La distribuzione di Gauss - Approfondimento

Abbiamo già visto come spesso, in presenza di una distribuzione di frequenze relative ad un insieme sufficientemente ampio di dati l'istogramma assume la classica forma a campana. Tale tipologia di distribuzione è chiamata distribuzione normale o distribuzione gaussiana. Tale distribuzione è caratterizzata dalla seguente funzione di densità di probabilità:

$$p(X) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{[-\frac{1}{2}(\frac{X-\bar{X}}{\sigma})^2]}$$

in cui μ è la media dei valori x, σ lo scarto quadratico medio, x il valore di cui si vuol conoscere la probabilità. Tale funzione ha l'interessante proprietà avere l'area totale tra la curva e l'asse delle x calcolata da $-\infty$ a $+\infty$ pari ad 1. (Il motivo della frazione anteposta alla funzione esponenziale è proporio quello di normalizzare la curva per ottenere l'area pari ad 1).

Variabili standardizzate

Siccome ogni singola analisi statistica produce una curva diversa e siccome il calcolo dell'area sottesa alla curva non è semplicissimo e coinvolge l'utilizzo degli integrali su limiti infiniti per utilizzare le informazioni della curva di Gauss è più agevole utilizzare la curva di Gauss standardizzata e le relative tabelle precalcolate da cui si può dedurre rapidamente l'area sottesa ad un determionato settore della curva per poi risolvere vari problemi di statistica. L'equazione della funzione di Gauss standardizzata è la seguente:

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

in cui il valore di x non è più un valore che appartiene alla distribuzione ma è un valore standardizzato che si ottiene mediante la seguente trasformazione:

$$z = \frac{X - \bar{X}}{\sigma}$$

Come interpretazione grafica della standardizzazione della curva si può dire che si è proceduto nel centrare il grafico della curva in modo che il valore corrispondente alla media coincida con $x = 0$ (traslazione orizzontale della curva con $X - \bar{X}$) e successiva regolazione dell'ampiezza della campana con il fattore $\frac{1}{\sigma}$.

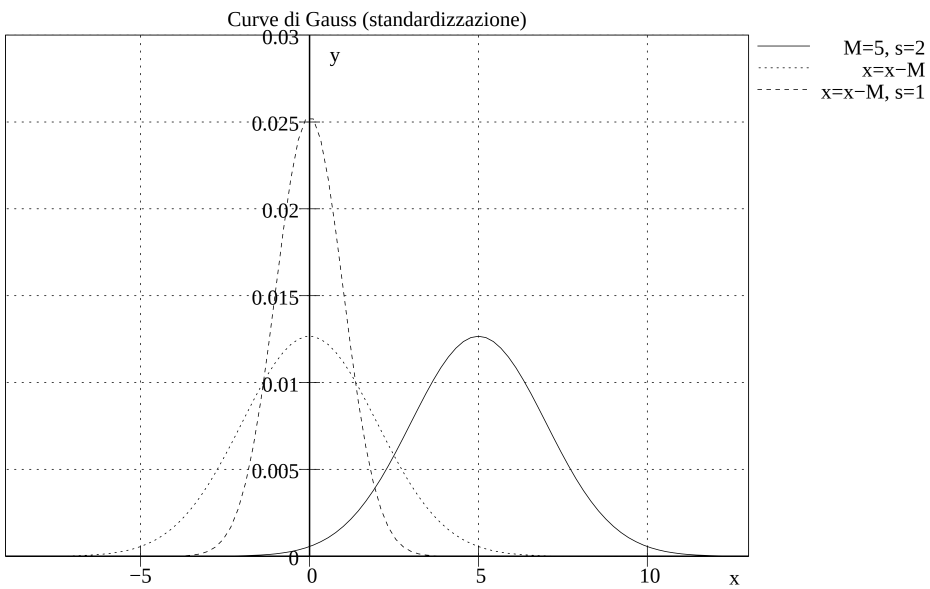


Tabella 1. Standardizzazione

Questo valore standardizzato viene usato sia per valutare le aree sotto la curva di Gauss sia per paragonare risultati statistici che hanno medie e scarti quadratici diferenti tra loro.

Comparazioni tra statistiche differenti

Esempio 1. Ad un esame finale di matematica la media dei voti è stata 72 e lo scarto quadratico medio 15; nella stessa sessione di esami la media dell'esame di fisica è stato 80 con uno scarto quadratico medio di 8. Se uno studente ha ottenuto il punteggio di 80 all'esame di matematica e di 84 a fisica, in quale dei due esami si è comportato meglio?

Il rendimento nei due esami non è direttamente confrontabile. Tuttavia è possibile convertire il rendimento in una variabile standardizzata e si ottiene

$$
\begin{aligned}
z_{\text{matematica}} &= \frac{80-72}{15} = 0.5333 \\
z_{\text{fisica}} &= \frac{84-80}{8} = 0.5
\end{aligned}
$$

Per cui si può affermare che seppur di poco, nell'esame di fisica il candidato si è comportato un po' meglio (rispetto alla media del corso).

Utilizzo delle variabili standardizzate per stimare valore statistici in base a medie e scarti quadratici

Esempio 2. La lunghezza media di 500 foglie di lauro di un certo cespuglio è di 151 mm e lo scarto quadratico medio è di 15 mm. Assumendo che la distribuzione delle foglie sia di tipo normale, trovate quante foglie hanno una lunghezza compresa tra 120 (minimo) e 155 (massimo) mm.

Si procede innanzitutto standardizzando le due lunghezze limite del problema:

$$
\begin{aligned}
z_{\text{minimo}} &= \frac{120-151}{15} = -2.066667 \\
z_{\text{massimo}} &= \frac{155-151}{15} = 0.2666667
\end{aligned}
$$

Cercando sulla tavola degli integrali della curva normale si vede che per 2.07 l'area è 0.4808 mentre per 0.27 è 0.1064. (Si noti come la curva a campana è simmetrica e quindi i valori negativi sono identici a quelli positivi!).

Nel caso qui trattato l'area che determina densità di probabilità di trovare foglie (e quindi indica la proporzione rispetto all'insieme globale di foglie) è data dalla somma della zona da -2.07 a 0 con la zona che va da 0 a 0.27 e quindi l'area sottesa alla curva normale è:

$$0.4808 + 0.1064 = 0.5872$$

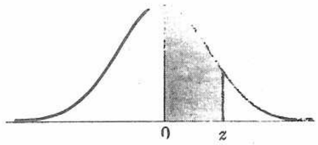
Siccome la curva è standardizzata e l'area totale è 1 questo vuol dire che il 58.72% delle foglie prese in considerazione sono all'interno della forchetta di lunghezze da 120 a 155 mm e quindi si hanno:

$$500 \cdot \frac{58.72}{100} = 293.6$$

Cioè si può ipotizzare che 294 foglie abbiano una lunghezza corrispondente alla forchetta di lunghezze che ci interessa.

Tavola degli integrali della curva normale

Aree sotto la curva normale standardizzata da 0 a z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Esercizi

Esercizio 1. Ricalcolare l'esempio delle foglie di lauro considerando che le foglie sono state misurate con un'approssimazione al millimetro e che quindi si può ipotizzare che quelle a cui è stata assegnata una lunghezza di 120 mm in realtà possono essere al minimo lunghe 119.5 mm e quelle a cui è stata assegnata la lunghezza di 155 mm possono al massimo essere lunghe 155.5 mm. Il risultato varia significativamente?

[− 2.10; 0.30; 0.6; 300 foglie]

Esercizio 2. Prendendo riferimento dall'esempio 1, si calcolino i valori standard di tre studenti che hanno ottenuto rispettivamente 60, 93 e 72 nell'esame di matematica. [− 0.8; 0; 1.4]

Esercizio 3. Prendendo rifermento sempre dall'esempio 1 si calcoli il punteggio ottenuto da due studenti che hanno ottenuto un valore standard di -1 e 1.6 nell'esame di matematica. [57; 96]

Esercizio 4. Trovate le aree sottese alla curva standard delimitate dai seguenti intervalli di z:

i. da $z=0$ a $z=1.2$;

ii. da $z=-0.68$ a $z=0$;

iii. da $z=0.46$ a $z=2.21$;

iv. da $z=0.81$ a $z=1.94$;

v. a sinistra di $z=-0.6$;

vi. a destra di $z=-1.28$;

vii. a destra di $z=2.05$ e a sinistra di $z=-1.44$;

Esercizio 5. I voti di un questionario di biologia andavano da 1 a 10 secondo il numero di risposte date a 10 domande: il voto assegnato consisteva in un punteggio intero. Il voto medio è stato 6.7 e lo scarto quadratico medio 1.2. Ipotizzando che i voti siano stati distribuiti in modo normale determinate:

i. la percentuale di studenti che hanno ottenuto il voto 6; [27%]

ii. il voto massimo del peggior 10% della classe; [5.2;5 all'unità]

iii. il voto minimo del miglior 10% della classe. [8.2; 8 all'unità]

(Visto che il punteggio assegnato è un valore intero si può ipotizzare che il voto 6 sia stato assegnato dal 5.5 al 6.5 ...).

Esercizio 6. Il diametro interno medio di un campione di 200 rondelle prodotte da una macchina è 0.502 cm, mentre lo scarto quadratico medio è di 0.005 cm. La. funzione a cui sono destinate queste rondelle permette che i limiti massimi di tolleranza per i diametri interni vadano da 0.496 a 0.508 cm. Qualora si esca da tali limiti, le rondelle sono considerate difettose. Determinate la percentuale di rondelle difettose prodotta dalla macchina, assumendo che i diametri siano distribuiti normalmente. Si calcoli sia il caso così come descritto, sia il caso in cui si consideri che la precisione di mnisurazione è al millesimo e che quindi il limite 0.496 debba avere il minimo teorico a 0.4955 e il massimo in ... Si valuti anche la differenza tra i due calcoli

[23%]