

Consideriamo ora non più una popolazione ridotta (campione reale) ma una popolazione generalizzata!

popolazione ideale (n= infinito)

un popolazione di dati ideale:

è

infinita

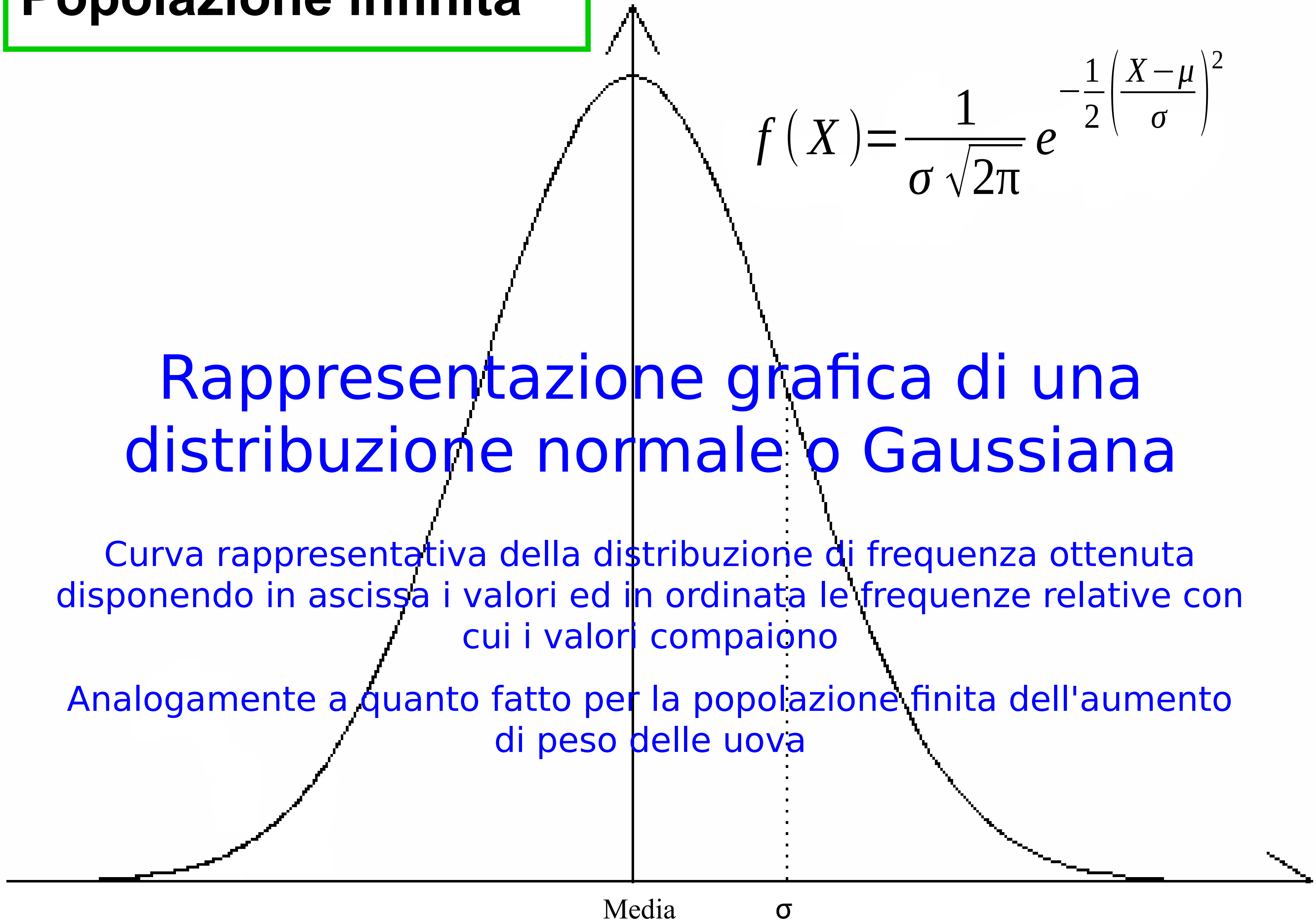
distribuita normalmente

L'esperienza, assai più della dimostrazione formale, ha dimostrato che la maggior parte dei dati di misurazione* nel campo della biologia animale possono considerarsi come estratti da popolazioni infinite distribuite normalmente

* la misurazione, specie se non è "diretta", può essere "distorta" dal "sistema" utilizzato per effettuare la misurazione stessa e quindi la popolazione di dati che si ottiene potrebbe non essere più distribuita normalmente non perché la misura "vera" non sia normalmente distribuita.

Popolazione infinita

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



Rappresentazione grafica di una distribuzione normale o Gaussiana

Curva rappresentativa della distribuzione di frequenza ottenuta disponendo in ascissa i valori ed in ordinata le frequenze relative con cui i valori compaiono

Analogamente a quanto fatto per la popolazione finita dell'aumento di peso delle uova

Equazione matematica che rappresenta la Distribuzione Normale

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

- **dove:**

σ $\equiv$ deviazione standard della popolazione dei dati

π $\equiv$ pi greco = 3,14….

e $\equiv$ base logaritmi naturali = 2,7183….

$(x-\mu)^2$ $\equiv$ scarti dalla media della distribuzione elevati al quadrato;

La media e la deviazione standard di una popolazione di origine è bene siano rappresentati con simboli diversi da quelli dei campioni per distinguerle meglio useremo:

popolazione

deviazione standard σ

media aritmetica μ

La media e la deviazione standard dei campione sono rappresentate con:

campione

media aritmetica $\bar{x}$ = media

deviazione standard d.s. = s = MS

Proprietà della Distribuzione Normale

1 La curva della distribuzione delle frequenze ha la forma di una campana è cioè perfettamente simmetrica quindi media aritmetica, mediana e moda hanno lo stesso valore centrale.

la curva è perfettamente simmetrica all'ordinata massima Y, cioè dove la funzione $f(X)$ raggiunge il suo punto più alto, che è in corrispondenza di $X_i = \mu$; questo fatto comporta che media, mediana e moda coincidano.

2 Il valore che appare con la frequenza massima è la media; i valori più bassi e più alti della media compaiono con minore frequenza e sono tanto meno frequenti quanto più differiscono dalla media.

è crescente per valori della X che vanno da -∞ (meno infinito) a μ (alla media) ed è decrescente per valori che vanno da μ a +∞ (più infinito).

3

Teoricamente la curva è infinita a destra ed a sinistra ma le frequenze diminuiscono in modo assai rapido così che dopo una certa distanza dalla media la frequenza dei dati diviene infinitesimale e può essere considerata trascurabile

la sua funzione di distribuzione *f*(x) è asintotica di x verso -∞ e +∞ (la curva si avvicina all'asse delle ascisse senza mai toccarla); tuttavia per x_i che dista più di 3 σ dalla media, la distanza tra la curva e l'asse delle x è estremamente piccola (tanto piccola da essere un valore trascurabile).

4

La concavità della funzione cambia: vicino alla media è rivolta verso il basso, lontano dalla media verso l'alto.

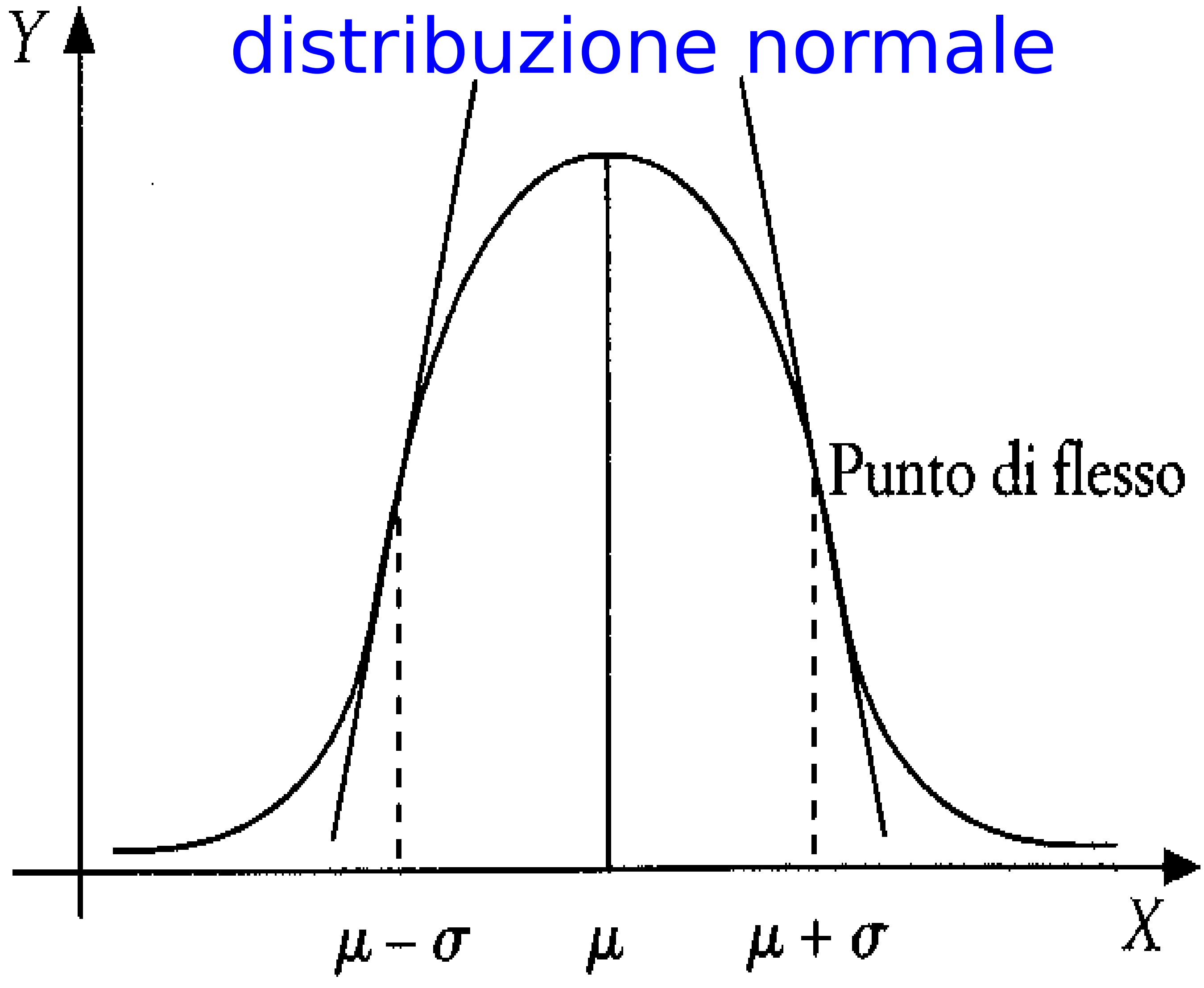
presenta due punti di flesso (punto nel quale la concavità della funzione cambia) **in corrispondenza di** $\mu+\sigma$ **e** $\mu-\sigma$**; cioè i punti in cui la curva da convessa diventa concava si trovano in corrispondenza a ±1** ***deviazione standard*** **dalla media.**

5

Per avere la conoscenza completa di una popolazione distribuita normalmente è sufficiente conoscere due soli valori: la media e la differenza fra le ordinate corrispondenti alla media aritmetica ed al punto di cambiamento di concavità che si chiama **deviazione standard della popolazione σ**

presenta due punti di flesso in corrispondenza di μ+σ e μ-σ; cioè i punti in cui la curva da convessa diventa concava si trovano in corrispondenza a ±1 *deviazione standard* dalla media ed è completamente caratterizzata da questi due parametri μ e σ, oltre che dalle due costanti (*π ed e*).

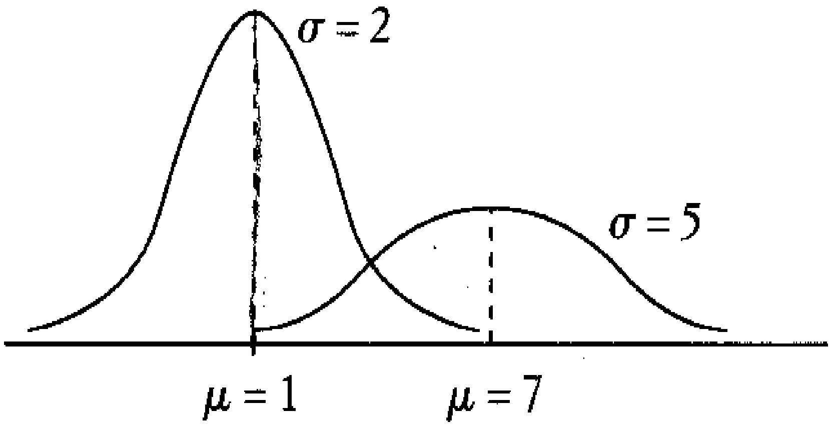
Rappresentazione grafica di una distribuzione normale



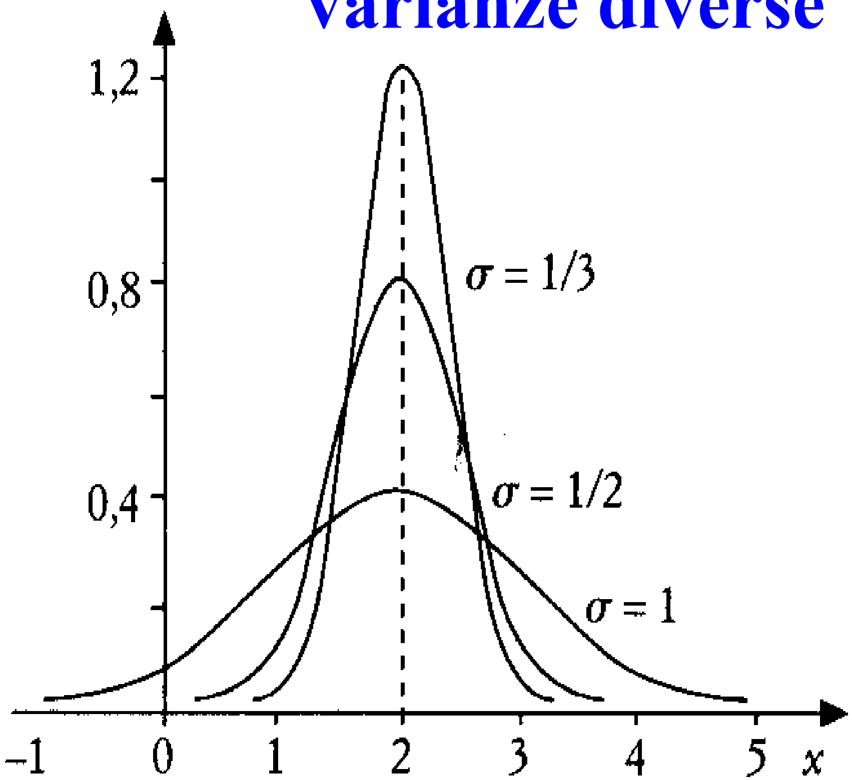
ogni distribuzione normale è univocamente definita dalla **media (μ) e dalla deviazione standard della popolazione (σ).**

Le distribuzioni normali mantengano costanti le "caratteristiche" ma possono differire per la **media (μ)** e la **deviazione standard della popolazione** (**σ**).

Medie diverse e varianze diverse



Stessa media e varianze diverse



Le distribuzioni normali possono differire per la media e la deviazione standard ma mantengano costanti le caratteristiche generali

La probabilità (numero di osservazioni) relativa ad intervalli di valori della funzione normale è così definita:

per un valore di X_i = a, la probabilità dell'intervallo di valori: $-\infty < X < a$ corrisponde all'integrale* seguente:

P(a) = probabilità dell'intervallo di valori tra - ∞ ed a

$$P(a) = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} dX_i$$

* l'integrale definito (fra due punti) corrisponde all'area sottesa dalla curva fra quei due punti.

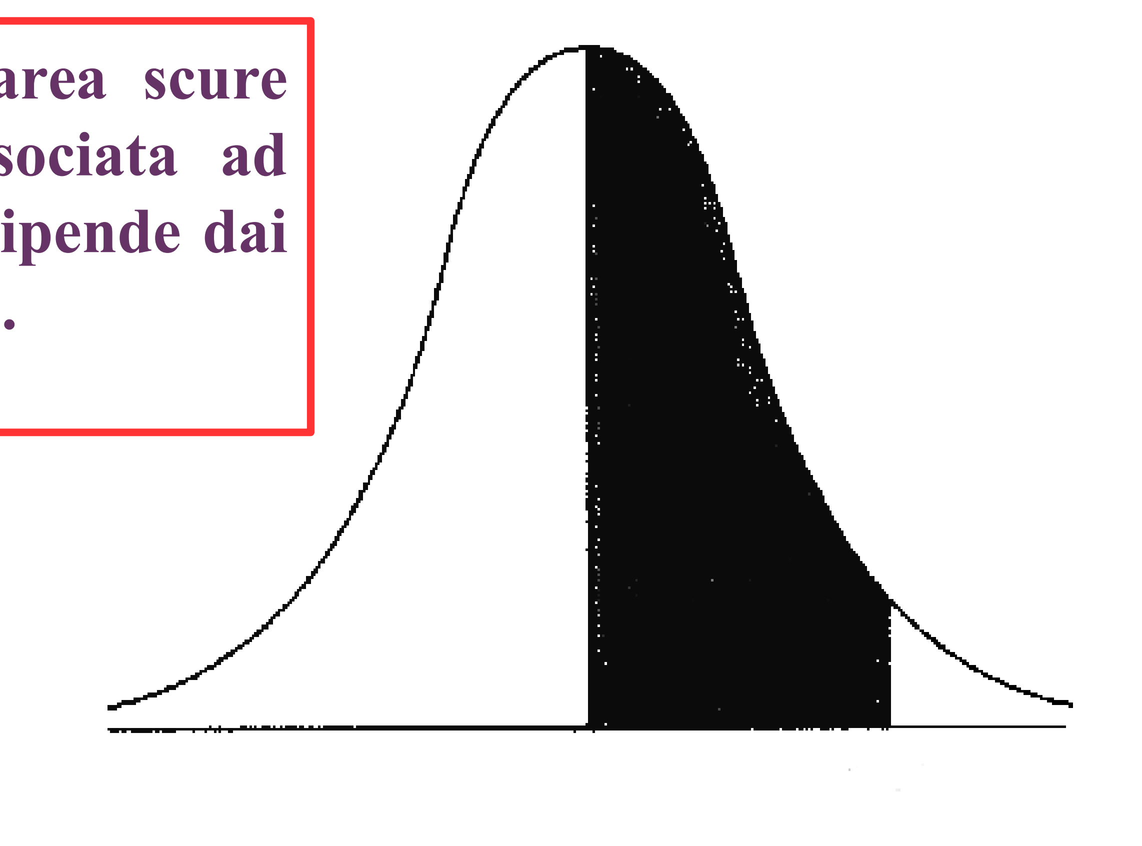
La distribuzione rappresentata dalla relazione precedente viene anche chiamata "*curva degli errori*" perché rappresenta la legge con cui si distribuiscono le variazioni naturali delle popolazioni detti *errori di natura accidentale.*

Il calcolo dell'integrale (a parte la sua complessità) dipende dai soli valori μ e σ^2 (pi greco ed "e" sono costanti = numeri fissi); pertanto si può dire che la probabilità associata ad un intervallo di valori X è funzione dei due soli parametri μ e σ^2.

graficamente:

graficamente:

Il calcolo dell'area scure (probabilità associata ad un intervallo) dipende dai soli valori μ e σ^2.



La conoscenza <u>vera</u> di una legge della biologia comporta avere acquisito la capacità di rappresentarla con una equazione matematiche universalmente applicabile.

La distribuzione normale contiene due parametri, μ e σ^2, che ne rendono difficile il calcolo. Il ricorso alla cosiddetta "***distribuzione standardizzata***" o "ridotta" o "generalizzata" consente invece di individuare le probabilità relative ai diversi intervalli utilizzando le *tavole di probabilità* nelle quali i valori della equazione sono stati calcolati risolvendo gli integrali definiti utilizzando come unità di misura σ.

La *distribuzione normale standardizzata* si ottiene con la trasformazione lineare dei punti grezzi in punti **z**, **utilizzando cioè come unità di misura la σ** :

$$z = \frac{(X - \mu)}{\sigma} \qquad z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

La "*distribuzione standardizzata*" o "ridotta" ha le stesse caratteristiche delle distribuzione normale

La curva normale standardizzata è quindi una curva normale (e tutte le curve normali hanno le stesse caratteristiche) ma caratterizzata da una media = 0 ed una deviazione standard = 1 (inserendo al denominatore **σ** "si misura tutto diviso **σ**")

Le tavole di probabilità della distribuzione normale nel nostro campo vengono utilizzate per calcolare l'area compresa tra due determinati valori della variabile oggetto di studio senza dover calcolare integrali della funzione esponenziale.

Le tavole di probabilità sono quindi relative alla probabilità totale compresa tra due limiti qualunque X_1 e X_2 di una variabile normale che è stata "standardizzata".

Da Tavole probabilità	Probabilità = area sottesa	
$\mu \pm \sigma =$	68,26%	31,74%
$\mu \pm 2\sigma =$	95,44%	4,56%
$\mu \pm 3\sigma =$	99,74%	0,26%
$\mu \pm 1{,}96\sigma =$	95%	5%
$\mu \pm 2{,}575\sigma =$	99%	1%
$\mu \pm 3{,}29\sigma =$	99,9%	0,1%

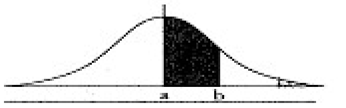
Punteggio standard o Z = la misurazione in unità di deviazioni standard

Le tavola di probabilità riportata nella diapositiva seguente è calcolata fra il punto 0 e + infinito (parte destra della curva).
Poiché la curva è simmetrica I valori della parte sinistra della curva sono esattamente identici e si possono ricavare dalla stessa tavola

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Tabella generata con la funzione distrib.norm.st del foglio di calcolo

=NORMSDIST(X)-0.5



Es:
Z = 1,63
Area = 0,4484

Es:
Area = 0,483
Z = 2,12

Come si usa

Osservando la tavola si troveranno i punti z nella prima colonna a sinistra con una cifra decimale; la seconda cifra decimale è posta nella prima riga in alto della stessa tavola (riga e colonna evidenziate in giallo).

A causa della simmetria della distribuzione queste tavole di solito riportano soltanto i valori delle probabilità comprese fra lo zero e l'ascissa +x, essendo quelle dell'altra metà della curva esattamente uguali.

n.b.

In alcuni libri le tavole possono essere rappresentate diversamente, **leggere le istruzioni per il loro corretto uso ed imparare ad usare le tavole che poi si porteranno per fare l'esame scritto**.

Esempi di uso:

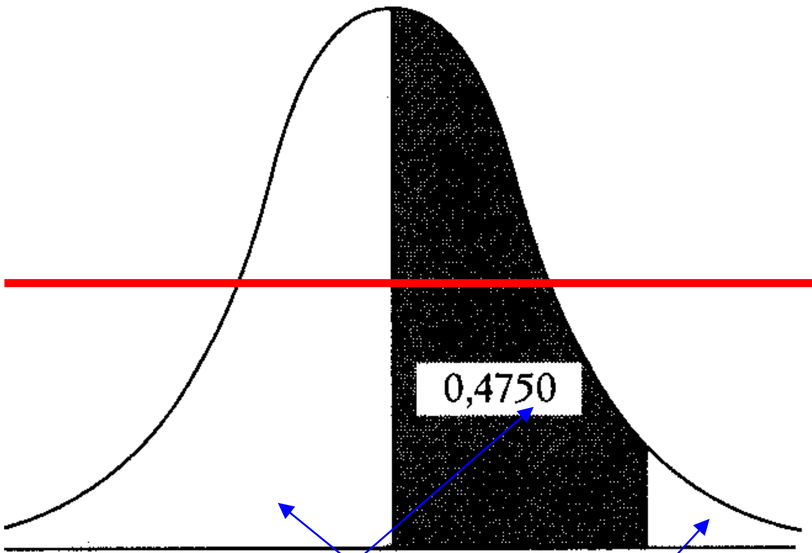
Vogliamo conoscere l'area compresa tra le ordinate corrispondenti a z=0 e z=1,96. Da tabella 0,4750

e tra z= -1,96 e z= +1,96?

0,4750 +

0,4750 =

0,9500 = 95%



	Probabilità	
$\mu + 1{,}96\sigma =$	47,5%	52,5%
$\mu \pm 1{,}96\sigma =$	95%	5%

e tra z=-INFINITO e z=+1,96?

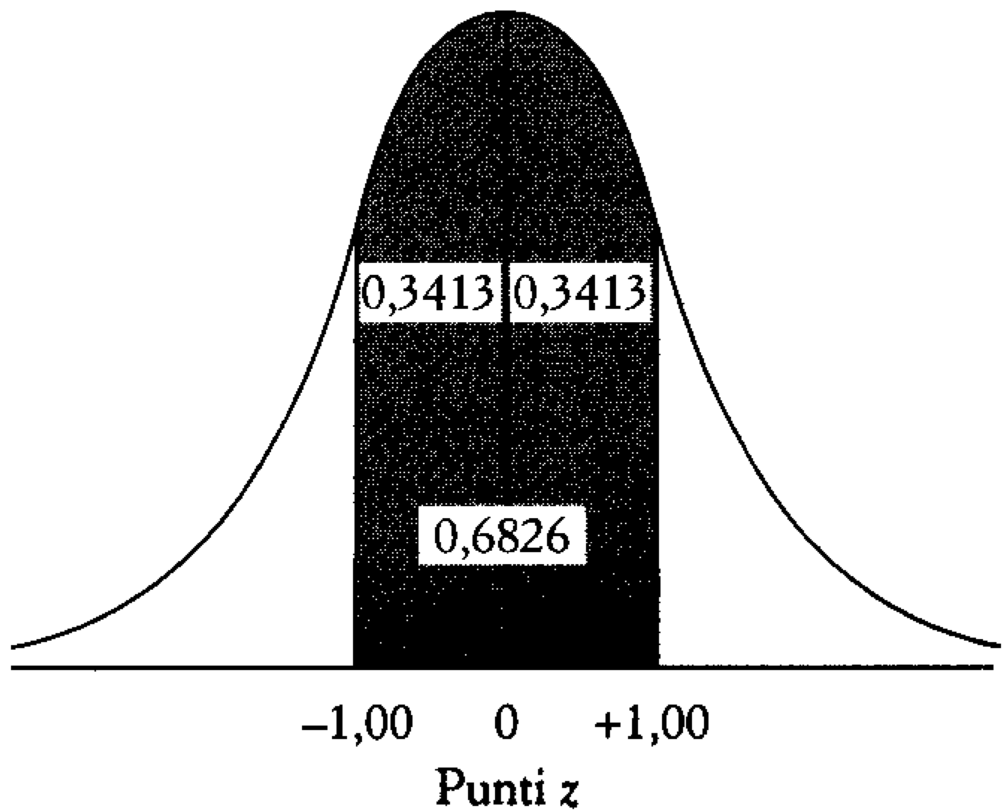
Osservando la colonna dei punti z, si deve scendere fino a trovare z = 1,9 e, rimanendo nella stessa riga fino a trovarsi in quella indicata con 6. Il punteggio che troverete in quel punto indica la porzione di area compresa tra le due ordinate: 0,4750. Poiché l'area totale sotto la curva alla destra dell'ordinata corrispondente a z = 0,00 è 0,500, l'area alla destra dell'ordinata di z = 1,96 sarà: 0,500-0,475=0,025.

Che è lo stesso di: 1- (0,500+0,475)=0,025.

fra lo zero e l'ascissa +X più lo zero e l'ascissa -X = 2 volte il valore trovato (simmetria della distribuzione, altra metà della curva del tutto uguale)

Vogliamo conoscere l'area compresa tra le ordinate corrispondenti a z= -1,0 e z= +1,0.

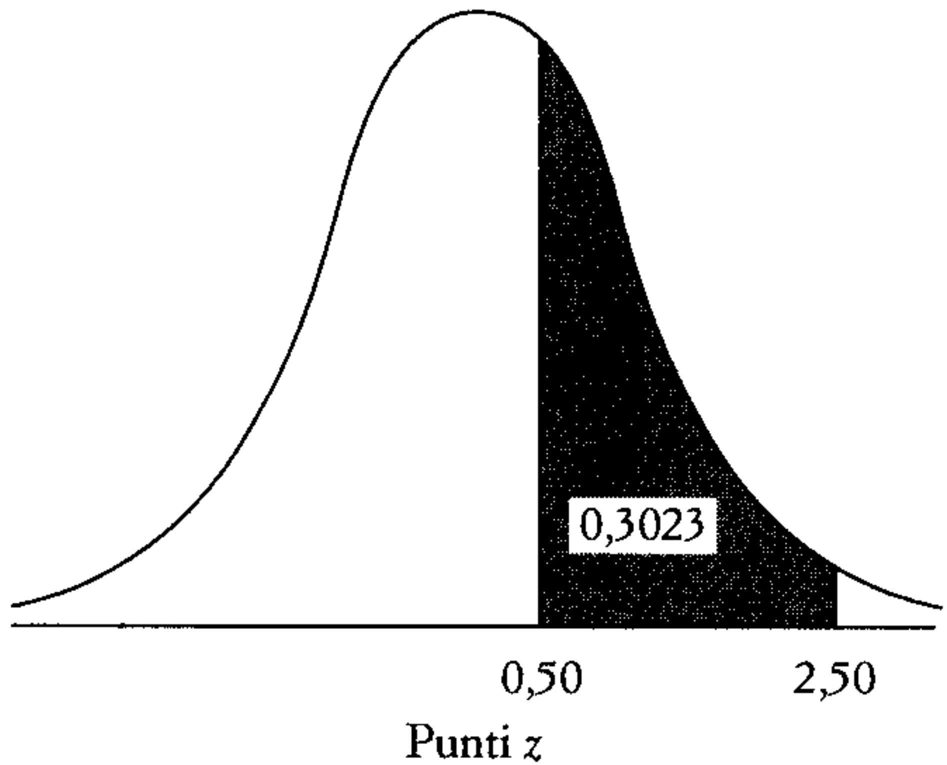
0,3413 +
0,3413 =
0,6826 = 68,26%



$\mu \pm \sigma =$ 68,26% 31,74%

Cercando nella tabella troverete che la porzione di area sotto la curva compresa z=0,00 e z=1,00 è 0,3413. Dalla porzione opposta della curva si troverà ovviamente lo stesso valore, quindi la proporzione di area si otterrà sommando i due valori: 0,3413+0,3413=0,6826 = 68,26%.

Vogliamo ora conoscere l'area compresa tra le ordinate corrispondenti a z = +0,50 e z = +2,50.



Le tavole danno solo le aree a partire dal punto z=0,00!

il calcolo richiede quindi i seguenti passaggi:

- l'area tra le ordinate corrispondenti a z=0,00 e z=0,50 è 0,1915;
- l'area tra z=0,00 e z=2,50 è 0,4938.
- l'area tra z=0,00 e z=2,50 meno l'area tra le z=0,00 e z=0,50 è pari all'area tra z=0,50 e z=2,50;
- Basta allora sottrarre le due precedenti aree: 0,4938-0,1915 = 0,3023. Il punteggio ottenuto è la proporzione di area ricercata.

N.B. Per fare questo tipo di esercizi disegnare sempre la curva della tabella distribuzione normale standardizzata in modo da individuare la porzione di area che deve essere individuata e ricordare che Le tavole danno solo le aree a partire dal punto $z = 0{,}00$

Molto spesso capita di disporre dei dati di una popolazione e voler conoscere le proporzioni di area in percentuale della popolazione.

In tal caso:

per **passare dalla nostra popolazione alla distribuzione normale standardizzata** basta **passare ai valori Z**

per **passare dalle aree alla percentuale** della nostra popolazione basta **moltiplicare le aree trovate per 100.**

Esempio di uso pratico: Un problema di "probabilità":

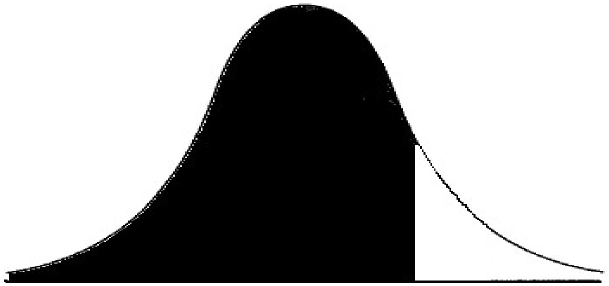
Popolazione animale di peso medio $\mu = 50$ kg e dev.st. Della popolazione $\delta = 15$

Vogliamo sapere che percentuale di animali pesa meno di 65 kg.

da $$z = \frac{(X - \mu)}{\sigma} \qquad \frac{(65 - 50)}{15} = 1$$

Cioè vogliamo conoscere l'area compresa tra le ordinate corrispondenti a:

z = -infinito e z = +1,0 vedi disegno



il calcolo richiede quindi i soliti passaggi:

•l'area tra le ordinate corrispondenti a z = 0,00 e z = 1,00 è 0,3413 (vedi precedente esempio);

•l'area tra z = 0,0 e z = -infinito è 0,5 (l'altra metà).

•Basta allora addizionare le due precedenti aree: 0,3413 + 0,5 = 0,8413. Il punteggio ottenuto moltiplicato per 100 è la proporzione di area ricercata.

• La percentuale della popolazione quindi è:

$$84,13\% = 0,8413 \times 100.$$

Se nella nostra stalla ci sono 500 animali (Stessa popolazione, media e dev. Standard 50 e 15) quanti animali ho a disposizione (posso scegliere) “che pesano più di 65 kg”?

n=	500	
M-pop =	50	
Dev.st.pop=	15	
x = animali che pesano più di	65	
Z=(x-Mpop)/d.s.pop	1	
Z=(x-Mpop)/d.s.pop	**1**	
da tabella Z	0.3413	
aggiungo 0,5	0.8413	
	%	**n**
parte"sinistra"area=	0.8413	421
parte"destra"area =	**0.1587**	**79**

Basta fare il seguente calcolo:

- 500 x 84,13% = 421 (= meno di 65kg)
- 500-421= 79 animali pesano più di 65 kg

Si sono cioè calcolate le caratteristiche di un campione a partire da una popolazione di origine nota (per esempio, quanti soggetti sono disponibili oltre un certo peso che possono essere scelti per la riproduzione).

Se voglio trovare il valore di Z corrispondente al 40% cioè nella tabella **area = 0,4000**?

Cerco 0,400 nella tabella, trovo:

(Tabella Z) 0,3997 = 1,28 e 0,4014 = 1,29.

Il valore di Z sarà fra 1,28 e 1,29, un po più grande di 1,28 ma di quanto devo aumentare 1,28? impostando la proporzione:

(0,4000-0,3997) : (0,4015-0,3997) = X: (1,29-1,28);

Differenza da area tabulata sta all'intervallo fra aree come "di quanto devo aumentare" sta a intervallo di Z

0,0003 : 0,0018 = X : 0,01; x = 0,0003*0,01/0,0018;

$X = 0{,}001\overline{66}$; quindi $Z = 1{,}28 + 0{,}001\overline{66} = 1{,}282$ (arrotondato)

addendum

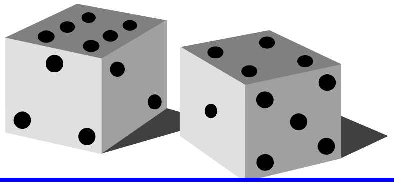
Molte distribuzioni non normali possono essere "adattate" in una distribuzione normale

Se prendiamo campioni di uguale numerosità da una distribuzione non normale la distribuzione delle medie di questi campioni sarà normale (sempre che i campioni siano sufficientemente grandi).

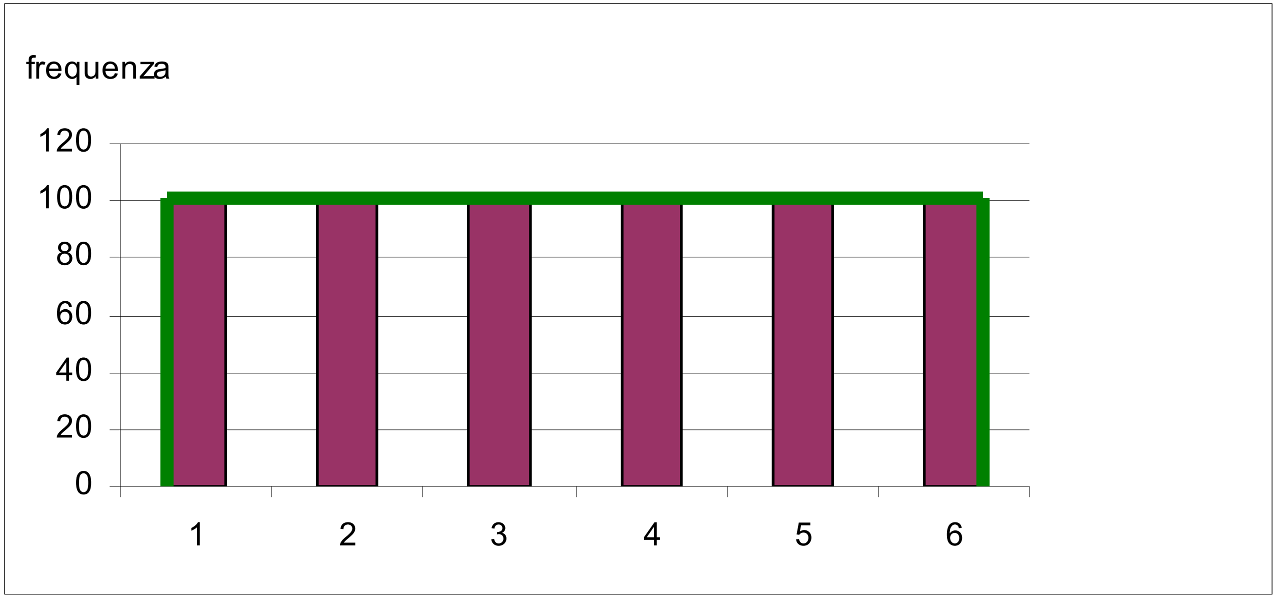
Di solito più di 20-30 ma spesso sono sufficienti 5-6

Tralasciando la dimostrazione matematica .....

Esempio del lancio di un dado



Se lanciamo un dado 600 volte (in realtà dovrebbero essere infinite), ci aspettiamo la seguente distribuzione



La distribuzione (linea verde) ha una forma "rettangolare"

Nessuna faccia appare più spesso di un'altra
(se il dado non è truccato all'infinito … …ogni faccia esce "un sesto")

Se lanciamo un dado 2 volte di seguito per 600 volte e facciamo la media dei due valori trovati ………

2 = 1+1

3 = 1+2 - 2+1

..

6 = 1+5 - 5+1 - 2+4 - 4+2 - 3+3

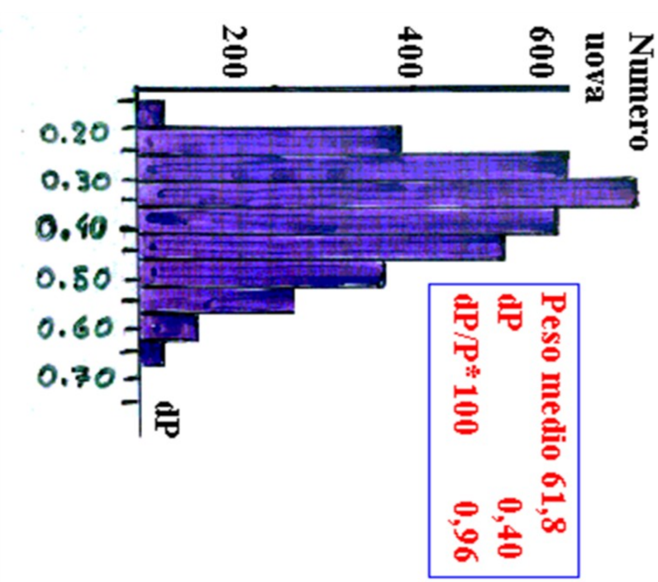
..

12 = 6+6

I numeri “in mezzo” escono più spesso!

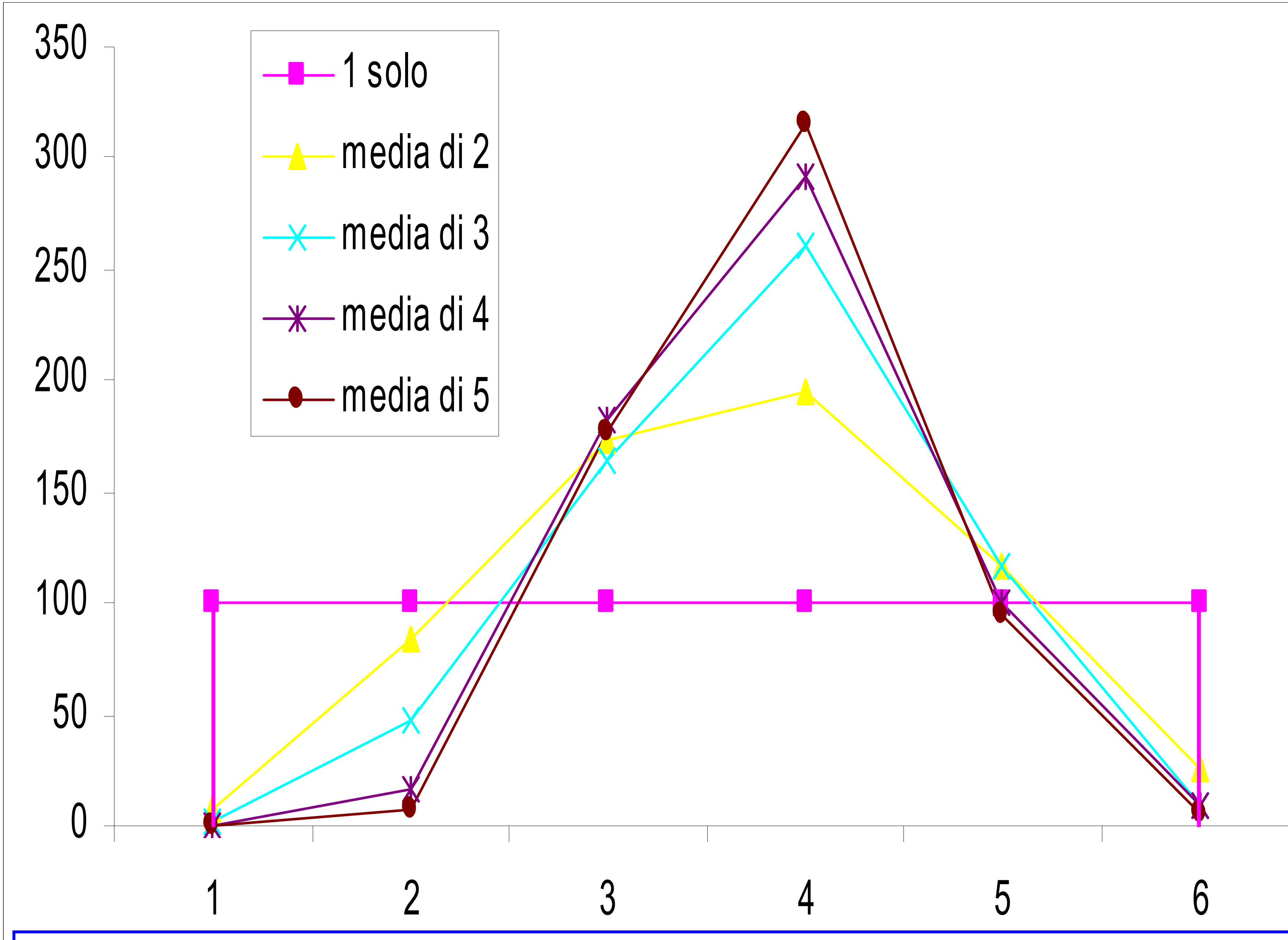
come esercizio calcola tutte le combinazioni di 2 lanci

somma dei due dadi							totale delle volte	media dei due valori					
2	1+1						1	1					
3	1+2	2+1					2	1.5	1.5				
4	1+3	2+2	3+1				3	2	2	2			
5	1+4	2+3	3+2	4+1			4	2.5	2.5	2.5	2.5		
6	1+5	2+4	3+3	4+2	5+1		5	3	3	3	3	3	
7	1+6	2+5	3+4	4+3	5+2	6+1	6	3.5	3.5	3.5	3.5	3.5	3.5
8	2+6	3+5	4+4	5+3	6+2		5	4	4	4	4	4	
9	3+6	4+5	5+4	6+3			4	4.5	4.5	4.5	4.5		
10	4+6	5+5	6+4				3	5	5	5			
11	5+6	6+5					2	5.5	5.5				
12	6+6						1	6					
totale =	6*6			=			36						



Nota già dalla somma di "2" la somiglianza con una distribuzione normale reale, non perfetta!

Simulazione al computer del lancio dei dadi tramite i numeri casuali (fino a 5 volte)



La distribuzione si avvicina sempre di più alla “forma” di una distribuzione normale

Oltre la media di 20 per gli statistici (anche un po' prima per i biologi = a fini pratici) la distribuzione che si ottiene mediando più valori è molto simile “cioè può essere statisticamente assimilata” ad una distribuzione normale (detta anche distribuzione Gaussaiana)!