Quale matematica per i fenomeni casuali?

I primi strumenti per descrivere posizione e dispersione dei dati

0. Introduzione

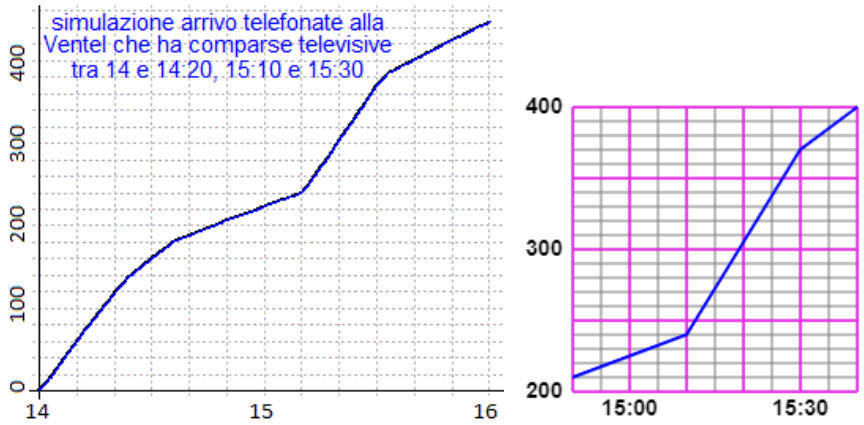
Riprendiamo e approfondiamo lo studio dei fenomeni casuali, che abbiamo già avviato nel biennio, in particolare nella → scheda 3 de *Le statistiche* e nella scheda → *Calcolo delle probabilità.* Prima di proseguire rileggi queste schede.

1. Una situazione problematica

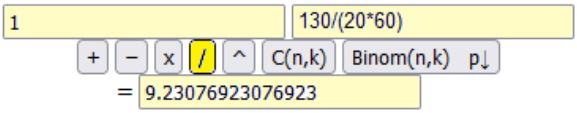
L'organizzazione di vendite televisive Ventel utilizza le strutture e il personale (centraliniste) di una agenzia specializzata (che offre i suoi servizi a diverse organizzazioni di vendita) per ricevere ordinazioni telefoniche tra le 14 e le 16. Le trasmissioni della Ventel vanno in onda tra le 14 e le 14:20 e tra le 15:10 e le 15:30.

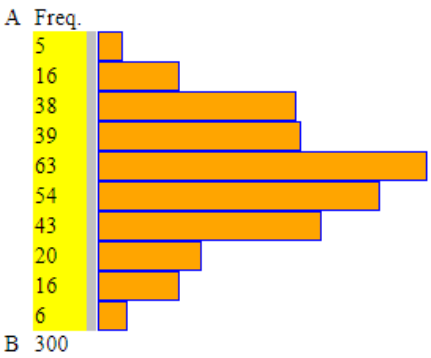
La Ventel vuole studiare quante linee (e centraliniste) conviene richiedere alla agenzia. Il servizio non prevede liste di attesa: se non c'è una linea libera il potenziale acquirente trova occupato. Per fare questo studio la Ventel chiede alla ditta specializzata Telstat di studiare i tempi di arrivo delle telefonate (la Telstat è in grado di individuare anche le telefonate che, arrivate al centralino, trovano occupato) e le durate delle telefonate che riescono a prendere la linea.

Possiamo simulare i rilevamenti effettuati dalla Telstat mediante alcuni programmi. Ecco, sotto, l'esito della "media" di una ventina di queste simulazioni, che, rispetto al fenomeno effettivo, assume un andamento abbastanza "liscio". A destra un ingrandimento.



1 Qual è l'intervallo di tempo in cui le telefonate arrivano più frequentemente? Perchè dai grafici posso ricavare che in questo intervallo vi sono circa una telefonata ogni 9 secondi?



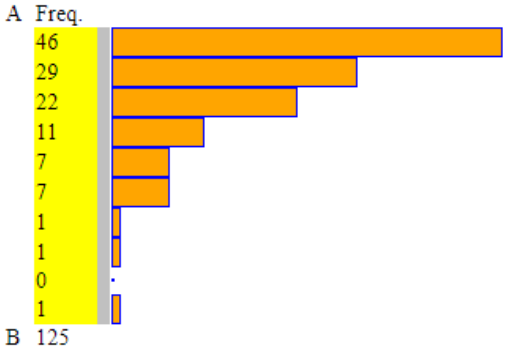


2 Occupiamoci, ora, della **durata** delle telefonate. La ditta Telstat, studiata la situazione, simula il rilevamento della durata delle telefonate. **Qui** → puoi accedere a uno script che genera una simulazione; copia gli esiti senza preocuuparti del codice dello script. Poi incollali nello script **Istogramma**. Ottieni esiti simili a quello a fianco (la forma dell'istogramma e i valori prodotti possono leggermente cambiare). La durata media di una telefonata (vedi "mean") è di circa 50 sec. Quante linee telefoniche sono necessarie per non perdere telefonate? Perchè?

A = 0 B = 100 intervals = 10 their width = 10 n=300 min=4.419016504560972
max=97.67555822104947 median=48.1385439715408
1^|3^ quartile = 35.32395609246026 | 62.33038200142828 mean=48.75626134811374

Nell'ipotesi che arrivi e durate delle telefonate abbiano esattamente questo regime, sembra che basti questo numero di linee: riesco infatti a prendere telefonate che arrivino ogni 9 secondi e che durino fino a 54 secondi (9·6 = 54), e 54>50. In altre parole, se si misura il tempo a partire dalla 1ª telefonata, al 9° sec arriva la 2ª telefonata e occupa la seconda linea, …, al 45° sec arriva la 6ª telefonata e occupa la 6ª linea, cioè l'ultima linea rimasta libera; al 50° sec si libera la prima linea, per cui la 7ª telefonata che arriva al 54° sec trova una linea in cui inserirsi; al 59° sec si libera la seconda linea, per cui …; e così via.

Ma, da una parte, possono capitare telefonate che durano meno della durata media e telefonate che durano di più, per cui possono rimanere delle linee libere o, viceversa, si possono perdere delle telefonate. D'altra parte anche il tempo tra una telefonata e la successiva non è sempre 9 secondi: anch'esso è variabile.



3 Se simulo con uno script **come** → questo i **tempi di arrivo** tra una telefonata e l'altra e poi li incollo nello script **Istogramma** ottengo uscite grafiche e numeriche simili a quelle qui, a sinistra e sotto, riprodotte. Quali sono le differenze principali tra i dati precedenti (durate delle telefonate) e questi (distanze temporali tra arrivi successivi)?

A = 0 B = 50 intervals = 10 their width = 5 n=125 min=0.010155104203901371
max=47.16979660110787 median=6.7132674552586
1^|3^ quartile = 2.6856144571588216 | 13.515376050890003 mean=9.7953528950112

La soluzione che abbiamo ottenuto nel quesito 1 non teneva conto della ***casualità*** dei tempi che passano tra una telefonata e la successiva e dei tempi di durata delle telefonate. Avevamo, infatti, erroneamente, schematizzato la situazione con un ***modello deterministico***: utilizzando i valori medi prevedevamo esattamente come al passare del tempo si sarebbe modificato lo stato del centralino.

La ***media aritmetica***, per il nostro problema, non è un concetto matematico sufficiente a caratterizzare tempi di arrivo e durate delle telefonate. Vediamo di individuare strumenti matematici più efficaci per i nostri scopi. Vedremo poi, più avanti, come è possibile approssimare istogrammi come i precedenti con i grafici di opportune funzioni.

2. Indici di posizione e di dispersione

Data una sequenza di informazioni di tipo numerico, eventualmente già classificate, i suoi *valori medi* (media, moda e mediana) vengono chiamati anche ***indici di posizione*** in quanto indicano, con diverse caratterizzazioni, la zona dell'asse numerico in cui tali dati cadono con maggiore frequenza.

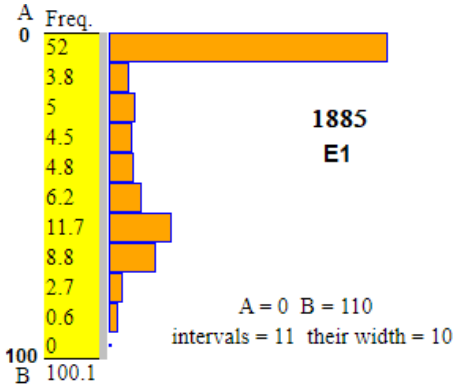
Abbiamo già osservato che il confronto tra i diversi indici di posizione può dare anche indicazioni sulla forma dell'istogramma di distribuzione. Ad esempio affinché la rappresentazione grafica sia simmetrica rispetto a un asse verticale è necessario (non sufficiente) che media e mediana coincidano. Invece se la rappresentazione grafica è più o meno a forma di campana ma allungata verso destra [sinistra], la media è maggiore [minore] della mediana.

Una ***interpretazione fisica*** del fenomeno è che la ***mediana*** rappresenta l'ascissa in cui praticare un taglio verticale che divida l'istogramma in due parti di area uguale, mentre la ***media*** è l'ascissa del *baricentro* dell'istogramma, ossia del punto dell'asse orizzontale per cui appenderlo in modo che, capovolto, rimanga con la base orizzontale.

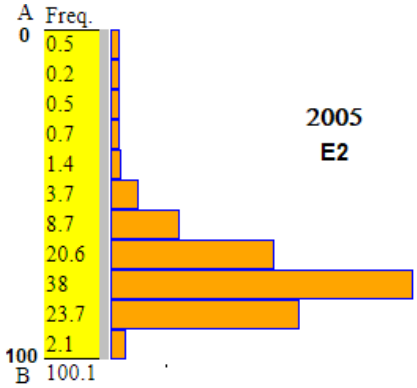
Nella figura riprodotta sotto sono rappresentate le distribuzioni delle età dei morti in Italia nel 1885 e, 120 anni dopo, nel 2005; indichiamole E1 ed E2.

Le rispettive *medie* sono, approssimativamente, 28 e 81: un morto nel 1885 aveva mediamente 29 anni e 81 nel 2005. Usando **M** per indicare la media: M(E1) = 28 e M(E2) = 81. Le età *mediane* di morte sono invece, approssimativamente, 5 e 85: Mediana(E1) = 5 e Mediana(E2) = 85. Abbiamo aggiunto "approssimativamente" in quanto abbiamo usato dati già classificati.

Il fatto che, nel 1885, la media abbia un valore molto maggiore della mediana (mascherando in parte il fenomeno della mortalità infantile) è dovuto alla lunga coda verso 100 che fa aumentare il risultato del calcolo della media. Nel 2006, invece, la media è inferiore alla mediana a causa della coda verso 0; la differenza in questo caso è lieve in quanto si tratta di una coda molto "sottile", e quindi non incide molto sul risultato.



1°quartile = 5 median = 5 3°quartile = 55
mean = 28.6064



1°quartile = 75 median = 85 3°quartile = 95
mean = 81.1339

Sono chiamati ***indici di dispersione*** alcuni indicatori numerici che danno un'idea quantitativa di come i dati sono più o meno sparpagliati. Riferiamoci alle stesse distribuzioni considerate sopra.

In 120 anni, oltre a uno spostamento verso 100 della zona in cui si concentrano le età di morte (testimoniato dall'aumento sia della media che della mediana), possiamo osservare un maggiore addensamento dei dati: l'istogramma assume una forma più tozza. Questa percezione intuitiva può essere precisata considerando *l'intervallo in cui si colloca il 50% centrale dei dati,* ossia i dati che vanno dal 25° al 75° percentile, ossia dal 1° al 3° quartile: da circa [5, 55] (il 25% dei morti aveva età che non superava i 5 anni e il 75% età che non superava i 55 anni) passa a circa [75, 95].

Questi valori sono scritti sotto agli istogrammi. Qui a destra sono rappresentati anche in un *box-plot* quelli relativi al 2005. La ampiezza dell'intervallo tra 1° e 3° quartile viene chiamata ***distanza interquartile*** e viene in genere indicata con il simbolo *IQR* (InterQuartile Range). Questo è l'indice di dispersione più usato.

Qui → vedi come ottenere le precedenti rappresentazioni.

4 Nei 120 anni considerati sopra la distanza interquartile passa da circa … a circa …

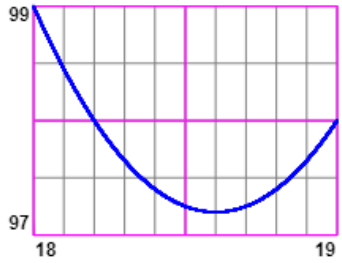
Un altro modo per valutare la dispersione di una sequenza di N dati $x_1, x_2, \dots x_N$ può essere quello di quantificare opportunamente il loro livello di concentrazione attorno a un indice di posizione **p**. Potremmo valutare gli scarti $x-\mathbf{p}$ dei singoli dati da **p** e farne la media, ma in questo modo scarti positivi e negativi si compenserebbero tra di loro. Per evitare ciò possiamo considerare la media **mQ** dei loro quadrati.

Consideriamo ad esempio i dati 13, 15, 18, 22, 25:

5 Osserva le seguenti uscite (il grafico è stato realizzato con **questo** script). Che cosa puoi notare? Quale potrebbe essere il valore di **p** per cui **mQ** è minimo?

```
function F(x) { y= x*x; return y }
function f(x) { y= F(13-x)+F(15-x)+F(18-x)+F(22-x)+F(25-x); return y}
aX = 18; bX = 19; aY = 97; bY = 99
Dx = 0.1; Dy = 1/2
```

```
(13+15+18+22+25)/5 = 18.6
```



In effetti si può dimostrare (vedi l'esercizio **e10**) che *la media dei quadrati degli scarti da **p** è minima* quando ***p** è la media dei dati*. Quindi posso considerare questo valore come un indice della dispersione dei dati attorno alla media. Esso viene chiamato *varianza*. In altre parole, per N dati $x_1, \dots, x_N$ di *media* $\boldsymbol{\mu}$ ("μ" è la lettera greca "mu", o "mi"), si pone:

$$\textbf{\textit{varianza}} = \frac{(x_1-\boldsymbol{\mu})^2 + (x_2-\boldsymbol{\mu})^2 + \dots (x_N-\boldsymbol{\mu})^2}{N}$$

La *varianza* è quindi la *media dei "quadrati" degli scarti* dalla media. Per ottenere un valore con ordine di grandezza confrontabile con quello degli scarti dobbiamo applicare alla varianza la "radice quadrata", ossia considerare:

$$\textbf{\textit{scarto quadratico medio}} = \sqrt{\textbf{\textit{varianza}}} = \left(\frac{(x_1-\boldsymbol{\mu})^2 + (x_2-\boldsymbol{\mu})^2 + \dots (x_N-\boldsymbol{\mu})^2}{N}\right)^{1/2}$$

6 Qual è lo scarto quadratico medio dei cinque dati del quesito precedente?

Nelle formule useremo **Var** e **sqm** per indicare la varianza e lo scarto quadratico medio (lo scarto quadratico medio, come vedremo meglio in una scheda successiva, è chiamato anche ***deviazione standard** teorica*).

Se con la nostra **grande CT** introduco `13,15,18,22,25` e clicco **[sqm]** ottengo:

```
scarto quad. medio (sq.root of var./theoret.st.dev.) = 4.409081537009714
```

3. Notazioni

Per evitare di usare i puntini ("...") per descrivere una somma di un numero variabile di addendi si usa il simbolo **Σ** (detto ***sommatoria*** e costituito dalla lettera maiuscola greca "sigma"). Ecco un esempio:

$\sum_{n=1}^{10} n^2 = 385$ si legge "la somma di n^2 per n da 1 a 10 è uguale a 385" e abbrevia la scrittura:
$1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2+10^2 = 385$

Per comodità di scrittura a volte si usano notazioni più compatte: $\sum_{n=1..10} n^2$ o $\sum_n n^2$ (se è chiaro dal contesto quali siano il valore iniziale e quello finale dell'indice n). In casi in cui i dati sono pochi (1 o 2 decine) posso ricorrere alle "nostre" CT. Con la "piccola" introduco:

```
1+2*2+3*3+4*4+5*5+6*6+7*7+8*8+9*9+10*10
```

Con la "grande" introduco 1,2,3,4,5,6,7,8,9,10, poi clicco **[data^2]**, poi, copiato l'esito, clicco **[sum]**.

```
1, 2, 3, 4, 5, 6, 7, 8, 9, 10
1, 4, 9, 16, 25, 36, 49, 64, 81, 100
sum = 385
```

Se la distribuzione X ha $x_1, \dots, x_N$ come valori e $f_1,\dots,f_N$ come frequenze,
il totale dei dati è $f_1 + \dots + f_N$,
la somma totale dei valori è $x_1 \cdot f_1 + \dots + x_N \cdot f_N$
e la sua media M(X) può essere descritta con:

$$M(X) = \sum_{k=1}^{N} (x_k \cdot f_k) \,/\, \sum_{k=1}^{N} f_k \quad \text{o con:} \quad \mathbf{M(X)} = \sum_{\mathbf{k=1}}^{\mathbf{N}} \mathbf{(x_k \cdot fr_k)}$$

se $\mathbf{fr_k}$ indica la frequenza relativa del valore x_k: $fr_k = f_k$ / Totale, Totale = $\Sigma_k f_k$.

Ad es. se so che in un cineclub il 70% degli spettatori sono soci e hanno pagato 3 € mentre gli altri hanno pagato 5 €, posso dire che mediamente uno spettatore ha pagato: $3\cdot 70\% + 5\cdot 30\% = 3\cdot 0.7 + 5\cdot 0.3 = 2.1+1.5 = 3.6$ €.

Con la **grande CT** basta che introduca 3*50, 5*30 e clicchi **[mean]**:

```
3*70, 5*30
mean=3.6
```

La varianza è la media di $(X-\mu)^2$ dove $\mu = M(X)$. Ossia è la media di $(X - M(X))^2$. In modo compatto può essere descritta come **Var(X) = M((X – M(X))²)**.

Con la **grande CT** basta usare i tasti **[var]** e **[sqm]**:

```
3*70, 5*30
```

variance = 0.8399999999999999 (cioè 0.84)
scarto quad. medio (sq.root of var./st.dev.) = 0.9165151389911679

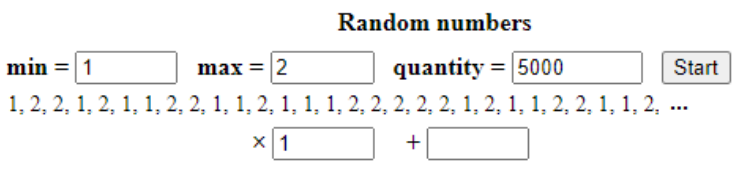
A = 0.5 B = 6.5 intervals = 6 their width = 1 n=5000 min=1 max=6 median=4 1^|3^ quartile=2|5 mean=3.491

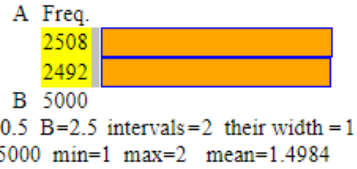
A = 0 B = 1 intervals = 10 their width = 0.1 n=3500 min=0.00003 max=0.99994 median=0.50177 1^|3^ quartile=0.24407|0.74124 mean=0.497698528571

4. Leggi di distribuzione (variabili discrete)

Nella scheda sul *Calcolo delle probabilità,* nel ➡ §3 e nel ➡ §4, abbiamo considerato sia variabili casuali che possono variare con continuità su tutto un intervallo di numeri reali, e che vengono dette ***variabili casuali continue***, sia variabili casuali che possono assumere solo valori "separati" l'uno dall'altro, elencabili in una successione, e che vengono dette ***variabili casuali discrete***.

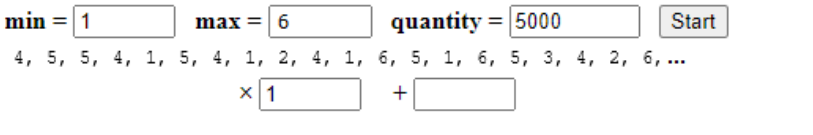
Ecco sotto gli esiti di una variabile casuale *discreta,* che simula il lancio di una moneta equa (1: testa, 2: croce). I valori sono stati generati con lo script **RandomNum** e poi analizzati con lo script **Istogramma**.

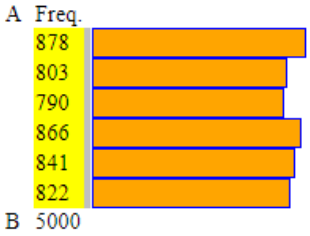




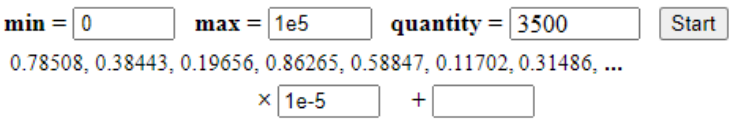
Con 100 prove le due colonne sarebbero di lunghezze molto diverse. Al crescere del numero delle prove (qui sono 5000) le due colonne hanno quasi la stessa altezza.

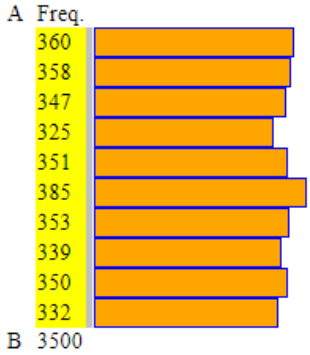
La simulazione, analoga, del lancio di un dado equo. Come si vede, il valor medio tende ad essere 3.5.





Sotto sono riportati gli esiti della generazione di molti numeri "reali" tra 0 ed 1. È una variabile *continua,* anche se nella simulazione è stata approssimata con un numero limitato (tra 0 e 0.99999). L'istogramma è stato suddiviso in 10 colonne, ma il numero di esse poteva essere diverso.





Prima il minimo delle uscite era 1 e il massimo era 2, o erano 1 e 6, ora sono 0.00003 e 0.99994. **RandomNum** genera numeri tra min e max, a meno che non siamo moltiplicati o aumentati utilizzando i box in fondo. Nel terzo caso abbiamo, per avere uscite non intere fra 0 e 1 con 5 cifre dopo il "." abbiamo messo **1e5** come max e poi abbiamo moltiplicato per **1e-5**.

Osserviamo che una variabile casuale discreta può essere ***non finita***. Pensiamo al numero **N** dei ***lanci di una moneta equa da effettuare fino ad ottenere l'uscita di "testa"*** (T).

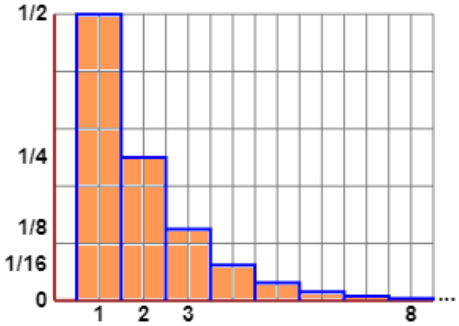
• Al 50% N=1, ossia viene T al primo lancio: **Pr**(N=1) = 1/2.

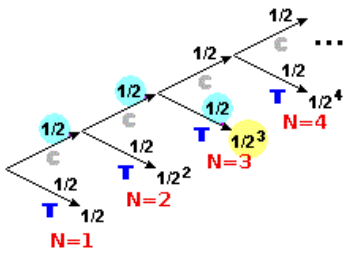
• La probabilità che venga T si mantiene la stessa nei lanci successivi, ma via via, ovviamente, rispetto all'inizio dei lanci essa si dimezza (vedi grafo sotto a destra): **Pr**(N=2) = (1/2)/2 = 1/4 = 25%.

• La probabilità **Pr**(N=3) che T venga al terzo lancio è (1/2)(1/2)(1/2) = $1/2^3$ = 1/8 = 12.5%.

• In generale: **Pr**(N = h) = $1/2^h$

A sinistra è tracciata parte dell'istogramma di distribuzione di N: è un esempio di ***figura illimitata*** (la base dell'istogramma prosegue senza fine a destra) con ***area finita*** (uguale a 1).





Nel caso statistico la ***media*** di una distribuzione X la possiamo ottenere sommando i prodotti dei valori x_k per le loro frequenze relative fr_k (corrispondenti alle aree delle colonne dell'istogramma sperimentale), nel caso di una variabile casuale X che possa assumere i valori x_1, x_2, … faremo analogamente la somma dei prodotti dei valori x_k per le loro probabilità $\mathbf{Pr}(X = x_k)$ (corrispondenti alle aree delle colonne dell'istogramma teorico):

$$\mathbf{M(X)} = \Sigma_k (x_k \cdot fr_k) \quad \text{diventa} \quad \mathbf{M(X)} = \Sigma_k (x_k \cdot \mathbf{Pr}(X = x_k))$$

La media di una variabile casuale X a volte viene chiamata anche *speranza matematica* o ***valore atteso*** ("expected value" in inglese) di X, e indicata **E**(X).

Qual è la media nel caso del numero N dei lanci da effettuare per ottenere testa considerato sopra?

$$\underset{1/2}{1/2} + \underset{1}{2 \cdot 1/(2^2)} + \underset{1.375}{3 \cdot 1/(2^3)} + \underset{1.625}{4 \cdot 1/(2^4)} + \underset{1.78125\ldots}{5 \cdot 1/(2^5)} \cdots + \underset{1.98828125}{10 \cdot 1/(2^{10})} + \mathbf{\ldots = 2}$$

In questo caso, a differenza di quelli all'inizio del paragrafo, la media non coincide con la mediana ma è più grande.

In questo esempio l'ultimo **"..."** sta ad indicare che la somma può proseguire all'infinito. È un'estensione del concetto di ***somma*** che, anche se implicitamente, abbiamo già incontrato più volte. Ad esempio la scrittura 1.111…, ad intendere che il numero prosegue con una successione infinita di "1", potrebbe essere sostituita da 1+1/10+1/100+1/1000+…. In questo caso si tratta di una somma che, calcolandola per un numero di addendi via via crescente, si avvicina sempre più ad un numero, appunto a 1+1/10+1/100+1/1000+…, che in questo caso potremmo scrivere anche in forma finita: 1+1/9; infatti 1/9 = 0.111…. Per un esempio analogo, 1.999… = 1+9/10+9/100+9/1000+… = 2. Ovviamente, non in tutti i casi una "somma infinita" è uguale ad un numero. Ad esempio 1+2+3+4+…, all'aumentare del numero di interi che aggiungo, cresce oltre ogni limite. È chiaro come, in casi simili a quelli richiamati negli esempi iniziali, si possono usare scritture come $\sum_{k=0 \ldots \infty} 1/10^k$, $1 + \sum_{k=1 \ldots \infty} 9/10^k$, …, in cui l'uso di $\sum$ viene esteso al caso di una somma di infiniti addendi. Su tutto ciò ci soffermeremo in futuro.

7 La variabile casuale X può assumere i valori 0, 1 e 2 con le probabilità 0.35, 0.45 e 0.20. Qual è la media di X?

8 Ho sei botti in cantina, 3 di barbera e 3 di dolcetto. Voglio del dolcetto ma non mi ricordo più in quali botti sia. Allora assaggio del vino da ogni botte, fino a che trovo quella giusta. Qual è il numero medio di assaggi che dovrò fare? [devi ottenere 1.7]

Come abbiamo visto nel ➡ §4 della scheda sul *Calcolo delle probabilità,* nel caso dell'uscita U del lancio di due dadi equi l'istogramma di distribuzione di U ha forma simmetrica rispetto alla retta di ascissa 7: quindi la media è M(U) = 7.
Osserviamo che le distribuzioni U1 e U2 delle uscite dei due singoli dadi hanno media M(U1) = M(U2) = 3.5, e 7 = 3.5+3.5. In effetti potevamo dedurre che M(U) = 7 da una proprietà più generale:

> se X e Y sono variabili casuali numeriche con medie M(X) e M(Y), la variabile casuale X+Y ha media
> **M(X+Y) = M(X)+M(Y).**

Questa proprietà è abbastanza evidente; si pensi ad un esperimento con *n* prove:

$$M(X+Y) = \mathbf{(}(x_1+y_1)+\ldots+(x_n+y_n)\mathbf{)} \,/\, n = (x_1+\ldots+x_n)/n + (y_1+\ldots+y_n)/n = M(X) + M(Y)$$

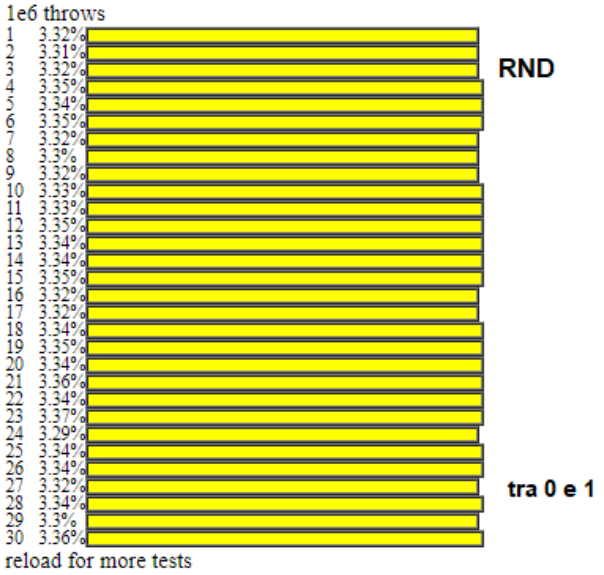
Nota. Data una variabile casuale numerica X diciamo che la media dei valori assunti da X in un certo numero *n* di "prove" è una ***media sperimentale*** (o *media empirica* o *media statistica*) di X. A volte questo numero viene indicato con $\mathbf{M_n(X)}$. Spesso tuttavia useremo al suo posto il simbolo **M(**…**)** che usiamo per indicare le medie "teoriche": dal contesto si comprende quale interpretazione darne.

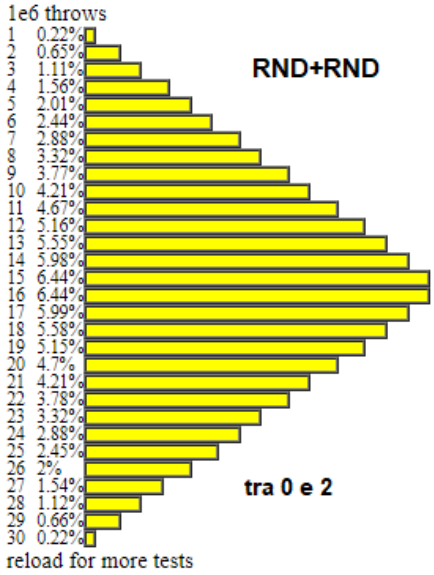
In modo del tutto analogo avviene il passaggio dalla ***varianza*** sperimentale a quella teorica, sostituendo le probabilità alle frequenze relative. Considerazioni analoghe valgono per la ***mediana***.

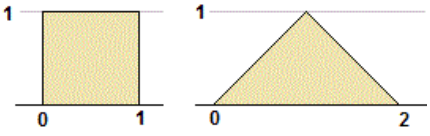
Vediamo quanto vale la varianza delle uscite di un *dado equo,* che abbiamo visto avere 3.5 come valor medio:
$\mathbf{(}(1-7/2)^2 + (2-7/2)^2 + (3-7/2)^2 + (4-7/2)^2 + (5-7/2)^2 + (6-7/2)^2\mathbf{)}/6 = 35/12$.

5. Leggi di distribuzione (variabili continue)

La variabili casuali considerate nel §1 (durate e tempi di arrivo delle telefonate) e nel terzo esempio illustrato nel §4 (le uscite del generatore di numeri casuali) erano praticamente continue ("praticamente" perché, in realtà, i tempi li misuriamo con un orologio, che non ci dà dei tempi esatti, ma delle approssimazioni, e il generatore di numeri casuali non ci fornisce un generico numero reale, ad infinite cifre, ma solo un numero limitato). Per un altro esempio di pensi alla somma di due uscite del generatore di numeri casuali. Sotto sono raffigurate le distribuzioni di questi due ultimi casi ("RND" sta per "random") simulati con 1 milione di casi e rappresentate suddividendo gli intervalli tra 0 e 1 (primo caso) e tra 0 e 2 (secondo caso) in 30 intervallini (le simulazioni sono ottenute con gli script presenti **qui**).







Nel caso discreto l'istogramma sperimentale all'aumentare delle prove tende a stabilizzarsi sull'istogramma teorico, che racchiude una superficie di area 1, nel caso continuo tende a stabilizzarsi su una curva che racchiude con l'asse x una superficie di area 1.

Nel questi due esempi si tratta, rispettivamente, di un rettangolo di base 1 e altezza 1 e di un triangolo di base 2 e altezza 1 (a lato sono illustrate le due situazioni).

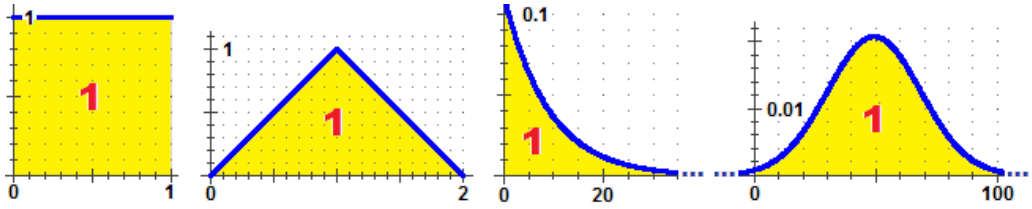
In questi casi è facile determinare l'area tra curva ed asse x. La cosa può essere fatta nel caso di una qualunque funzione continua **F** definita in un intervallo I = [**a, b**]: il suo valore viene indicato $_a\int^b \mathbf{F}$ o $\int_a^b \mathbf{F}$ o $\int_I \mathbf{F}$ e chiamato ***integrale*** *di F tra a e b (o su I).*

Quando la funzione non è descritta con un nome ma direttamente con un'espressione, come $x \to x^2$, si usa l'espressione $\int_I \mathbf{x^2\,dx}$, o, ad esempio, $\int_I \mathbf{u^2\,du}$.

Rinviamo alla ➡ scheda sulla ***integrazione*** come effettuare il calcolo in questi casi.

Come abbiamo visto, l'*integrale* si può calcolare anche per vari tipi di ***funzioni non continue***. Tieni dunque presente (anche se non approfondiremo questo aspetto) che anche l'area di un istogramma può essere interpretata come calcolo di un integrale.

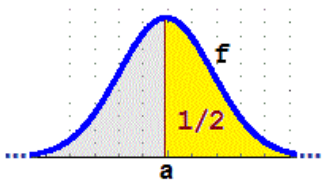
L'eventuale funzione sul cui grafico (aumentando il numero delle prove e riducendo l'ampiezza degli intervallini) si stabilizza l'istogramma sperimentale di una data variabile casuale numerica si chiama ***funzione di densità***. L'area che sta tra il suo grafico e l'asse x, nell'intervallo in cui la variabile è definita, vale 1 (il nome è una naturale estensione del termine *densità di frequenza* con cui abbiamo chiamato la frequenza relativa unitaria). Sotto a destra sono rappresentati i grafici delle funzioni su cui tendono a stabilizzarsi gli istogrammi dei tempi tra le telefonate e delle durate delle telefonate considerati nel primo paragrafo. Vedremo in una prossima scheda come descrivere tali funzioni mediante formule.

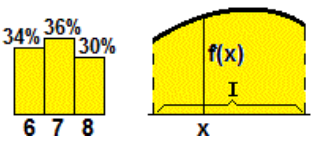


9 U è una variabile casuale continua a valori in [1,3] con legge di distribuzione uniforme. Traccia il grafico della sua funzione densità.

10 V ha la stessa legge di distribuzione della variabile casuale U del quesito precedente. Sia W = U+V. Traccia il grafico della funzione densità di W.

L'integrazione ci consente di estendere il calcolo dell'area di un istogramma a quello della superficie che sta sotto ad una curva. Ad esempio nel caso di una variabile casuale U con una distribuzione come quella raffigurata a lato abbiamo $\mathbf{Pr}(a \le \mathbf{U} \le \infty) = {}_a\int^{\infty} \mathbf{f} = 1/2$. Ci consente, inoltre, di estendere al caso continuo i concetti di media e di varianza. Vediamo come.





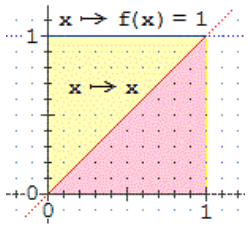
Sia **f** la densità di U. Posso definire la *media* M(U) di U in analogia al caso discreto:

– se U fosse stata a valori in $\{v_1, v_2, v_3, \ldots\}$ avrei avuto $M(U) = \Sigma v_i \cdot Pr(U=v_i)$; nel caso a sinistra avrei $6\cdot 34\% + 7\cdot 36\% + 8\cdot 30\% = 6.96$.

– nel caso continuo analogamente ho $\mathbf{M(U) = \int_I x\cdot f(x)\,dx}$

Posto $\boldsymbol{\mu} = M(U)$ ho che $Var(U) = M((U-\boldsymbol{\mu})^2)$, quindi $\mathbf{Var(U) = \int_I (x-\mu)^2\cdot f(x)\,dx}$ per quanto trovato sopra per il calcolo di M.

Consideriamo ad esempio la *distribuzione* ***uniforme*** in [0,1), già discussa sopra, che ha come densità **f**: x → 1, e calcoliamone la media (che sappiamo essere 1/2) usando la formula ora vista:



μ = $_0\int^1$ x·f(x) dx = $_0\int^1$ x dx = 1/2 (è l'area del triangolo raffigurato a **destra**).

Calcoliamone la varianza **V**:

V = $_0\int^1$ (x−μ)²·f(x) dx = $_0\int^1$ (x−1/2)² dx = 1/12

[$_0\int^1$ (x−1/2)² dx = $[(x-1/2)^3/3]_{x=1}-[(x-1/2)^3/3]_{x=0}$ = 1/2³/3+1/2³/3 = 1/12]

da cui: **sqm** = √**V** = √(1/12) = 1/√12.

Vediamo come si potrebbe calcolare l'ultimo integrale con lo script **IntegrPol**:

f(x) = (h*x^4 + k*x^3 + p*x^2 + q*x + u)^e

h 0 k 0 p 1 q -1 u 0.25 | e 1

a 0 b 1 Try with n = 10000, 20000, 40000, ...

n = 1e5, I = 0.08333333332500076

n = 2e5, I = 0.08333333333125177

n = 4e5, I = 0.08333333333281225

[0.08333… = (8+1/3)/100 = 25/3/100 = 1/(4·3) = 1/12]

11 Calcola (seguendo le indicazioni che ti dà l'insegnante) la media e lo scarto quadratico medio della prima variabile casuale considerata nel paragrafo (quella con funzione densità ad andamento "triangolare").

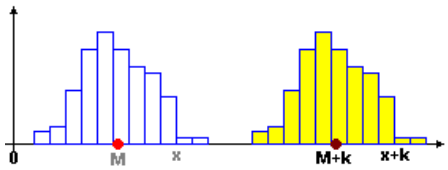
In una successiva unità didattica svilupperai ulteriormente questi argomenti.

6. Approfondimenti

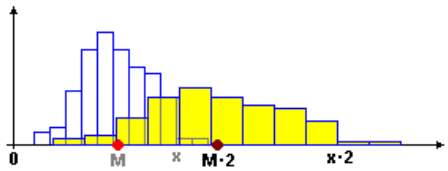
Uno

Se X è una *distribuzione* e k è un valore *costante* diversa da 0, con X+k, X-k, kX e X/k possiamo indicare le distribuzioni aventi i valori, rispettivamente, aumentati, diminuiti, moltiplicati o divisi per k, e le stesse frequenze di X. Abbiamo:

• **M(X+k) = M(X)+k**: se sostituisco ogni dato x con x+k anche la media viene variata di k (l'istogramma si sposta orizzontalmente di k, con il suo baricentro - clicca l'immagine per ingrandirla).



• **M(kX) = M(X)·k**: se sostituisco ogni dato x con kx anche la media si moltiplica per k (ad es., se dilato l'istogramma raddoppiando le ascisse – e dimezzando le ordinate: l'area deve rimanere = 100% = 1 – anche l'ascissa del baricentro raddoppia - clicca l'immagine per ingrandirla).



Ad esempio se X è la distribuzione: 980 con frequenza 3, 990 con freq. 5, 1010 con freq. 7, 1030 con freq. 5,

posso indicare con X-1000 la distribuzione:

-20 con freq. 3, -10 con freq. 5, 10 con freq. 7, 30 con freq. 5,

e con (X-1000)/10 la distribuzione:

-2 con freq. 3, -1 con freq. 5, 1 con freq. 7, 3 con freq. 5.

Per calcolare M(X) posso ricondurmi al calcolo della media di questa nuova distribuzione, ossia al calcolo di M((X-1000)/10):

(-2·3-1·5+1·7+3·5)/(3+5+7+5) = (-6-5+7+15)/20 = 11/20

e poi fare: 11/20 · 10 + 1000 = 5.5 + 1000 = 1005.5.

Due

Nel caso dell'uscita U del lancio di *due dadi equi*, qual è la varianza? (sappiamo che ➡ nel caso di un dado equo è 35/12)

Anche per la varianza si ha: ***Var(X+Y) = Var(X)+Var(Y)***. Nel nostro caso:

Var(U1+U2) = Var(U1)+Var(U2) = 35/12 + 35/12 = 35/6.

Questa proprietà vale, però, ***se X e Y sono indipendenti***, non in generale.

Si pensi, come "controesempio", al caso in cui X sia il numero sulla faccia superiore di un dado equo e Y sia quello sulla faccia inferiore. Se lancio il dado sia X che Y di distribuiscono uniformemente, con media 3.5 e varianza 35/12. Le facce opposte di un dado hanno numeri che sommati danno 7, per cui X+Y vale sempre 7. La media è dunque 7, in accordo col fatto che 3.5+3.5=7, ma la varianza è 0 in quanto X+Y ha valore costante.

Analogamente se X e Y sono *indipendenti* vale anche ***M(X·Y) = M(X)·M(Y)***. Se si pensa alla definizione di indipendenza la cosa non sorprende.

7. Esercizi

e1 Nella tabella a lato sono riportati gli esiti dei rilevamenti della pressione arteriosa massima in un gruppo di maschi quarantenni (nella colonna 1 i valori, nella 2 le frequenze assolute). I dati sono espressi in millimetri di mercurio (mm Hg) e arrotondati alle cinquine.
Determinane (usando al più una calcolatrice non programmabile) mediana, distanza interquartile, media, varianza e s.q.m.. Controlla eventualmente i risultati utilizzando opportuno software.

95	1	140	19
100	1	145	14
105	2	150	13
110	3	155	8
115	5	160	5
120	6	165	4
125	10	170	3
130	15	175	3
135	21	180	2

e2 **Qui** trovi i dati (arrotondati) relativi alle altezze e ai pesi di un gruppo di alunni maschi di 2ª media di una scuola della provincia di Genova. Analizzali statisticamente.

e3 Un sacchetto di semi scaduto ne contiene sei ancora buoni e quattro che non lo sono più. Se si prendono tre semi a caso dal sacchetto, qual è il "valore atteso" di quelli buoni tra questi?

e4 La variabile casuale X può assumere i valori 0, 1 e 2 con le probabilità 0.2, 0.5 e 0.3. Sia $Y = X^2$. Qual è la media di X? e quella di Y?

e5 Ho scritto N lettere e ho scritto i rispettivi indirizzi su N buste. Mi cade tutto. Rimetto le lettere a caso nelle buste. Ipotizzando che l'inserimento sia del tutto casuale (ossia che una lettera possa finire con uguale probabilità in tutte le buste), qual è il numero medio di lettere che vengono messe nella busta corretta?

e6 Calcola la varianza e lo scarto quadratico medio del punteggio di un dado non truccato.

e7 Un ricercatore rileva il tempo in ore di vita di 50 batteri ottenendo la media 1.34 e la varianza 0.22. Esprimendo il tempo in minuti, quali sarebbero la media e la varianza?

e8 X varia casualmente in [0,2], con legge di distribuzione avente come funzione densità **f** tale che **f**(x) = x/2 per ogni x in [0,2]. Qual è la sua mediana?
(A) 1 (B) 1/2 (C) √2 (D) 1/√2 (E) 1/3

e9 Se a un certo insieme di dati numerici ne aggiungo uno uguale al minimo di essi, la media e la varianza aumentano, diminuiscono, rimangono invariate o dipende dai casi? E se ne aggiungo uno uguale alla loro media?

e10 Abbiamo osservato nel quesito 5 che la differenza dei quadrati degli scarti di un insieme di dati da un fissato numero **p** è minima quando p è uguale alla loro media. Prova a dimostrare questa cosa.

1) Segna con l'evidenziatore, nelle parti della scheda indicate, frasi e/o formule che descrivono il significato dei seguenti termini:
indici di posizione (§2), *indici di dispersione* (§2), *distanza interquartile* (§2), *varianza* (§2), *scarto quadratico medio* (§2), *variabili casuali continue* (§4), *variabili casuali discrete* (§4), *media di una variabile casuale* (§4), *funzione densità* (§5), *media e varianza di una variabile casuale continua* (§5).

2) Su un foglio da "quadernone", nella prima facciata, esemplifica l'uso di ciascuno dei concetti sopra elencati mediante una frase in cui esso venga impiegato.

3) Nella seconda facciata riassumi in modo discorsivo (senza formule, come in una descrizione "al telefono") il contenuto della scheda (non fare un elenco di argomenti, ma cerca di far capire il "filo del discorso").

script: piccola CT grande CT isto isto con % boxplot striscia 100 ordina Grafici GraficD divisori Indet distanza Triang eq.polinomiale eq.nonPolin sistemaLin moltPolin sempliciEq divisori fraz/mcd opFraz SumPro sin LenArc Poligono Circ3P Inscr3P IntegrPol Istogramma morti RandomNum RND/RND+RND