17. LA DISTRIBUZIONE NORMALE E LA FUNZIONE DI GAUSS

17.1 LA DISTRIBUZIONE DEI DATI

Nel trattare gli errori casuali abbiamo utilizzato il concetto di media aritmetica ed il concetto di deviazione standard e abbiamo associato il loro valori a specifici parametri per la descrizione degli errori in una misura.
I concetti di media e di deviazione standard hanno anche un significato statistico intrinsecamente legato alla distribuzione dei dati che si sta analizzando.
La distribuzione dei dati è una caratteristica che descrive la probabilità di ottenere quel dato quando lo si estrae dal campione che stiamo analizzando.

Supponiamo di avere tre campioni ognuno di 10 palline aventi massa rispettivamente:

Campione 1: 10 palline con lo a stessa massa pari a 21 g
Campione 2: 5 palline con massa 18 g e 5 palline con massa 24 g
Campione 3: 1 pallina 18 g, 2 palline 19 g, 3 palline 21 g, una 22 g, una 23 g e due 24 g
In numeri:
Campione 1 : 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ; 21 g ;
Campione 2 : 18 g ; 18 g ; 18 g ; 18 g ; 18 g ; 24 g ; 24 g ; 24 g ; 24 g ; 24 g ;
Campione 3 : 18 g ; 19 g ; 19 g ; 21 g ; 21 g ; 21 g ; 22 g ; 23 g ; 24 g ; 24 g ;

media Campione 1 : 21 g	**ds** Campione 1 : 0.0 g
media Campione 2 : 21 g	**ds** Campione 2 : 1.1 g
media Campione 3 : 21 g	**ds** Campione 3 : 1.1 g

Le medie dei tre campioni sono uguali mentre le **ds** sono uguali solamente per il campione 2 e 3, per il primo campione la **ds** è nulla in quanto le masse sono uguali tra loro.

Le distribuzioni dei tre campioni sono mostrate negli istogrammi di figura 6

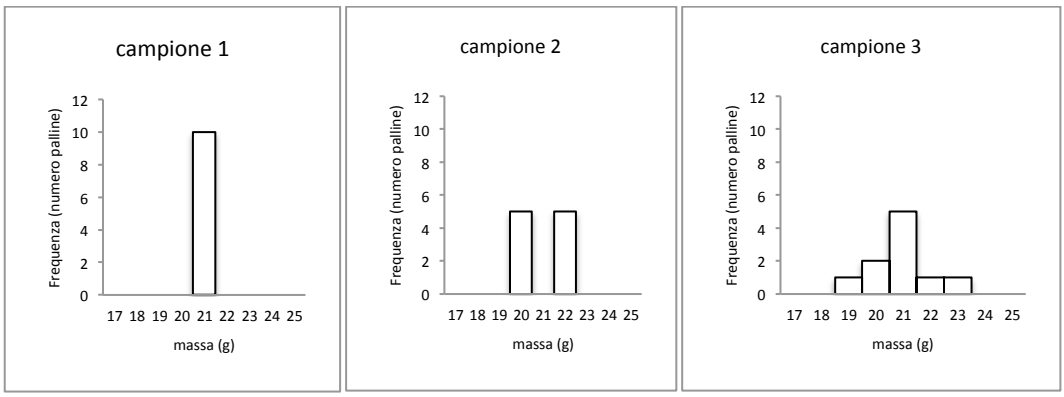

Figura 6. Distribuzione delle frequenze delle masse dei tre campioni

Questi istogrammi mettono in evidenza che i campioni sono decisamente diversi, risultano diversi anche il campione 2 e 3 che hanno gli stessi valori della **media** e della **ds**.
Quindi i due parametri (media e ds) non sono sufficienti a caratterizzare completamente il campione.

17.1 LA DISTRIBUZIONE DELLE FREQUENZE O PROBABILITA

Per capire meglio il legame tra media, deviazione standard e la distribuzione dei dati abbiamo bisogno di generalizzare il concetto di distribuzione introducendo il concetto di distribuzione delle frequenze relative, dove per frequenza relativa intendiamo il numero di volte che un dato compare rispetto al numero totale dei dati.
Se il dato *i-esimo* compare $\mathbf{n_i}$ volte allora la sua frequenza relativa $\mathbf{f_i}$ sarà data da:

$$f_i = \frac{n_i}{N}$$

dove *N* è il numero di elementi che costituisce il campione.
Le frequenze assumono valori compresi tra 0 e 1. La frequenza esprime la probabilità di trovare quel valore di massa nelle palline di ciascun campione. Si noti che numericamente la frequenze e la probabilità hanno stesso valore numerico. In tabella 1 e tabella 2 sono riportati i dati delle distribuzioni delle masse e delle frequenze dei tre campioni.

Tabella 1. Distribuzione delle frequenze relative delle masse dei tre campioni

	campione 1	campione 2	campione 3
massa (g)	n°	n°	n°
17	0	0	0
18	0	0	0
19	0	0	1
20	0	5	2
21	10	0	5
22	0	5	1
23	0	0	1
24	0	0	0
25	0	0	0
N	10	10	10

Tabella 2. Distribuzione delle frequenze relative con sui si presentano le masse dei tre campioni

	campione 1	campione 2	campione 3
massa (g)	f	f	f
17	0	0	0
18	0	0	0
19	0	0	0,1
20	0	0,5	0,2
21	1	0	0,5
22	0	0,5	0,1
23	0	0	0,1
24	0	0	0
25	0	0	0
F	1	1	1

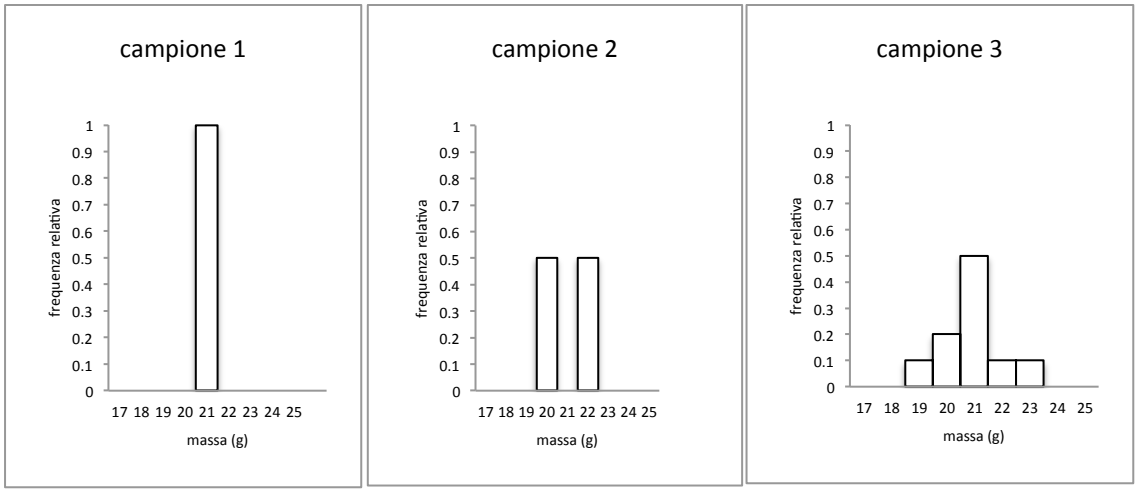

Figura 7. Distribuzione delle frequenze relative o della probabilità con sui si presentano le masse dei tre campioni

La differenza tra le distribuzioni di figura 6 e di figura 7 sta solo nella scala numerica dell'asse delle ordinate (numero palline nel primo caso, frequenza o probabilità nel secondo caso) ma il loro aspetto ovvero la forma delle distribuzioni sono sovrapponibili.

17.3 LA DISTRIBUZIONE NORMALE

Quando il valore di una grandezza assume valori diversi e l'origine di queste differenze che chiameremo fluttuazioni, sono completamente casuali allora la distribuzione delle frequenze con cui questi valori si presenteranno ha una distribuzione nota a priori. Se un suo campione è costituito da un numero di elementi ***N*** elevatissimo allora la ***densità della probabilità*** con cui si presenteranno questi valori è descritta dalla funzione di Gauss:

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

dove:
$\bar{x}$ è la media della grandezza del campione e σ è la **ds** (deviazione standard del campione).

Questa funzione ha alcune caratteristiche molto importanti:

1) Ha il massimo quando la variabile *x* assume un valore uguale alla media $\bar{x}$;
2) È simmetrica rispetto al massimo
3) È sempre positiva e tende zero per valori di *x* molto maggiori e molto minori della media.
4) Presenta due flessi a destra e sinistra del massimo nei punti $\bar{x}$ - σ e $\bar{x}$ + σ
5) $\int_{-\infty}^{+\infty} f(x)dx = 1$ ovvero l'area sottesa dalla curva è adimensionale e vale 1.
6) Al variare della media la curva trasla sull'asse delle *x*.
7) Il valore della deviazione standard σ determina la larghezza della curva

Nella figura 8 viene riportata a titolo di esempio la densità di probabilità di un campione gaussiano di oggetti aventi massa media pari a 8.0 g e deviazione standard σ = 1.2 g. Le figure 9 e 10 mostrano come si modifica la distribuzione di probabilità gaussiana al variare della media e della deviazione standard.
E' necessario sottolineare che il punto 5) fornisce anche l'informazione che il prodotto $f(x)dx$ è adimensionale per cui se dx esprime una differenza di massa che misuriamo in [g] allora l'unità di misura della densità di probabilità $f(x)$ è [1/g].
Quindi le dimensioni dei valori della funzione densità di probabilità *f(x)* saranno :

1/[m] = [m^{-1}] se la variabile *x* è una lunghezza,
1/[s] = [s^{-1}] se la variabile *x* è un tempo
1/[kg] = [kg^{-1}] se la variabile *x* è una massa
e così via.

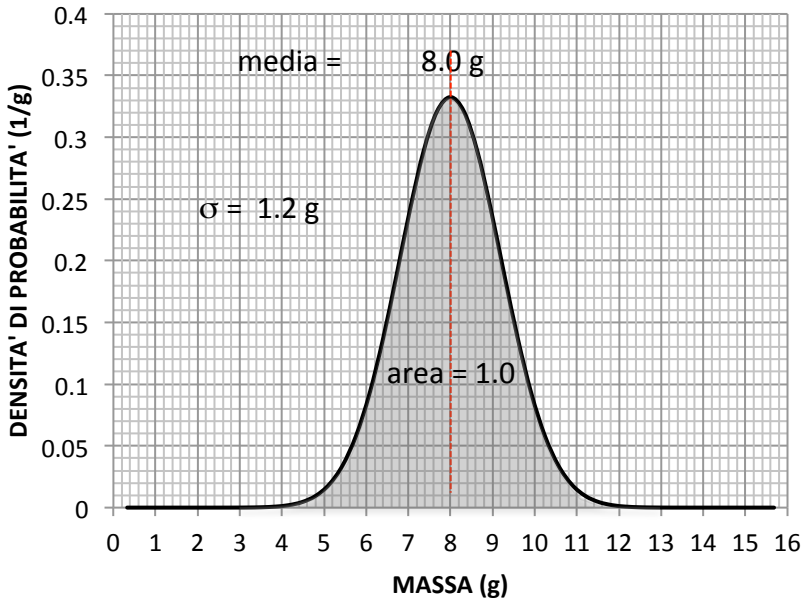

Figura 8. Distribuzione delle frequenze o della probabilità normalizzata ovvero area = 1.0

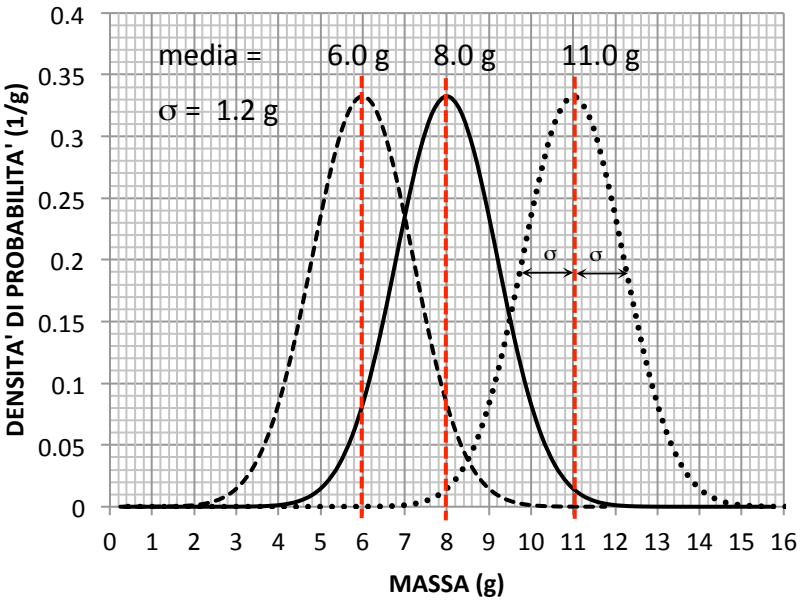

Figura 8. Distribuzione delle frequenze o della probabilità al variare della media $\bar{x}$

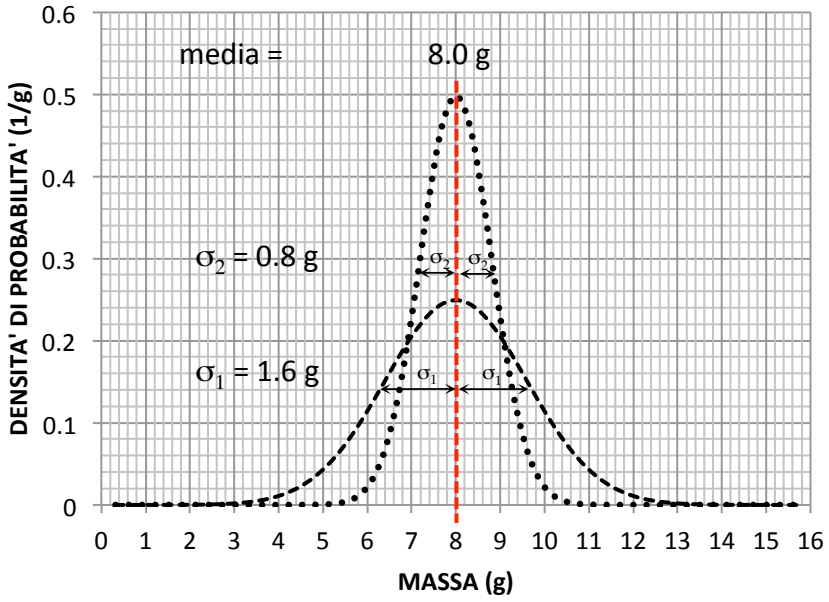

Figura 9. Distribuzione delle frequenze o della probabilità al variare della deviazione standard σ

Il coefficiente $\frac{1}{\sigma\sqrt{2\pi}}$ della funzione di Gauss (1), serve a normalizzare il suo integrale: con quel coefficiente l'integrale della funzione di Gauss tra - ∞ e +∞ vale esattamente 1.

Si noti che è proprio questo coefficiente che determina l'unità di misura della densità di probabilità *f(x)* infatti tutto il resto dell'espressione è adimensionale.
Questa osservazione ci porta ancora a concludere che le unità di misura dei valori della funzione densità di probabilità *f(x)* sono il reciproco delle unità di misura della deviazione standard σ che sono ovviamente anche quelle della media.

Ora, come facciamo trasformare il valore della densità di probabilità *f(x)* fornito dalla funzione di Gauss in un valore di frequenza o di una probabilità?

La risposta è molto semplice: è sufficiente moltiplicare il suo valore per un intervallo Δ*x*.
Infatti il prodotto *f(x).* Δ*x* è un numero puro ovvero non ha dimensioni e rappresenta una frequenza o una probabilità.
In particolare, questo prodotto rappresenta la frequenza o la probabilità con cui nei dati del nostro campione compaiono valori contenuti nell'intervallo x±(Δ*x*)/2.

Nella figura 10 è mostrato come calcolare la probabilità o la frequenza che un elemento del campione assuma un valore compreso tra 6,8 g e 7,2 g (7,0 ± 0,2 g) *Δx = 0.4 g*.
Si calcola il valore della densità di probabilità *f(x)* per *x=7,0 g* poi lo si moltiplica per l'intervallo *Δx = 0.4 g.* Il risultato 0,092 è una probabilità o una frequenza. (Questo valore ci dice che il 9,2% (infatti: 0,092/1 = 9,2/100) dei dati del campione assume un valore nell'intervallo considerato)

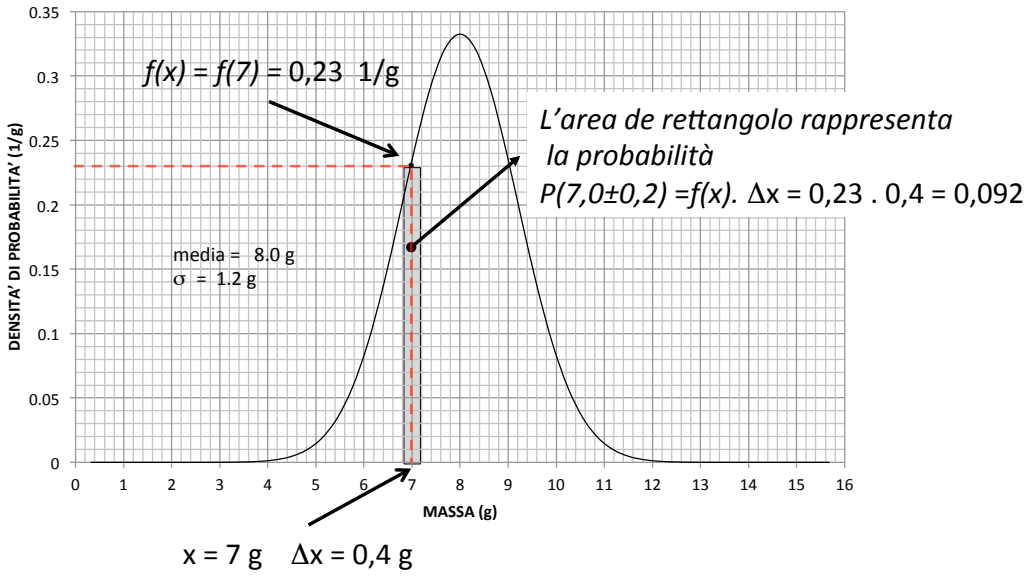

Figura 10. Determinazione grafica della probabilità che un valore del campione abbia un valore compreso tra 6,8 g e 7,2 g
in altri termini un valore compreso 7,0 g ± 0,2 g

Questa tecnica approssimata di calcolo ci permette di stimare con rapidità le frequenze previste per valori del campione negli intervalli di nostro interesse. Nella figura 11 è riportato un semplice esempio per calcolare la frequenza o lo probabilità in intervalli a nostro piacere sommando tra loro le aree di rettangoli simili a quelli di figura 10.

Esempio di calcolo approssimato delle frequenze che i dati del campione assumano valori compresi tra 8.0 g ± 1.2 g
Il valore viene ottenuto sommando tutte le aree dei rettangoli evidenziati in figura.
Ogni rettangolo ha una larghezza di base pari a 0,4 g
Frequenza o probabilità = (0.23 + 0.29 + 0.325 + 0.325 + 0.29 + 0.23)• 0,4 = 0.676

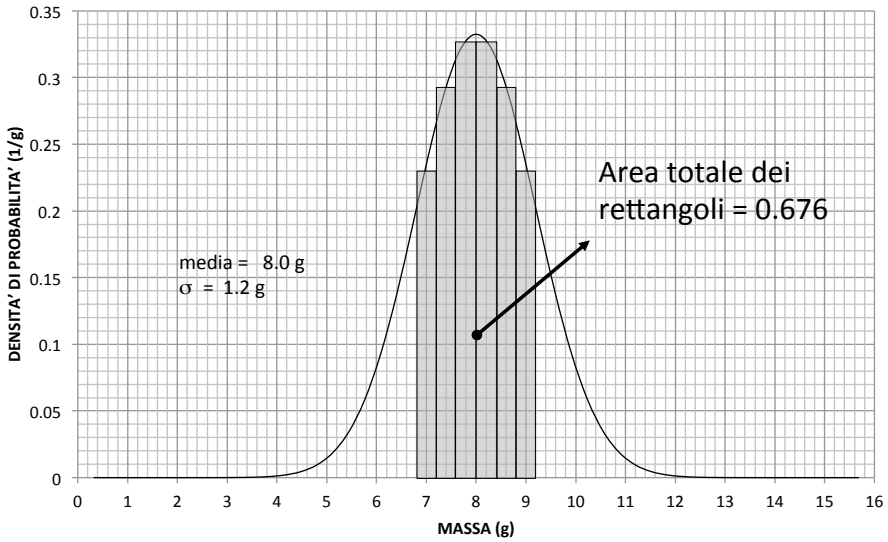

Figura 11. Determinazione grafica della probabilità che un valore del campione abbia un valore compreso tra 6,8 g e 9,2 g
in altri termini un valore compreso 8,0 g ± 1,2 g (media ± σ)

Questi esempi di calcolo approssimato delle frequenze o probabilità mette in evidenza che l'area sottesa dalla curva di Gauss rappresenta proprio la frequenza o probabilità che i dati del nostro campione abbiano valori inclusi tra gli estremi dell'area presa un considerazione. La figura 12 mette in evidenza i valori di tre aree, quella per cui x assume valori < 6,8 g; quella per cui 6,8 g < x < 9,2 g e quella per cui x > 9,2 g. Dal punto di vista matematico queste aree possono essere calcolate con gli integrali sotto riportati.

$$\int_{-\infty}^{+6.8} f(x)dx = 0.165 \quad ; \quad \int_{+6.8}^{+9.2} f(x)dx = 0{,}670 \quad ; \quad \int_{+9.2}^{+\infty} f(x)dx = 0.165$$

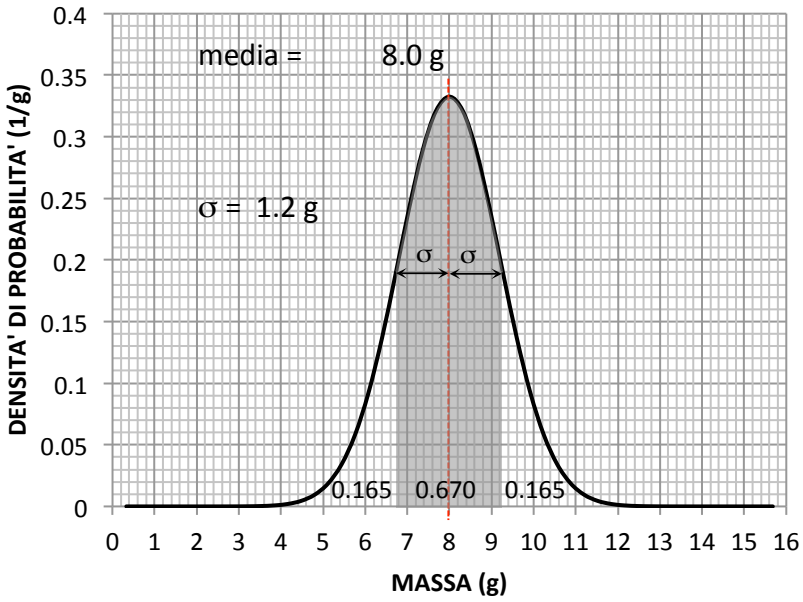

Figura 12. Nella figura sono riportati i valori della probabilità che un dato del campione abbia un valore minore di 6,8 g (8,0 g - 1,2 g (media - σ); che un dato del campione abbia un valore compreso tra 6,8 g e 9,2 g in altri termini un valore compreso 8,0 g ± 1,2 g (media ± σ); che un dato del campione abbia un valore maggiore di 9,2 g (8,0 g + 1,2 g (media + σ).

17.4 VERIFICA DELL'IPOTESI CHE UNA DISTRIBUZIONE DEI DATI SIA GAUSSANA

Le informazioni fornite dalla funzione di Gauss possono essere utilizzate per verificare se le distribuzioni delle frequenze dei dati sperimentali si adattano a meno alla previsione che abbiano una distribuzione GAUSSIANA o NORMALE (questi due termini sono sinonimi). Prendiamo in considerazione i campioni 2 e 3 le cui distribuzioni sono presentate in figura 7, entrambi hanno la stessa media e la stessa deviazione standard. Se prendiamo questi due valori (media = 21.0 g e d.s.= 1.1 g) e li mettiamo nella funzione di Gauss otteniamo la distribuzione riportata in figura 13 (istogramma di colore rosso). La curva rossa non rappresenta minimamente il campione 2, il campione è invece un po più simile anche se le differenze sono evidenti. Il livello di sovrapposizione delle due distribuzioni (quella del campione e quella di Gauss) può essere utilizzato per quantificare la somiglianza.

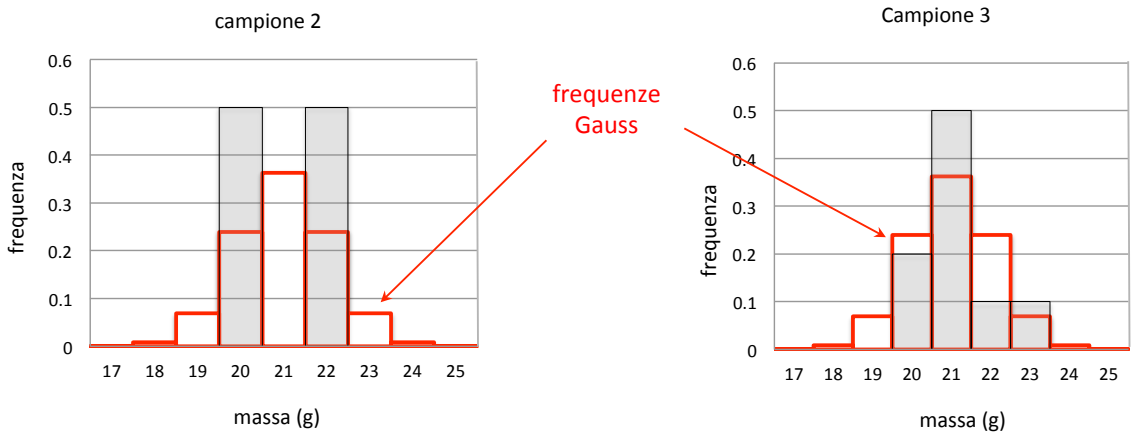

Figura 13. Confronto fra le frequenze dei campioni 2 e 3 e delle frequenze calcolate utilizzando la distribuzione di Gauss (tratto rosso). La distribuzione dei campioni non assomigliano molto a quella prevista dalla distribuzione Gaussiana, in particolar modo quella del campione 2.

Chiamiamo fc_1, fc_2, fc_3 …… fc_n le frequenze del campione e con fg_1, fg_2, fg_3 …… fg_n le frequenze calcolate con la funzione di Gauss.
Moltiplicando il valore della frequenze previste dalla funzione di Gauss (fg_i)per il numero di elementi del campione N (10 nel nostro caso) otterremo il numero di elementi che il campione dovrebbe avere per ogni valore di massa ovvero:
$ng_i = N \cdot fg_i$

Nella figura.14 viene riportato il confronto tra la distribuzione dei campioni 2 e 3 con quelle previste dalla distribuzione di Gauss.

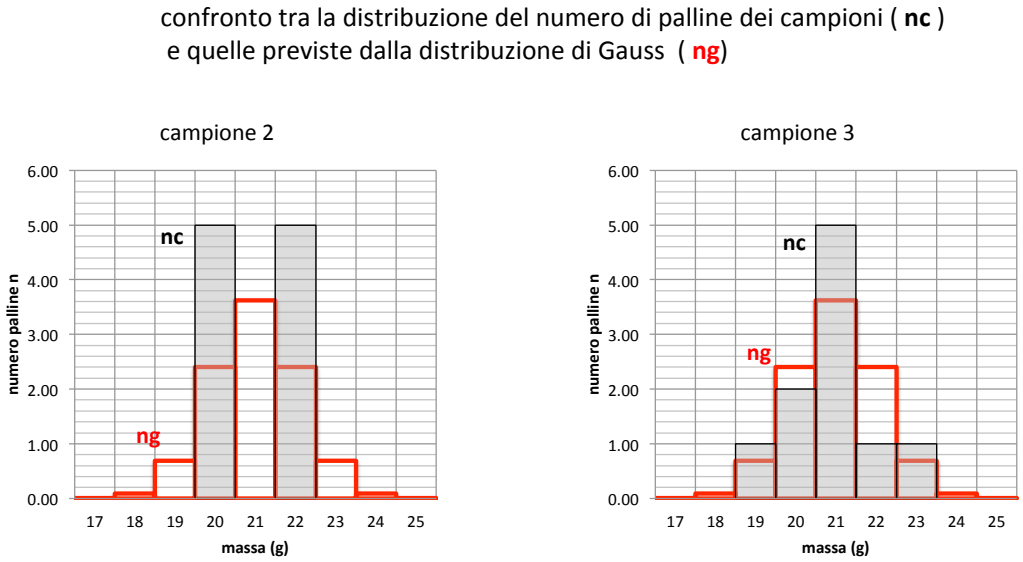

Figura 14. Confronto fra le distribuzioni delle palline dei campioni 2 e 3 e delle distribuzioni delle palline calcolate utilizzando la funzione di Gauss (tratto rosso). Anche in questo caso le distribuzione non assomigliano molto a quelle previste dalla distribuzione di Gauss, in particolar modo quella del campione 2.

La differenza tra il numero di elementi dei campione e quelli calcolati con la distribuzione di Gauss potrebbe essere un indicatore di somiglianza delle due distribuzioni, infatti se **nc** e **ng** fossero tutte uguali allora le differenze sarebbero tutte nulle. Ma come si vede in figura 15 le differenze ci sono positive quando **nc** > **ng** e negative quando **nc**< **ng** per cui la somma delle differenze si potrebbe annullare anche se se le frequenze sono molto diverse tra loro. In ogni caso possiamo già capire che le differenze sono maggiori nel caso del campione 2. Quindi il campione 3 segue meglio il profilo previsto della distribuzione di Gauss

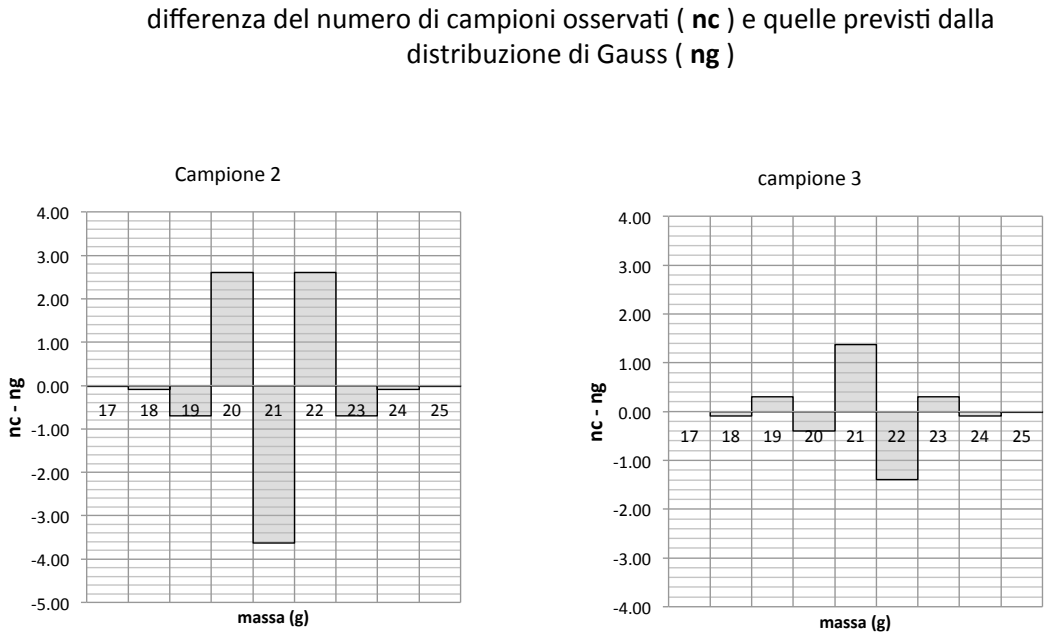

Figura 15. Confronto fra le differenze delle distribuzioni dei campioni e quelle calcolate utilizzando la funzione di Gauss le differenze sono visibilmente minori nel caso del campione 3.

Per rendere più significative queste differenze è stato introdotto il parametro:

$$C_i = \frac{(nc_i - ng_i)^2}{ng_i}$$

La figura 16 riporta il valore di questo parametro nel confronto dei nostri campioni.

La somma di tutti i valori di C_i rappresenti dall'area delle barre degli istogrammi di figura 16 prende il nome di χ^2 (chi quadro).

$$\chi^2 = \sum_{i=1}^{n} \frac{(nc_i - ng_i)^2}{ng_i}$$

in questo caso n è il numero delle classi per le quali abbiamo calcolato il valore del numero di campioni previsti dalla distribuzione di Gauss ovvero (7) che corrisponde al numero di rettangoli grigi della figura 16.

andamento del parametro $C_i = \frac{(nc_i - ng_i)^2}{ng_i}$ nel confronto dei campioni

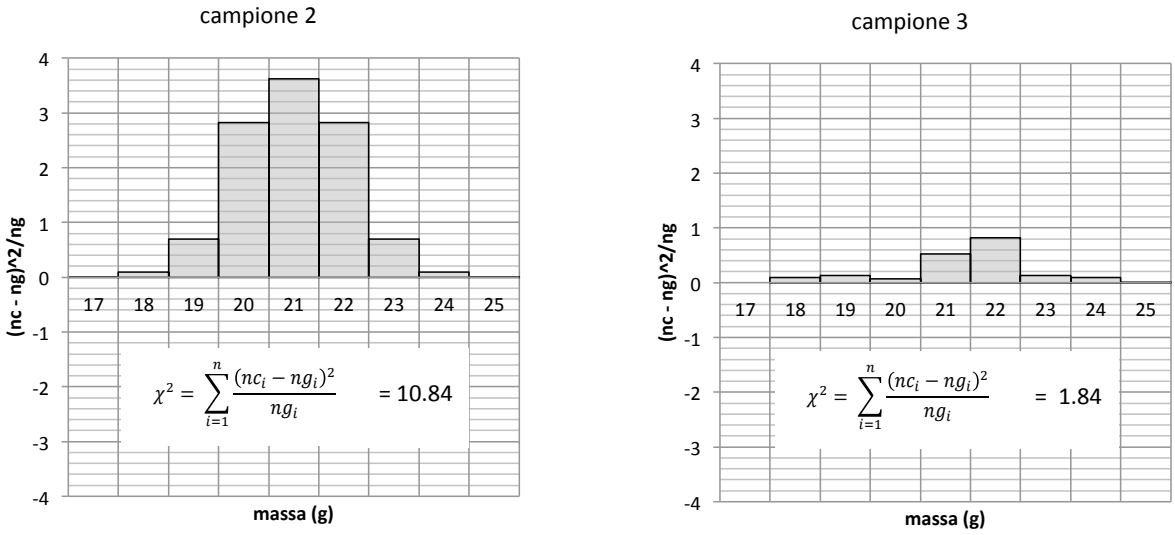

Figura 15. Confronto del parametro *Ci* fra distribuzioni dei campioni 2 e 3 e confronto della loro somma che rappresenta il χ^2 delle due distribuzioni . Questo parametro è significativamente minore per il campione 3.

Il valore del χ^2 risulta inferiore nel caso del campione 3 il che sta ad indicare che la probabilità che la distribuzione dei dati del campione 3 è più somigliante alla distribuzione gaussiana di quanto non lo sia quella del campione 2.
Una semplice regola per accettare o rifiutare l'ipotesi che la distribuzione sia o meno gaussiana è la seguente:

se χ^2 ≤ al numero delle classi (n. di colonne dell'istogramma) l'ipotesi è accettabile
se χ^2 >> del numero delle classi (n. di colonne dell'istogramma) l'ipotesi va rigettata

Nell'esempio di fig.15 il numero di classi è 7 quindi per il campione 3 l'ipotesi è accettabile.
Questo criterio può essere formulato in un modo più generale introducendo alcuni concetti statistici come il concetto di gradi di libertà che però non affronteremo in questa dispensa.