Edoardo Milotti
novembre 2009

Introduzione ai metodi di misura e analisi dati in Fisica

1. Introduzione

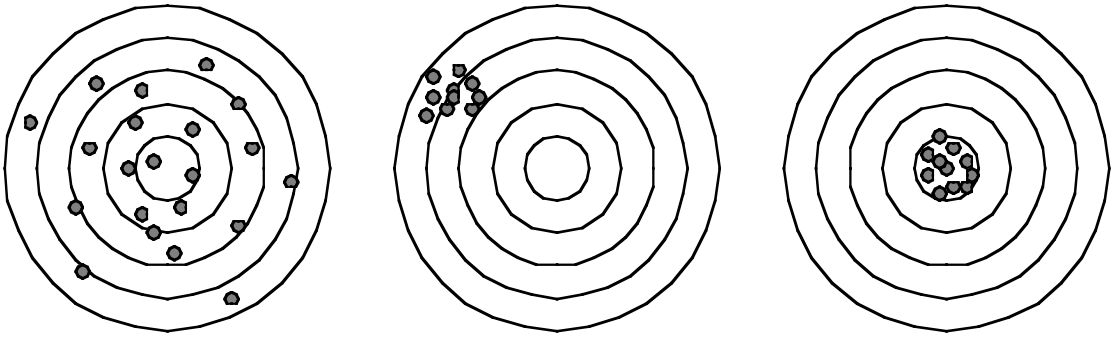
Lo scopo della Fisica è quello di costruire modelli matematici predittivi per fenomeni del mondo fisico nel quale viviamo. I modelli matematici possono essere costruiti solo a partire da misure quantitative, che costituiscono la base ed il punto di partenza per la formulazione di qualunque teoria fisica, che poi deve anche essere in grado di fare predizioni e di confrontarsi con successo con misure effettuate a posteriori.
È chiaro allora che le metodologie di misura e le tecniche matematiche di analisi dei dati sperimentali costituiscono una parte importante dello studio della Fisica. In queste note vengono delineati alcuni dei concetti necessari all'analisi ed all'esecuzione delle misure.

Prima di iniziare a studiare alcuni degli strumenti matematici impiegati nell'analisi dei dati, facciamo alcune considerazioni preliminari sulle incertezze di misura.
Le sorgenti di incertezza possono essere di diversi tipi:

1. Ci sono delle incertezze statistiche: l'operazione di misura è a sua volta un processo fisico, e l'esecuzione di una misura potrebbe non essere esattamente ripetibile perché potremmo non essere in grado di rimettere il sistema fisico ogni volta *esattamente* nello stesso stato. In tal caso possiamo pensare che i dati siano valori particolari di una o più variabili aleatorie, che possiedono una certa distribuzione di probabilità. Questi dati potrebbero essere distribuiti con una variabilità più o meno grande rispetto ad un certo valore medio. Questa variabilità statistica associata alla misura si chiama *precisione* della misura. Tanto minore è la variabilità statistica e tanto più precisa è la misura.
2. Ci sono delle incertezze ripetibili ma ignote - dette incertezze *sistematiche* - dovute ai meccanismi della misura. Queste incertezze determinano l'*accuratezza* della misura. Tanto minori sono le incertezze ripetibili e tanto più accurata è la misura.

3. In molte apparecchiature moderne, i meccanismi fisici di misura sono talmente complessi da rendere impossibile la separazione tra incertezza statistica ed incertezza sistematica. In tal caso si parla semplicemente di *accuratezza*, intendendo con ciò l'incertezza combinata statistica+sistematica.

La figura che segue mostra nella prima parte un bersaglio con dei colpi accurati ma imprecisi (grande incertezza statistica, trascurabile componente sistematica), nella seconda parte dei colpi precisi ma inaccurati (piccola incertezza statistica, elevata componente sistematica) e nella terza parte dei colpi precisi e accurati (piccola incertezza statistica, trascurabile componente sistematica).



2. Istogrammi ed indicatori statistici

Da quanto detto sopra si capisce che una descrizione matematica corretta delle misure fisiche richiede dei concetti probabilistici. In questa sezione vediamo come si possono introdurre dei concetti statistici per l'analisi dei dati sperimentali. Introduciamo anzitutto gli istogrammi, che ci permettono di dare una rappresentazione grafica dei dati ottenuti in un esperimento.

Un istogramma è una rappresentazione grafica della frequenza con cui osserviamo una certa grandezza. Supponiamo ad esempio di voler fare un istogramma della frequenza con cui i segni 1, X, 2 compaiono sulla schedina del Totocalcio. Per le partite giocate in una certa giornata potremmo allora trovare qualcosa del genere:

- il segno 1 compare 6 volte,
- il segno X compare 4 volte,
- il segno 2 compare 3 volte.

Con queste frequenze costruiamo allora la figura che segue.

Ad ogni segno associamo una sbarra verticale la cui lunghezza dà la frequenza con cui compare un certo segno sulla schedina. La rappresentazione grafica che abbiamo ottenuto è un istogramma: naturalmente l'istogramma non dà nessuna informazione in più rispetto la tabellina riportata sopra, però permette di apprezzare la distribuzione dei dati con una sola occhiata.

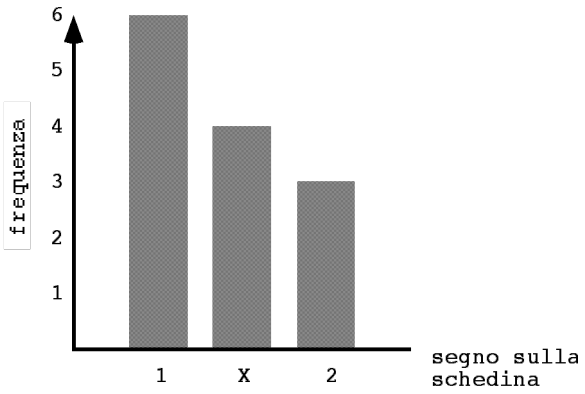


Figura 1

Nel caso in cui vogliamo studiare una grandezza che può assumere solo valori interi, la costruzione dell'istogramma è semplicissima. Partiamo da una tabella di dati, ad esempio:

Tabella 1

n	*x*
1	5
2	6
3	4
4	6
5	7
6	4
7	5
8	5
9	6
10	5

n indica il numero progressivo della misura, ed *x* è la grandezza misurata, che, come si vede, assume solo valori interi. Il valore minimo di *x* nella tabella è 4, mentre il

valore massimo è 7. Contiamo adesso la frequenza - cioè il numero di volte - con cui compaiono i valori da 4 a 7:

Tabella 2

valore	frequenza
4	2
5	4
6	3
7	1

Disegniamo adesso l'istogramma: tracciamo gli assi e disegniamo delle sbarre di lunghezza proporzionale ai valori della frequenza riportati in tabella:

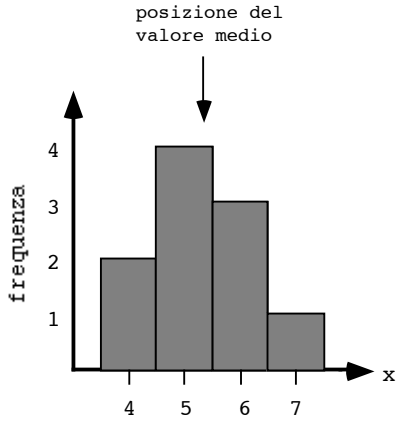


Figura 2

Se confrontiamo i dati riportati nella prima tabella con questo istogramma apprezziamo immediatamente il grande vantaggio della rappresentazione grafica. È immediatamente evidente che in questo caso i dati si accumulano intorno ad un certo valore, che adesso cerchiamo di definire con precisione.

Se guardiamo l'istogramma qui sopra vediamo subito che la frequenza è massima in corrispondenza al valore x=5. Questo numero è un primo indicatore del valore centrale dell'istogramma. Un altro indicatore è il valore medio: supponiamo che la variabile x possa assumere gli n valori $x_1,...,x_n$, con frequenza $f_1,...,f_n$; allora il valore medio è definito dalla seguente formula

$$\textit{valore medio di } x = \bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n}.$$

Se applichiamo questa definizione ai valori ed alle frequenze della variabile *x* nella seconda tabella troviamo che il valore medio di *x* è

$$\textit{valore medio di } x = \bar{x} = \frac{4 \cdot 2 + 5 \cdot 4 + 6 \cdot 3 + 7 \cdot 1}{2 + 4 + 3 + 1} = 5.3$$

Osserviamo anche che il numero totale di misure fatte è proprio la somma di tutte le frequenze *f*

$$\textit{numero totale di misure} = N = f_1 + f_2 + \ldots + f_n,$$

e che nel caso che stiamo considerando il numero totale di misure è *N*=2+4+3+1=10.

Abbiamo visto che il valore della variabile *x* che corrisponde alla frequenza massima (questo valore di *x* è detto "valore modale" o "moda" dell'istogramma) oppure il valore medio indicano approssimativamente la posizione del "centro" dell'istogramma. Consideriamo ora gli istogrammi mostrati nella figura seguente

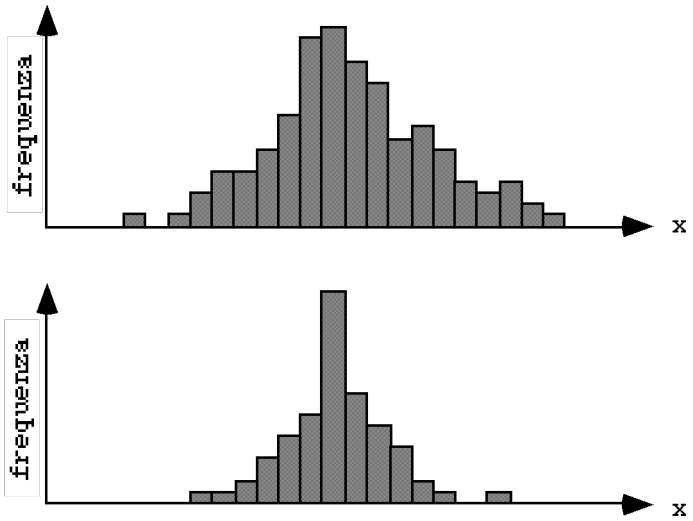


Figura 3

Questi due istogrammi hanno lo stesso valore modale, e quasi lo stesso valore medio, però sono chiaramente diversi tra loro, il primo è più "allargato" del secondo. È chiaro che non basta specificare la posizione del centro dell'istogramma, si deve anche dare una misura della sua "larghezza".

Se x_{min} e x_{max} sono il valore minimo e massimo che la variabile x può assumere allora la grandezza (x_{max}-x_{min}) è un indicatore della larghezza dell'istogramma: il suo significato è chiaro, l'istogramma è tanto più "largo" quanto più grande è (x_{max}-x_{min}). Ad esempio, se supponiamo che nel caso mostrato in figura 2 la variabile x non possa assumere altri valori al di fuori di quelli mostrati, allora (x_{max}-x_{min})=7-4=3.

Il principale difetto dell'indicatore di larghezza (x_{max}-x_{min}) è che noi non siamo sempre in grado di dire quale sia il valore minimo e quale il valore massimo che x può assumere, ed è per questo motivo che questo indicatore viene utilizzato piuttosto raramente.

Un altro indicatore della larghezza dell'istogramma di uso assi più frequente è la varianza, che dà una misura della "deviazione dal valore medio". Se la varianza di una certa quantità x è grande, allora i valori di x sono molto dispersi attorno al valore medio, e, viceversa, se la varianza è piccola i valori sono poco dispersi.

Consideriamo ad esempio la prima misura, x_1: la deviazione dal valore medio è data dalla quantità $(x_1 - \bar{x})$. Ora, se x_1 è più piccola del valore medio, questa quantità è negativa: però a noi non interessa il segno della deviazione, ci interessa solo sapere quanto è grande. Perciò, invece di $(x_1 - \bar{x})$, consideriamo la quantità $(x_1 - \bar{x})^2$, che è sempre positiva, e facciamo la media di tutte queste deviazioni per ottenere la varianza:

$$\textit{varianza di } x = \mathrm{var}(x) = \frac{(x_1 - \bar{x})^2 f_1 + (x_2 - \bar{x})^2 f_2 + \ldots + (x_n - \bar{x})^2 f_n}{f_1 + f_2 + \ldots + f_n}.$$

La varianza è una specie di media delle deviazioni al quadrato, quindi per riottenere una media “semplice”, prendiamo la radice quadrata della varianza. Questa quantità si chiama “dispersione” o “deviazione standard” o “scarto quadratico medio”, e si indica di solito con la lettera greca sigma (σ):

$$\textit{deviazione standard di } x = \sigma = \sqrt{\mathrm{var}(x)} = \sqrt{\frac{(x_1 - \bar{x})^2 f_1 + (x_2 - \bar{x})^2 f_2 + \ldots + (x_n - \bar{x})^2 f_n}{f_1 + f_2 + \ldots + f_n}}$$

Nel caso dell'istogramma mostrato nella figura 2

$$\textit{varianza di } x = \mathrm{var}(x) = \frac{(4-5.3)^2 \cdot 2 + (5-5.3)^2 \cdot 4 + (6-5.3)^2 \cdot 3 + (7-5.3)^2 \cdot 1}{2+4+3+1} = 0.81$$

e

$$\textit{deviazione standard di } x = \sigma = \sqrt{\mathrm{var}(x)} = \sqrt{0.81} = 0.9\,.$$

Se l'istogramma è molto largo e i dati sono molto dispersi allora la deviazione standard è grande, e viceversa se i dati sono poco dispersi la deviazione standard è piccola.

Finora abbiamo sempre considerato una variabile x che può assumere solo valori interi: vediamo ora come si fa un istogramma nel caso generale in cui x abbia anche valori non interi.

Anche in questo caso partiamo da una tabella di dati, ad esempio:

Tabella 3

n	x	x'	x''
1	5.4	5	5.5
2	6.3	6	6.5
3	4.7	5	5.0
4	6.6	7	6.5
5	7.2	7	7.0
6	4.2	4	4.0
7	5.1	5	5.0
8	5.9	6	6.0
9	5.7	6	5.5
10	5.8	6	6.0
11	5.9	6	6.0
12	6.4	6	6.5
13	4.7	5	4.5
14	6.2	6	6.0
15	6.9	7	7.0
16	3.9	4	4.0
17	5.5	6	5.5
18	4.9	5	5.0
19	6.4	6	6.5
20	4.8	5	5.0

n indica il numero progressivo della misura, ed *x* è la grandezza misurata, che, come si vede, ora non assume solo valori interi. Definiamo una variabile *x*' che è l'arrotondamento di *x* al più vicino numero intero: ad esempio, la dodicesima misura ci dà *x*=6.4, allora il suo valore arrotondato all'intero più vicino è *x'* = 6. La variabile *x*' assume solo valori interi e possiamo costruire un istogramma per *x'* proprio come abbiamo fatto sopra. Quella che segue è la tabella delle frequenze della variabile *x'*

Tabella 4

x'	*frequenza*
4	2
5	6
6	9
7	3

Questo non è il solo modo in cui possiamo arrotondare i nostri dati, ad esempio possiamo arrotondare al più vicino valore semi-intero, cioè ai valori 4.0, 4.5, 5.0, 5.5, 6.0, ... : tornando al caso di prima vediamo che ora la dodicesima misura *x* = 6.4 viene arrotondata a *x"* = 6.5. Procedendo in questo modo otteniamo una nuova variabile *x"*, le cui frequenze sono riportate nella seguente tabella

Tabella 5

x"	*frequenza*
4.0	2
4.5	1
5.0	4
5.5	3
6.0	4
6.5	4
7.0	2

La figura seguente mostra gli istogrammi corrispondenti alle variabili *x'* e *x"*.

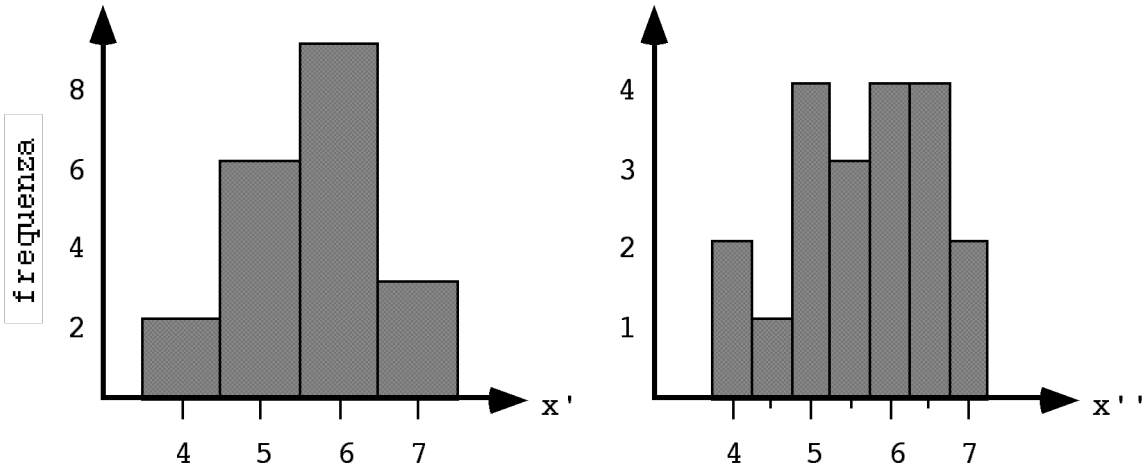


Figura 4

Le variabili *x'* e *x"* sono state costruite a partire dagli stessi dati, eppure gli istogrammi sembrano abbastanza diversi: come mai?

Il motivo è che abbiamo fatto troppo poche misure: un istogramma è una rappresentazione efficace dei dati solo quando ce ne sono abbastanza. Una regola utilizzata in statistica è che la rappresentazione è efficace se le celle dell'istogramma corrispondono ad almeno 5 misure. Facendo uso di questa regola si vede (figura 4) che l'istogramma a sinistra è migliore di quello a destra.

Così se suddividiamo molto le celle di un istogramma riusciamo a rappresentare dettagli fini della distribuzione dei dati, ma d'altra parte decresce anche il contenuto di ciascuna cella, e quindi un "buon" istogramma deriva da un compromesso tra finezza dei dettagli che vogliamo rappresentare e quantità dei dati in ciascuna cella. La conclusione è che per ottenere un istogramma che mostra anche i dettagli fini della distribuzione dobbiamo raccogliere molti dati.

Per finire ecco un utile esercizio: prendiamo 10 monete uguali, un bicchiere di plastica ed un panno morbido. Mettiamo le monete nel bicchiere, scuotiamole un po' e lanciamole sul panno. Contiamo il numero delle teste che sono uscite in questo primo lancio e registriamolo su un foglio di carta. Lanciamo ancora, contiamo le teste, e così via, registrando ogni volta il numero di teste che è uscito. In questo modo costruiamo una tabella di valori per la variabile x = "numero di teste uscite in un lancio delle monete".

Costruiamo l'istogramma delle frequenze di x. Possiamo definire un valore massimo ed uno minimo per x? Qual è il valore medio di x? Qual è la larghezza dell'istogramma?
Sulla base di considerazioni generali, si può dire quale sia la distribuzione di probabilità della variabile aleatoria x?

3. Probabilità

A partire dalle frequenze osservate è possibile definire la probabilità di un trovare un certo valore della variabile aleatoria x:

$$\text{probabilità di trovare il valore } x_n = p_n = \frac{f_n}{\sum_n f_n}$$

Questa è la *definizione frequentista* della probabilità, ed in realtà non è ancora completa: abbiamo già visto nella sezione dedicata agli istogrammi, che se ci sono pochi dati gli istogrammi possono essere irregolari, e questo è vero anche per le probabilità definite in questo modo. E dunque, se vengono effettuate N misure, la corretta definizione frequentista delle probabilità è la seguente

$$\text{probabilità di trovare il valore } x_n = p_n = \lim_{N\to\infty} \frac{f_n}{\sum_{n=1}^{N} f_n} = \lim_{N\to\infty} \frac{f_n}{N}$$

Quest'ultima definizione contiene uno strano limite, che non è un limite in senso matematico, ma sperimentale: si immagina di poter effettuare infinite misure. Con questa definizione si può dimostrare che i concetti relativi alle probabilità sono tutti consistenti, ma ci si trova in una situazione sperimentalmente impossibile da realizzare. In queste brevi note non esaminiamo ulteriormente il problema, ma si deve sapere che il problema esiste, e che è stato affrontato in vari modi. Per ora noi ci accontentiamo della nostra definizione ingenua, senza il limite, con la sola ipotesi che il numero di misure sia sufficientemente alto da produrre dei valori di probabilità abbastanza stabili.

Da queste definizioni si vede anche che un istogramma può essere trasformato in un istogramma normalizzato, dividendo tutti i valori di frequenza per *N*: allora le barre dell'istogramma rappresentano le probabilità, e l'istogramma diventa una *distribuzione di probabilità* sperimentale.

4. Un modello matematico di distribuzione di probabilità: la distribuzione binomiale

Se facciamo delle ipotesi sulla struttura degli eventi possiamo costruire modelli matematici delle probabilità che osserviamo.

Un modello semplice ed estremamente importante è la cosiddetta distribuzione binomiale. In questo caso studiamo la probabilità di eventi composti da eventi elementari: gli eventi elementari sono di due tipi, A e B, mutuamente incompatibili (si verifica A oppure B, e questo esaurisce tutte le possibilità), e gli eventi composti sono sequenze di A e B.

Esaminiamo un esempio importante, quello di un esperimento in cui lanciamo una moneta. Il singolo lancio della moneta dà due eventi tra loro incompatibili, testa e croce, ciascuno con una sua probabilità p_T e p_C (probabilità che sono normalizzate: $p_T + p_C = 1$), ma il nostro esperimento consiste ora in una serie di *n* lanci concatenati, e ci chiediamo quale sia la probabilità di avere un evento con *m* teste e *n-m* croci. Ci sono ovviamente $\binom{n}{m}$ modi in cui si possono posizionare le *m* teste all'interno della sequenza di *n* lanci, e ciascuna di queste sequenze (indipendentemente dall'ordine) ha la probabilità $p_T^m p_C^{n-m}$, e per l'incompatibilità di queste sequenze, la probabilità è la somma delle probabilità delle singole sequenze, e quindi

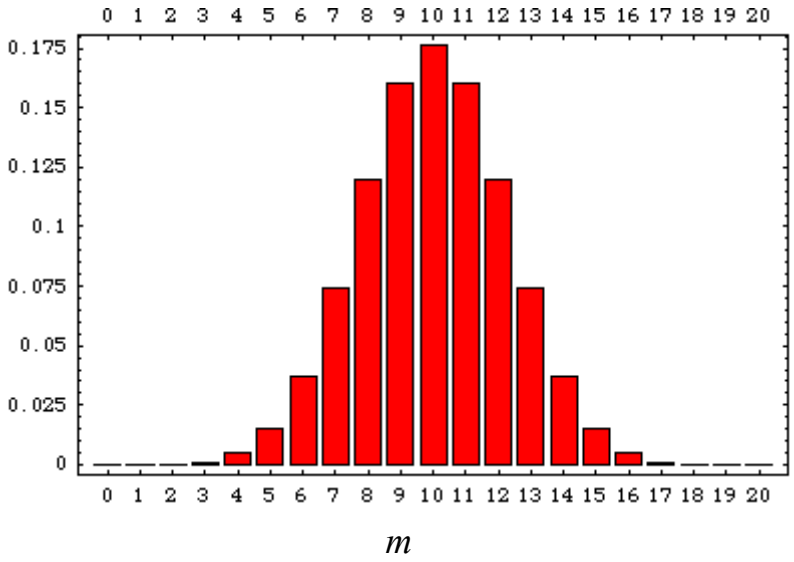
$$P(m) = \binom{n}{m} p_T^m p_C^{n-m}$$

Queste probabilità sono già correttamente normalizzate, ed infatti si trova:

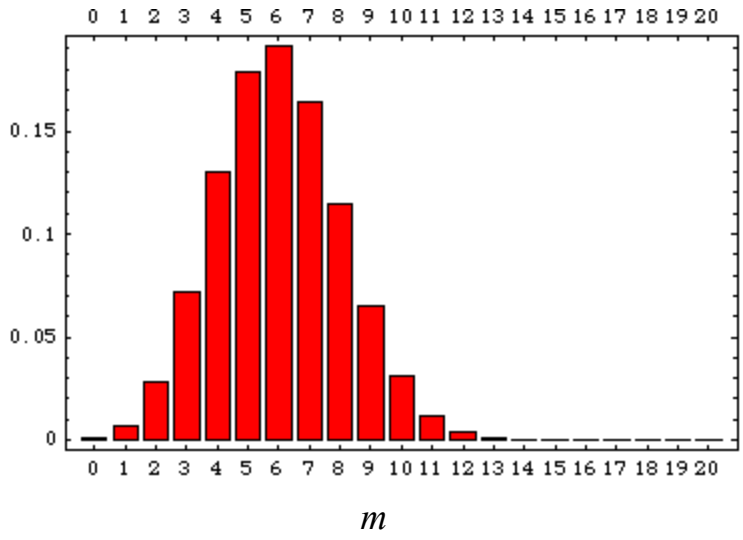
$$\sum_{m=0}^{m=n} P(m) = \sum_{m=0}^{m=n} \binom{n}{m} p_T^m p_C^{n-m} = (p_T + p_C)^n = 1^n = 1$$

Per quest'ultima formula è stato utilizzata l'espansione del binomio di Newton, e la distribuzione di probabilità è detta appunto *distribuzione binomiale*.

La figura seguente mostra i valori di $P(m)$ per $p_T = p_C = 0.5$ e n = 20



mentre la figura seguente mostra i valori di $P(m)$ per $p_T = 0.3, p_C = 0.7$ e n = 20.



Quando eseguiamo l'esperimento di lancio multiplo della moneta noi non possiamo prevedere con esattezza il risultato del nostro esperimento, ma data la distribuzione binomiale, possiamo dire con che probabilità otterremo ciascun risultato possibile.

Nel caso di modelli matematici delle probabilità come la distribuzione binomiale, si possono calcolare esattamente gli indicatori statistici che abbiamo introdotto nella sezione sugli istogrammi. Ad esempio, il valore medio della distribuzione è dato da

$$\begin{aligned}
\langle m \rangle &= \frac{\sum_{m=0}^{m=n} m f_m}{\sum_{m=0}^{m=n} f_m} = \sum_{m=0}^{m=n} m P(m) \\
&= \sum_{m=0}^{m=n} m \binom{n}{m} p_T^m p_C^{n-m} = \sum_{m=1}^{m=n} m \frac{n!}{m!(n-m)!} p_T^m p_C^{n-m} \\
&= np_T \sum_{m=0}^{m=n} \frac{(n-1)!}{(m-1)!\big((n-1)-(m-1)\big)!} p_T^{m-1} p_C^{(n-1)-(m-1)} \\
&= np_T \sum_{m=1}^{m=n} \binom{n-1}{m-1} p_T^{m-1} p_C^{(n-1)-(m-1)} = np_T \sum_{m=0}^{m=n-1} \binom{n}{m} p_T^m p_C^{n-1-m} \\
&= np_T
\end{aligned}$$

(usiamo la parentesi triangolare per denotare i valori medi teorici: in questo caso il valore medio della quantità m, il numero di teste). Questo è il valore medio del modello matematico, che possiamo confrontare con il valore medio sperimentale.

Analogamente si trova la varianza teorica del modello binomiale. Prima di procedere notiamo che

$$\begin{aligned}
\operatorname{var} x &= \sum_n \big(x_n - \langle x \rangle\big)^2 p_n = \sum_n x_n^2 p_n - 2\langle x \rangle \sum_n x_n p_n + \langle x \rangle^2 \sum_n p_n \\
&= \langle x^2 \rangle - 2\langle x \rangle^2 + \langle x \rangle^2 \\
&= \langle x^2 \rangle - \langle x \rangle^2
\end{aligned}$$

Per questo motivo, per calcolare la varianza, cominciamo con il calcolare il valore medio di m^2:

$$\begin{aligned}\left\langle m^2\right\rangle &= \sum_{m=0}^{m=n} m^2\, P(m) \\ &= \sum_{m=0}^{m=n} m^2 \binom{n}{m} p_T^m p_C^{n-m} = \sum_{m=1}^{m=n} \left[m(m-1)+m \right] \frac{n!}{m!(n-m)!} p_T^m p_C^{n-m} \\ &= n(n-1) \sum_{m=2}^{m=n} \frac{(n-2)!}{(m-2)!(n-m)!} p_T^m p_C^{n-m} + \langle m \rangle \\ &= n(n-1) p_T^2 + n p_T\end{aligned}$$

e quindi

$$\operatorname{var} m = \left(n(n-1) p_T^2 + n p_T \right) - \left(n p_T \right)^2 = n p_T - n p_T^2 = n p_T p_C$$

5. Variabili aleatorie continue: la distribuzione Gaussiana

Fino ad ora abbiamo sempre considerato variabili aleatorie che possono assumere solo certi valori ben definiti, discreti, ma nel caso di molte misure fisiche le variabili aleatorie possono variare con continuità. Possiamo discretizzare i valori di una variabile aleatoria x continua introducendo una variabile aleatoria discreta x_k tale che il valore $x_k = k\Delta x$ corrisponde all'insieme di valori $\left(k\Delta x - \Delta x/2 , k\Delta x + \Delta x/2 \right)$, dove Δx è la larghezza dell'intervallo di discretizzazione. A questa variabile aleatoria possiamo associare la probabilità p_k, ma a questo punto possiamo introdurre anche una densità di probabilità, definita da

$$p(x) = \frac{dP}{dx} = \lim_{\Delta x \to 0} \frac{p_k}{\Delta x}$$

In altre parole, la prob. di trovare la variabile aleatoria continua x nell'intervallo $(x, x+dx)$ è

$$dP(x) = p(x)\, dx$$

Qui nel seguito studiamo una densità di probabilità particolarmente importante, la densità di probabilità Gaussiana.

Supponiamo che l'incertezza statistica sia dovuta all'effetto combinato di molti processi fisici, e per semplificare i calcoli, supponiamo che ciascuno di questi abbia esattamente la stessa distribuzione di probabilità. In particolare, supponiamo che l'incertezza statistica sia dovuta alla somma di un gran numero di effetti che possono far aumentare o diminuire il valore osservato sempre della stessa (piccola) quantità.
Se questo è vero, allora la deviazione dal valore medio è data dalla somma

$$x = \sum_{i=1}^{N} e_i$$

dove le e_i sono delle variabili aleatorie indipendenti che possono assumere con uguale probabilità i valori $\pm\varepsilon$. Perciò la deviazione dal valore medio è una variabile aleatoria con distribuzione binomiale, e la probabilità di osservare una deviazione dovuta a k errori (microscopici) positivi ed N-k errori negativi $x = k\varepsilon - (N-k)\varepsilon = (2k-N)\varepsilon$, è

$$\begin{aligned}
\ln P(x) = \ln P(k) &= \ln\left[2^{-N}\binom{N}{k}\right] = \ln\left[2^{-N}\frac{N!}{(N-k)!k!}\right] \\
&\approx -N\ln 2 + \left(N\ln N - N\right) - \left[(N-k)\ln(N-k) - (N-k)\right] - (k\ln k - k) \\
&= -N\ln 2 + N\ln N - (N-k)\ln(N-k) - k\ln k \\
&= -N\ln 2 + N\ln N - \frac{1}{2}\left(N - \frac{x}{\varepsilon}\right)\ln\left[\frac{1}{2}\left(N - \frac{x}{\varepsilon}\right)\right] - \frac{1}{2}\left(N + \frac{x}{\varepsilon}\right)\ln\left[\frac{1}{2}\left(N + \frac{x}{\varepsilon}\right)\right] \\
&= -\frac{1}{2}\left(N - \frac{x}{\varepsilon}\right)\ln\left(1 - \frac{x}{N\varepsilon}\right) - \frac{1}{2}\left(N + \frac{x}{\varepsilon}\right)\ln\left(1 + \frac{x}{N\varepsilon}\right) \\
&\approx \frac{1}{2}\left(N - \frac{x}{\varepsilon}\right)\frac{x}{N\varepsilon} - \frac{1}{2}\left(N + \frac{x}{\varepsilon}\right)\frac{x}{N\varepsilon} = -\frac{x^2}{N\varepsilon^2}
\end{aligned}$$

dove si è usata l'approssimazione di Stirling (v. appendice), l'equivalenza data dalla formula $k = \frac{1}{2}\left(N + \frac{x}{\varepsilon}\right)$, e l'espansione in serie di Taylor del logaritmo naturale (troncata al primo ordine). Dunque, esponenziando la formula per la probabilità, si ottiene:

$$P(x) \propto \exp\left(-\frac{x^2}{N\varepsilon^2}\right)$$

dove si è utilizzato il simbolo di proporzionalità per mettere l'accento sul fatto che la probabilità appena calcolata non è normalizzata. Si noti che questa è anche la probabilità che x appartenga all'intervallo $[(2k-N)\varepsilon,(2k+2-N)\varepsilon)$ (l'unico valore che può assumere è quello che corrisponde all'estremo sinistro dell'intervallo). Introduciamo ora la variabile aleatoria $y = x/\sqrt{N}$, così che la probabilità scritta sopra è proprio la probabilità che y appartenga all'intervallo $\left[(2k-N)\varepsilon/\sqrt{N},(2k+2-N)\varepsilon/\sqrt{N}\right)$, e dunque introducendo un opportuno fattore di normalizzazione A, possiamo scrivere una densità di probabilità:

$$\frac{dP}{dy} = A\exp\left(-\frac{y^2}{\varepsilon^2}\right)$$

che è nota con il nome di densità di probabilità Gaussiana.

L'integrale di normalizzazione della Gaussiana si trova con un trucco semplicissimo. Noi vogliamo calcolare

$$I = \int_{-\infty}^{+\infty} A\exp\left(-\frac{x^2}{2\sigma^2}\right)dx$$

e invece di affrontare direttamente l'integrazione, calcoliamo il quadrato di I:

$$\begin{aligned} I^2 &= \int_{-\infty}^{+\infty} A\exp\left(-\frac{x^2}{2\sigma^2}\right)dx \cdot \int_{-\infty}^{+\infty} A\exp\left(-\frac{y^2}{2\sigma^2}\right)dy \\ &= A^2 \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \end{aligned}$$

Ora si passa dalle coordinate cartesiane alle coordinate radiali con la trasformazione

$\begin{cases} x = r\cos\vartheta \\ y = r\sin\vartheta \end{cases}$ e quindi $dx\,dy = r\,dr\,d\vartheta$ e $x^2+y^2 = r^2$. Perciò

$$I^2 = A^2 \int_0^{2\pi} d\vartheta \int_0^{\infty} r\,dr \exp\left(-\frac{r^2}{2\sigma^2}\right) = 2\pi A^2 \int_0^{\infty} r\,dr \exp\left(-\frac{r^2}{2\sigma^2}\right)$$
$$= \pi A^2 \int_0^{\infty} d(r^2) \exp\left(-\frac{r^2}{2\sigma^2}\right) = 2\pi\sigma^2 A^2$$

Per normalizzare la probabilità questo deve valere 1, e quindi $A = \frac{1}{\sqrt{2\pi\sigma^2}}$.

Esercizi:

1. dare una rappresentazione geometrica della trasformazione di coordinate.
2. calcolare il valore medio della densità di probabilità $\frac{dP}{dy} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right)$
3. calcolare la varianza della densità di probabilità $\frac{dP}{dy} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right)$

Noi abbiamo introdotto la distribuzione Gaussiana con delle ipotesi piuttosto restrittive, ma in realtà la si può ottenere con delle ipotesi molto più generali, e questa particolare distribuzione compare in innumerevoli casi di interesse fisico.

6. Trattazione matematica della "propagazione degli errori".

Quando si eseguono delle misure si ottengono certi valori per delle variabili fisiche. D'altra parte noi potremmo essere interessati solo indirettemente a queste variabili e potremmo invece voler valutare una loro funzione: come si fa a stimare l'incertezza statistica di questa funzione a partire dalle incertezze statistiche delle variabili osservate? Il metodo matematico che si utilizza per stimare questa nuova incertezza è detto *propagazione degli errori*. Per capire esattamente come funzioni cominciamo con il dimostrare un teorema relativo alla varianza di una somma di variabili dipendenti.

Due variabili aleatorie X e Y sono indipendenti se e solo se la probabilità di osservare due valori particolari valori x e y è uguale a $P_X(x)P_Y(y)$, dove $P_X(x)$ è la probabilità di

osservare il valore x estraendo a caso dalla distribuzione di probabilità della variabile aleatoria X (e analoga per Y). Allora, se le due variabili aleatorie sono variabili continue, la probabilità di osservare dei valori compresi tra x e $x+dx$ e y e $y+dy$ è data da $dP_X(x)dP_Y(y)$. Questo significa che il valore medio della somma delle due variabili aleatorie è

$$\begin{aligned}\langle x+y\rangle &= \int_{D_X}\int_{D_Y}(x+y)\,dP_X\,dP_Y \\ &= \int_{D_X} x\,dP_X \int_{D_Y} dP_Y + \int_{D_X} dP_X \int_{D_Y} y\,dP_Y \\ &= \int_{D_X} x\,dP_X + \int_{D_Y} y\,dP_Y = \langle x\rangle + \langle y\rangle\end{aligned}$$

dove D_X e D_Y sono i dominii su cui sono definite le variabili aleatorie X e Y e si è fatto uso degli integrali di normalizzazione

$$\int_{D_X} dP_X = \int_{D_Y} dP_Y = 1$$

Il valore quadratico medio si calcola nel modo seguente:

$$\begin{aligned}\left\langle (x+y)^2\right\rangle &= \int_{D_X}\int_{D_Y}(x+y)^2\,dP_X\,dP_Y \\ &= \int_{D_X}\int_{D_Y}(x^2+2xy+y^2)\,dP_X\,dP_Y \\ &= \int_{D_X} x^2\,dP_X \int_{D_Y} dP_Y + 2\int_{D_X} x\,dP_X \int_{D_Y} y\,dP_Y + \int_{D_X} dP_X \int_{D_Y} y^2\,dP_Y \\ &= \left\langle x^2\right\rangle + 2\langle x\rangle\langle y\rangle + \left\langle y^2\right\rangle\end{aligned}$$

Perciò la varianza della somma è

$$\begin{aligned}\mathrm{Var}(x+y) &= \left\langle (x+y)^2\right\rangle - \left(\langle x\rangle + \langle y\rangle\right)^2 \\ &= \left[\left\langle x^2\right\rangle + 2\langle x\rangle\langle y\rangle + \left\langle y^2\right\rangle\right] - \left[\langle x\rangle^2 + 2\langle x\rangle\langle y\rangle + \langle y\rangle^2\right] \\ &= \left(\left\langle x^2\right\rangle - \langle x\rangle^2\right) + \left(\left\langle y^2\right\rangle - \langle y\rangle^2\right) = \mathrm{Var}\,x + \mathrm{Var}\,y\end{aligned}$$

In altre parole se le variabili sono indipendenti la varianza della somma è uguale alla somma delle varianze. Si noti che se le variabili aleatorie non fossero indipendenti il valore medio del prodotto calcolato in (3) non sarebbe uguale al prodotto delle medie, e quindi la varianza (4) sarebbe diversa dalla somma delle varianze.
È chiaro che il teorema si estende in modo ovvio alla somma di più di due variabili indipendenti.

Adesso consideriamo una funzione $f = f(x_1,\ldots,x_n)$, e notiamo che per piccole variazioni rispetto il valore medio delle variabili indipendenti possiamo scrivere:

$$f(\overline{x}_1 + \Delta x_1,\ldots,\overline{x}_n + \Delta x_n) \approx f(\overline{x}_1,\ldots,\overline{x}_n) + \sum_{i=1}^{i=n} \left.\frac{\partial f}{\partial x_i}\right|_{\{x_i=\overline{x}_i\}} \Delta x_i$$

Assumiamo che le variabili Δx_i siano indipendenti, allora, tenendo anche conto del fatto che il loro valore medio è 0 (si spieghi perché....), il valore medio e la varianza della funzione *f* sono

$$\overline{f} \approx f(\overline{x}_1,\ldots,\overline{x}_n)$$

e

$$\mathrm{Var}\, f \approx \sum_{i=1}^{i=n} \left(\left.\frac{\partial f}{\partial x_i}\right|_{\{x_i=\overline{x}_i\}} \right)^2 \sigma_i^2$$

dove σ_i^2 è la varianza relativa all'i-esima variabile.
Se le variabili non sono indipendenti, è possibile ottenere una formula simile a quella scritta sopra, anche se più complicata.

Esercizio: date due variabili aleatorie indipendenti *x* e *y*, con varianza σ_x^2 e σ_y^2, si calcolino le varianze delle seguenti funzioni:

a. $x + y$

b $x - y$

c. $x \cdot y$

d. x / y

e. $x^{\alpha} y^{\beta}$

f. $\sin(x \cdot y)$

Infine si calcoli la deviazione standard del valore medio definito da $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, dove N è il numero totale di misure della variabile aleatoria x, supponendo che la varianza di ciascuna misura sia sempre la stessa e valga σ^2 .

7. Il metodo dei minimi quadrati

Molto spesso in Fisica, si vogliono trovare i parametri che caratterizzano un modello fisico estraendoli dai dati. Per questo vengono utilizzati molti metodi statistici.

In molti casi il modello è un semplice modello lineare, del tipo $y = ax + b$, dove x è una variabile indipendente, ed y è una variabile dipendente, funzione di x e dei parametri a e b. Nell'esperimento, in corrispondenza a valori fissati x_i della variabile indipendente, noi osseviamo dei valori y_i della variabile dipendente con deviazione standard σ_i, e vogliamo determinare i parametri a e b che caratterizzano il nostro modello teorico.

Dati dei valori prefissati a e b, il nostro modello lineare prevede che in corrispondenza a x_i noi osserviamo $y_i^{(teor.)} = ax_i + b$. La differenza tra questi valori e quelli osservati è $\Delta y_i = y_i - y_i^{(teor.)} = y_i - (ax_i + b)$. Questa distanza tra valore osservato e valore calcolato non è molto significativa se l'incertezza della misura è grande, mentre è molto significativa se l'incertezza della misura è piccola, perciò questa distanza viene pesata per mezzo della deviazione standard. Inoltre la distanza complessiva tra valori osservati e valori calcolati ha senso se i contributi dovuti alle singole misure sono tutti additivi, e quindi si ottiene una specie di distanza tra valori osservati e valori calcolati sommando i quadrati delle differenze pesate, e cioè

$$S = \sum_{i=1}^{N} \frac{\left[y_i - (ax_i + b)\right]^2}{\sigma_i^2}$$

I parametri che meglio corrispondono ai dati sono quelli che minimizzano questa strana distanza: per questo si devono trovare le derivate di S rispetto i parametri ed uguagliarle a zero:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum_{i=1}^{N} x_i \dfrac{\left[y_i - (ax_i + b)\right]}{\sigma_i^2} = 0 \\ \dfrac{\partial S}{\partial b} = -2\sum_{i=1}^{N} \dfrac{\left[y_i - (ax_i + b)\right]}{\sigma_i^2} = 0 \end{cases}$$

Si ottiene allora un sistema lineare 2×2 rispetto ai parametri *a* e *b*:

$$\begin{cases} a\sum_{i=1}^{N} \dfrac{x_i^2}{\sigma_i^2} + b\sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} = \sum_{i=1}^{N} \dfrac{x_i y_i}{\sigma_i^2} \\ a\sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} + b\sum_{i=1}^{N} \dfrac{1}{\sigma_i^2} = \sum_{i=1}^{N} \dfrac{y_i}{\sigma_i^2} \end{cases}$$

Questo sistema si risolve facilmente, e si trova:

$$\begin{cases} a = \dfrac{\sum_{i=1}^{N} \dfrac{x_i y_i}{\sigma_i^2} \sum_{i=1}^{N} \dfrac{1}{\sigma_i^2} - \sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \dfrac{y_i}{\sigma_i^2}}{\sum_{i=1}^{N} \dfrac{x_i^2}{\sigma_i^2} \sum_{i=1}^{N} \dfrac{1}{\sigma_i^2} - \left(\sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} \right)^2} \\ b = \dfrac{\sum_{i=1}^{N} \dfrac{x_i^2}{\sigma_i^2} \sum_{i=1}^{N} \dfrac{y_i}{\sigma_i^2} - \sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} \sum_{i=1}^{N} \dfrac{x_i y_i}{\sigma_i^2}}{\sum_{i=1}^{N} \dfrac{x_i^2}{\sigma_i^2} \sum_{i=1}^{N} \dfrac{1}{\sigma_i^2} - \left(\sum_{i=1}^{N} \dfrac{x_i}{\sigma_i^2} \right)^2} \end{cases}$$

Esercizio: si calcolino le varianze dei parametri *a* e *b* ottenuti per mezzo del metodo dei minimi quadrati, utilizzando il metodo della propagazione degli errori.

Appendice: L'approssimazione di Stirling

Questa è un'approssimazione molto utile in Meccanica Statistica, ed in generale nel calcolo combinatorio: si tratta di un'approssimazione della funzione fattoriale. Noi qui dimostriamo una versione "minima" dell'approssimazione, che però è sufficientemente buona per la maggior parte delle applicazioni. Prendiamo il logaritmo del fattoriale e sottoponiamolo ad alcune semplici trasformazioni:

$$\ln N! = \ln \prod_{n=1}^{n=N} n = \sum_{n=1}^{n=N} \ln n \approx \int_1^N \ln x \, dx = \left(x \ln x - x \right) \Big|_1^N = N \ln N - N + 1$$

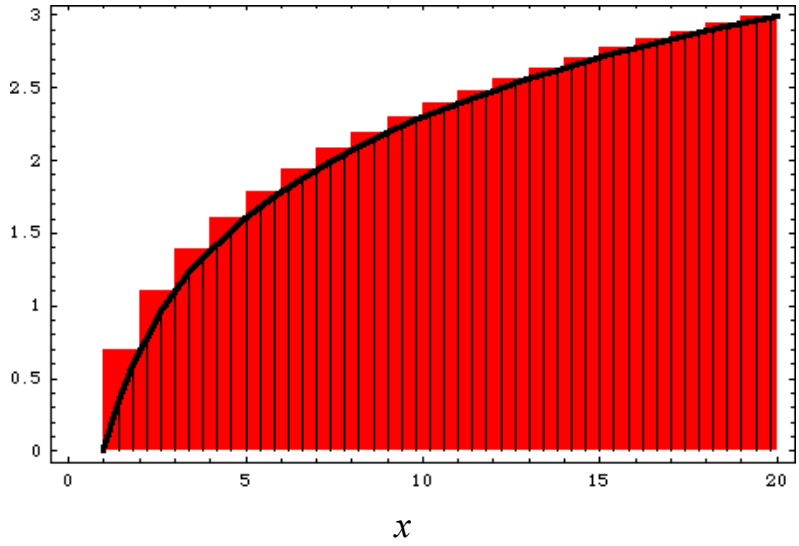
il passo "critico" è l'approssimazione della somma con un integrale. Questa formula può anche essere scritta nella forma

$$N! \approx N^N e^{-(N-1)}$$

La tabella seguente confronta il logaritmo del fattoriale calcolato esattamente con l'approssimazione di Stirling; si vede che l'approssimazione segue molto bene il risultato esatto, a parte una differenza costante (che corrisponde ad un fattore fisso che abbiamo trascurato)

n	$\ln n!$	$n \ln n - n + 1$
1	0.	0.
2	0.693147	0.386294
3	1.79176	1.29584
4	3.17805	2.54518
5	4.78749	4.04719
6	6.57925	5.75056
7	8.52516	7.62137
8	10.6046	9.63553
9	12.8018	11.775
10	15.1044	14.0259

Il grafico seguente mostra graficamente la differenza tra la somma e l'integrale: la somma che si cerca è l'area segnata in rosso, mentre l'integrale corrisponde all'area tratteggiata. La differenza corrisponde alla somma dei triangolini che sono stati trascurati.



L'area del k-esimo triangolino è approssimativamente $\frac{1}{2}\ln\frac{k+1}{k}$, quindi la differenza tra le due aree è

$$\sum_{k=1}^{k=n-1}\frac{1}{2}\ln\frac{k+1}{k}\approx\frac{1}{2}\sum_{k=1}^{k=n-1}\frac{1}{k}\approx\frac{1}{2}\ln n$$

e sommando questa differenza alla formula precedente si ottiene una migliore approssimazione del fattoriale

$$\ln N!\approx N\ln N-N+1+\frac{1}{2}\ln N$$

Ora si può costruire una tabella come prima:

n	$\ln n!$	$n\ln n$-n+1+1/2
1	0.	0
2	0.693147	0.732868
3	1.79176	1.84514
4	3.17805	3.23832
5	4.78749	4.85191
6	6.57925	6.64644
7	8.52516	8.59433
8	10.6046	10.6753
9	12.8018	12.8736
10	15.1044	15.1771

e si vede che la nuova approssimazione $N!\approx N^{N+1/2}e^{-(N-1)}$ è significativamente migliore di quella precedente.