

La distribuzione “normale” (di Gauss).

La distribuzione normale è piuttosto alla base dell’analisi statistica ed ha, invece, un ruolo più limitato nell’analisi probabilistica, non fosse altro che per il fatto che essa, non è esprimibile in forma chiusa (non si possono fare operazioni algebriche) ma solo come funzione integrale della sua densità. Dunque non è manipolabile più di tanto all’interno di formule fondamentali come la distribuzione totale.

Nel linguaggio dell’analisi probabilistica, si può dire che la funzione densità normale è la formalizzazione di un “modo di ripartizione” delle probabilità sullo spazio continuo delle realizzazioni di una variabile aleatoria fondata su due assunzioni:

1. Esiste una “realizzazione tipica”, nello spazio delle possibili realizzazioni, ed è quella che ricorre più frequentemente.
2. Esiste un “livello di dispersione ben preciso”, più o meno grande, che dà ragione di scostamenti più o meno grandi dalla realizzazione tipica e, comunque, simmetrici.

La ratio delle precedenti assunzioni si spiega con il fatto che, storicamente, con la densità normale si voleva descrivere un modo di ripartizione degli errori in esperimenti di misura di grandezze fisiche che poggiava sull’idea che l’errore tipico doveva essere nullo; ma ripetendo le misure era inevitabile la presenza di “disturbi di entità limitata”, ma assolutamente casuali e senza un verso preferenziale.

All’ipotesi 1. si può associare, in questa sede, l’interpretazione che la realizzazione tipica sia una sorta di durata vera e propria del fenomeno d’interesse, determinata da fattori endogeni. Mentre, con l’ipotesi 2., si aggiunge il fatto che, in assenza di ulteriori informazioni, si accetta l’idea che anticipi e ritardi (praticamente finiti) sulla durata del fenomeno siano equiprobabili.

Per comodità di trattazione, le ipotesi 1. e 2., saranno per il momento formalizzate imponendo che la forma di densità cercata sia riferita ad una variabile aleatoria, Z , standard, cioè di media nulla, $E[Z]=0$, e varianza unitaria, $VAR[Z]=1$. Poi si estenderà la funzione trovata al caso generico e, soprattutto, più naturale in questa sede, di una X con $E[X]=\mu$ e $VAR[X]=\sigma^2$ (si pone al quadrato perché, spesso, si lavora con la radice della varianza, detta deviazione standard e quindi pari a σ).

Con le ulteriori condizioni:

per z che tende a infinito
la funzione tende a 0

funzione crescente per $z < 0$ $\frac{d f_Z(z)}{dz} = \begin{cases} > 0 & z < 0 \\ < 0 & z > 0 \end{cases}$ e $f_Z(z \rightarrow \pm \infty) \rightarrow 0$

Più:

condizione per punto di massimo o minimo $\frac{d f_Z(z)}{dz} \Big|_{z=0} = 0$ e $\frac{d^2 f_Z(z)}{dz^2} \Big|_{z=0} < 0$ concava

lasciate all'interpretazione dello studente, si può riconoscere che la più semplice equazione differenziale che le rappresenta tutte è la seguente:

$$\frac{d f_Z(z)}{dz} = -f_Z(z) z; \quad \text{con } f_Z(z) > 0 \text{ per } -\infty < z < +\infty$$

Integrando la precedente si ottiene:

$$f_Z(z) = k \exp\{-z^2/2\}, \quad \text{con } k > 0$$

La costante k si determina imponendo che risulti pari ad uno l'integrale della f esteso a tutto lo spazio delle possibili realizzazioni della Z .

Si pone pari a 1 perché f è una densità ed il suo integrale tra $-\infty$ e $+\infty$ è pari a 1

Dunque: il due e 0 +infinito è perché è simmetrica

Preso il secondo
integrale di Eulero
$$1 = \int_{-\infty}^{+\infty} k \exp\{-z^2/2\} dz = k \frac{2}{\sqrt{2}} \int_{t=0}^{+\infty} t^{1/2-1} \exp\{-t\} dt \quad (t \triangleq z^2/2) \int_{t=0}^{+\infty} t^{\alpha-1} \exp\{-t\} dt = \Gamma(\alpha)$$

si ottiene il risultato

e poiché riconosciamo la “funzione gamma” nell'ultimo integrale indefinito:

$$\int_{t=0}^{+\infty} t^{1/2-1} \exp\{-t\} dt = \Gamma(1/2) = \sqrt{\pi}$$

concludiamo che la **densità normale standard** è:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\{-z^2/2\}, \quad -\infty < z < +\infty$$

e da qui la **distribuzione normale standard**, in forma di funzione integrale:

$$F_Z(z) = \int_{-\infty}^{u=z} \frac{1}{\sqrt{2\pi}} \exp\{-u^2/2\} du, \quad -\infty < z < +\infty \quad (\text{rimane funzione integrale})$$

I valori della “normale standard” sono ottenuti per integrazione numerica e sono tabulati. Alcune tabelle riportano la probabilità cumulativa della normale standard, ovvero la probabilità cade tra $-\infty$ and un dato valore $u > 0$, i.e., $P(-\infty \leq z \leq u)$:

Osservazione: $F_Z(-z) = 1 - F_Z(z)$ $z \rightarrow 0$ c'è simmetria rispetto all'origine

We shall now prove that the $f_X(x)$ is a probability density function by showing that if

dimostrazione alternativa
dell'integrale precedente

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2} dx$$

then $I = 1$. Direct evaluation of the integral is difficult. Hence we square I and write

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(1/2)x^2} dx \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(1/2)y^2} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(1/2)(x^2+y^2)} dx dy \end{aligned} \quad (6.4.6)$$

Substituting $x = r \cos(\theta)$ and $y = r \sin(\theta)$ in Eq. (6.4.5) we obtain,

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} r e^{-(1/2)r^2} dr d\theta = 1 \quad (6.4.7)$$

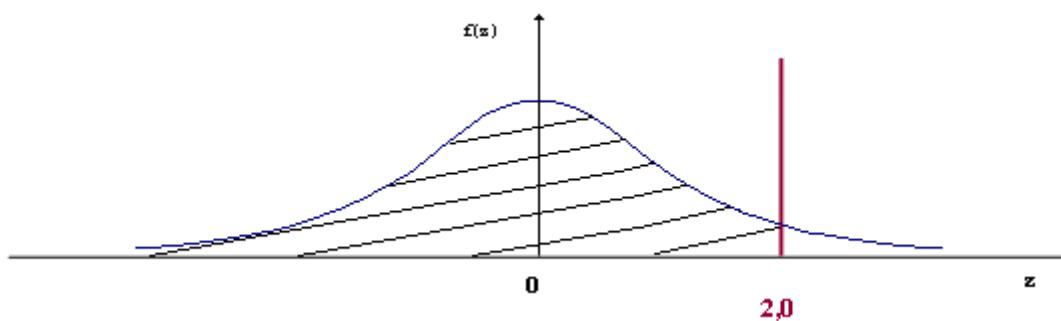
and the result is proven.

Example 6.4.1 The number of malfunctioning computers is Gaussian-distributed with $\mu = 4$ and $\sigma = 3$. We want to find the probability that the number of bad computers is

Tabella Cumulativa della $Z \sim N(0,1)$										
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Esempio 1

Qual è la probabilità che il valore di una variabile normale standard z sia ≤ 2 ?



La probabilità corrisponde all'area sottesa alla curva da $-\infty$ a 2. Il valore corrispondente a tale area è leggibile direttamente dalla tabella sopra riportata in corrispondenza dell'entrata riga-colonna (2,0 - 0,00), ovvero

$$P(-\infty \leq z \leq 2) = 0,9772.$$

Per liberarsi dalle ipotesi restrittive di media nulla e varianza unitaria, si può definire una nuova variabile aleatoria, X, (funzione della Z) alla seguente maniera:

$$X \hat{=} \sigma Z + \mu, \quad \text{con } -\infty < \mu < \infty \text{ e } \sigma > 0$$

Riconoscendo quanto segue:

$$F_X(x) \hat{=} \Pr\{X \leq x\} = \Pr\{\sigma Z + \mu \leq x\} = \Pr\left\{Z \leq \frac{x - \mu}{\sigma}\right\} \hat{=} F_Z\left(\frac{x - \mu}{\sigma}\right)$$

$$\text{grazie al fatto che } \Pr\{X \leq x\} = \Pr\left\{Z \leq \frac{x - \mu}{\sigma}\right\}$$

si deduce che il valore della gaussiana non standard nel (generico) punto "x" coincide con quello assunto dalla gaussiana standard nel punto " $(x - \mu)/\sigma$ ". Dunque, basta avere le sole tabelle relative alla gaussiana standard per valutare una gaussiana non standard.

Per assicurarsi che la forma "a campana" della densità (gaussiana) associata alla variabile aleatoria X non cambia con la trasformazione lineare definita, basta riconoscere la seguente relazione:

$$f_X(x) \hat{=} \frac{d}{dx} F_X(x) = \frac{d}{dx} F_Z\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma} f_Z\left(\frac{x - \mu}{\sigma}\right)$$

che permette di scrivere la funzione di densità della X come segue:

$$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\}, \quad -\infty < x < +\infty$$

Lo studio delle trasformazioni di variabili aleatorie non è compreso in questo corso.

Piuttosto, ci limitiamo ad osservare che la densità della X è centrata sul valore di ascissa pari a μ , con la larghezza della "campana" determinata dal fattore σ e i due punti di flesso corrispondenti alle ascisse $\mu \pm \sigma$. Peraltro si vede facilmente che alla X sono associate una media pari a μ e una varianza pari a σ^2 .

Infatti:

$$E[X] = \sigma E[Z] + \mu \quad \text{e} \quad \text{VAR}[X] = \sigma^2 \text{VAR}[Z] + \text{VAR}[\mu] = \sigma^2$$

in quanto risulta:

$$E[\sigma Z] = \int_{-\infty}^{+\infty} \sigma \cdot z \cdot f_Z(z) dz = \sigma \cdot \int_{-\infty}^{+\infty} z \cdot f_Z(z) dz = \sigma \cdot E[Z]$$

$$\text{VAR}[\sigma Z] = \int_{-\infty}^{+\infty} (\sigma z - E[\sigma Z])^2 f_Z(z) dz = \int_{-\infty}^{+\infty} \sigma^2 (z - E[Z])^2 f_Z(z) dz = \sigma^2 \text{VAR}[Z]$$

e poi il valore atteso di una costante (μ) è essa stessa mentre la sua varianza è nulla!

DETALGO TECNICO_1

dimostrazione del perché esce sigma al quadrato nella varianza

PER DEFINIZIONE DI VARIANZA:

$$\text{VAR}[x] \triangleq E[(x - E[x])^2]$$

Allora:

$$\text{VAR}[\alpha x] = E[(\alpha x - E[\alpha x])^2]$$

$$= E[(\alpha x - \alpha E[x])^2]$$

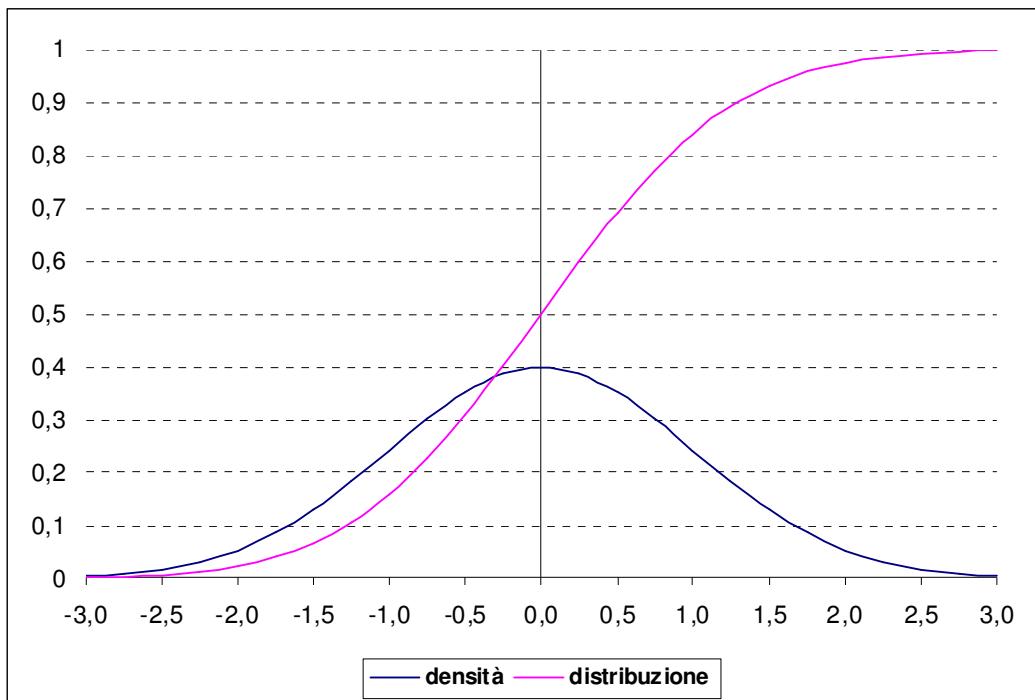
$$= E[\alpha^2 \cdot (x - E[x])^2]$$

$$= \alpha^2 E[(x - E[x])^2]$$

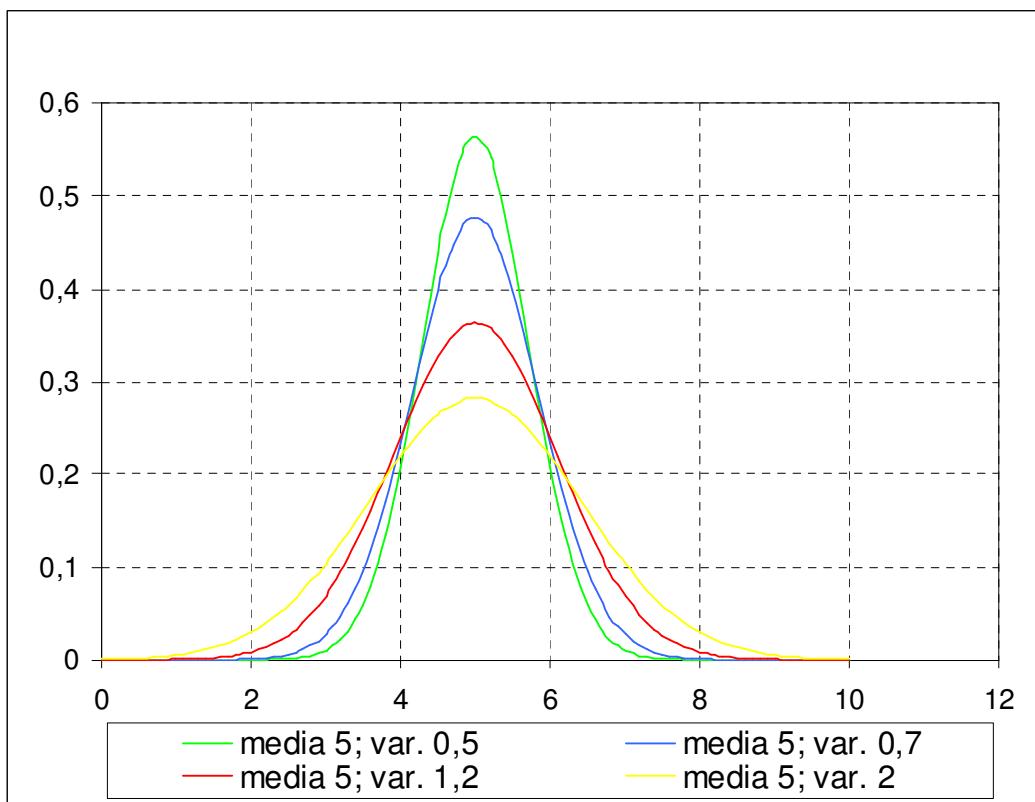
$$= \alpha^2 \text{VAR}[x]$$

Rappresentazioni della normale

è "larga"
3 deviazioni
standard poi
tende a 0



La normale standard.



Esempi di normale NON standard.

Risultati fondamentali sulla normale

L'importanza della distribuzione normale è in larga parte dovuta all'esistenza dei due risultati che saranno presentati adesso. Il primo è noto come teorema di riproducibilità della distribuzione normale perché stabilisce che la "forma" normale si mantiene rispetto all'operazione di somma (di variabili aleatorie), mentre il secondo è noto come teorema limite centrale perché stabilisce che la legge normale è la forma a cui tende la somma di variabili aleatorie di media e varianza note ed ha, per questo, un ruolo centrale nell'analisi statistica.

Teorema di riproducibilità della normale

Sia X_1, X_2, \dots, X_n una collezione di variabili aleatorie indipendenti e ciascuna distribuita secondo una legge normale di media μ_i e varianza σ_i^2 ($i = 1, \dots, n$). Allora, introdotte le costanti reali a_1, a_2, \dots, a_n , la variabile aleatoria

$$S_n \triangleq \sum_{i=1}^n a_i X_i$$

è distribuita con una legge normale di media

$$= \sum_{i=1}^n a_i \mu_i \quad \text{e varianza} \quad \sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2.$$

Teorema limite centrale

Sia X_1, X_2, \dots, X_n una collezione di variabili aleatorie indipendenti e ciascuna distribuita secondo una legge di media μ_i e varianza σ_i^2 ($i = 1, \dots, n$) entrambe finite e di forma qualsivoglia. Allora la variabile aleatoria:

$$\tilde{Z}_n \triangleq \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \quad \begin{array}{l} \text{Si sta standardizzando} \\ \text{ad una Gaussiana!} \end{array}$$

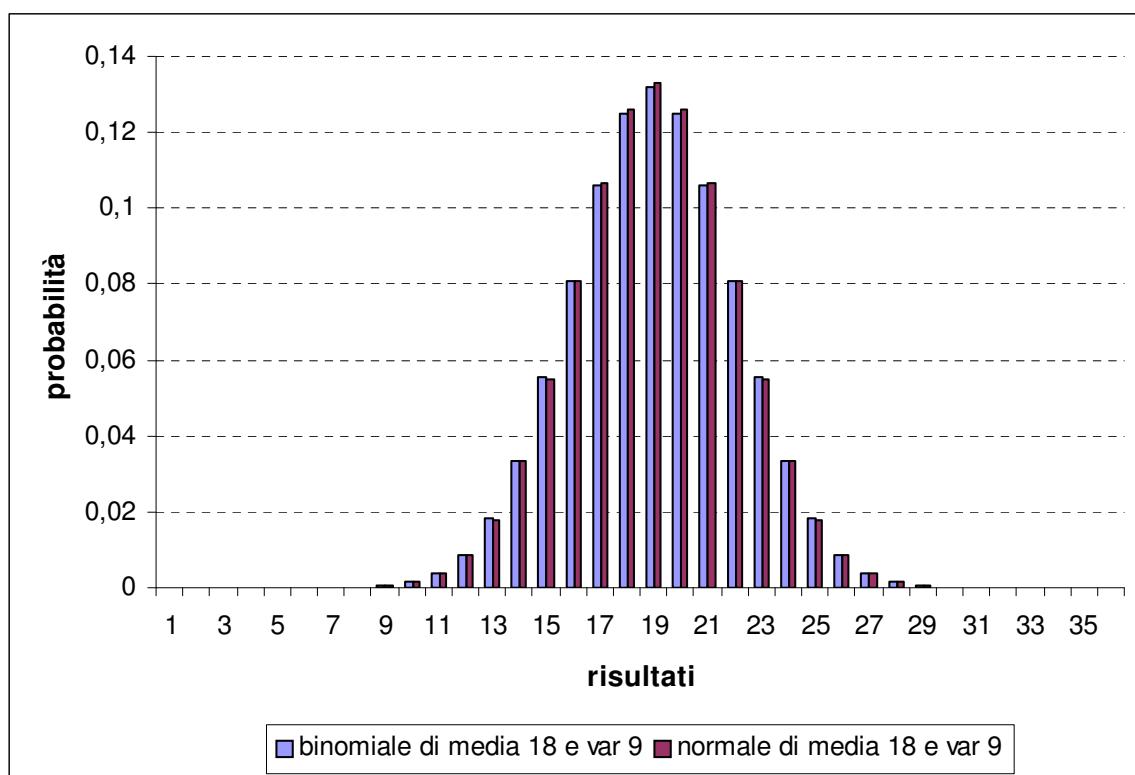
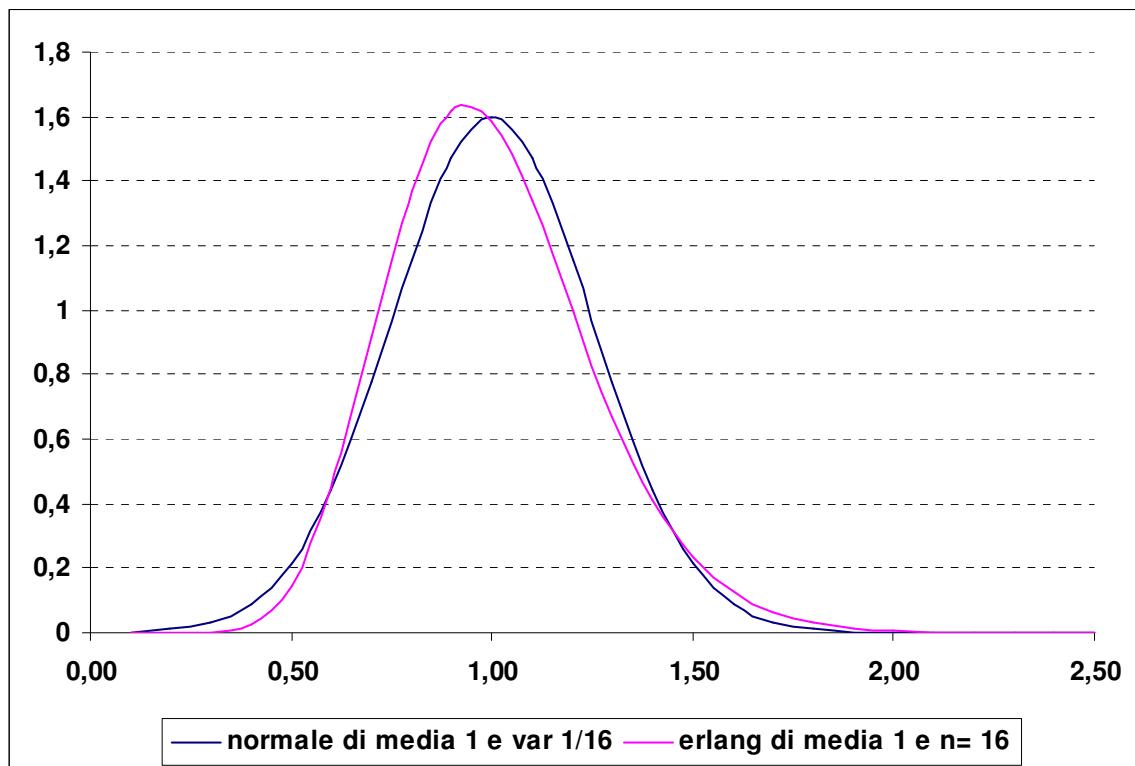
che ha, per costruzione, media nulla e varianza unitaria tende ad assumere la forma della normale standard al crescere di n , cioè:

$$n \rightarrow \infty \Rightarrow F_{Z_n}(z) \rightarrow \int_{-\infty}^{u=z} \frac{1}{\sqrt{2\pi}} \exp\{-u^2/2\} du, \quad -\infty < z < +\infty$$

La dimostrazione dei due teoremi è tralasciata.

In giallo si ha appunto una distribuzione della normale di Gauss!

Illustrazione del teorema limite centrale



Funzioni di variabili aleatorie

Abbiamo visto durante il corso le variabili aleatorie: esponenziale, di Weibull e Gaussiana. I matematici con il concetto di funzione di funzione, sono riusciti a costruire dei modelli diversi. Ma che vuol dire funzione di variabili aleatorie?

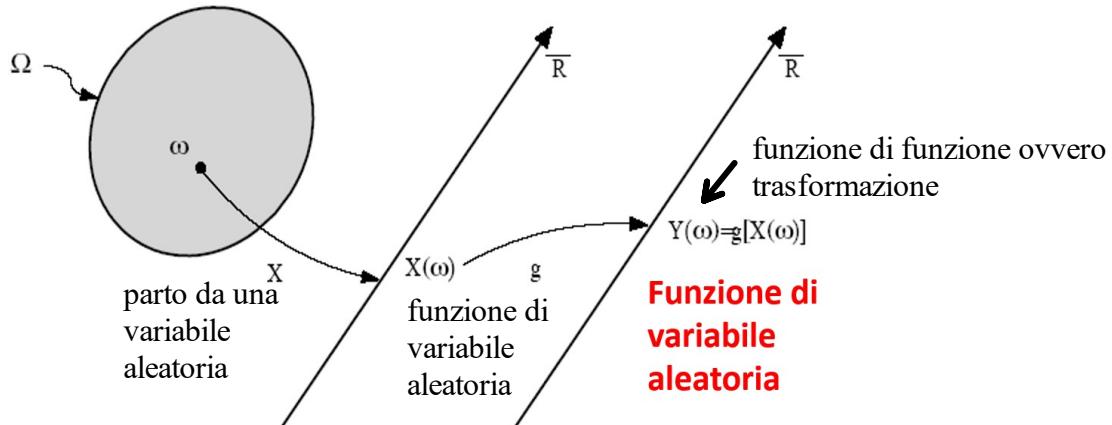
All'interno della letteratura scientifica si usa il termine trasformazioni di una variabile aleatoria. Nel nostro corso utilizzeremo il termine funzione al posto di trasformazione. Una trasformazione consiste in un'operazione algebrica di variabili aleatorie

Dal libro del prof. Gelli, "Probabilità e informazione"

86

Trasformazioni di una variabile aleatoria

"Trasformazione" = "funzione"



Il caso più facile è quello in cui g è funzione differenziabile e strettamente monotona (invertibile su tutti il dominio)

Il caso più difficile è quello in cui g è funzione differenziabile con punti a derivata nulla e interi intervalli a derivata nulla!

A ω risultato elementare, viene associato un numero e dopo tramite g se ne associa un altro. La domanda che nasce è può cambiare la probabilità di ω ?

La risposta è no, non cambia semplicemente la si rappresenta diversamente. Vediamo ciò tramite la proposizione di Miller seguente:

4.6 Transformations of Random Variables

$$F_Y(y) = \Pr(g(X) \leq y) = \Pr(X \leq g^{-1}(y)) = F_X(g^{-1}(y)). \quad (4.20)$$

probabilità che Y sia

minore o uguale ad un valore y

Note that this can also be written as

probabilità che X sia

minore o uguale all'inverso

della trasformazione applicata

a y

$F_X(x) = F_Y(g(x))$. "Si cambia la forma
ma non la sostanza"

Differenziale?

Si ma non detto
esplicitamente

(4.21)

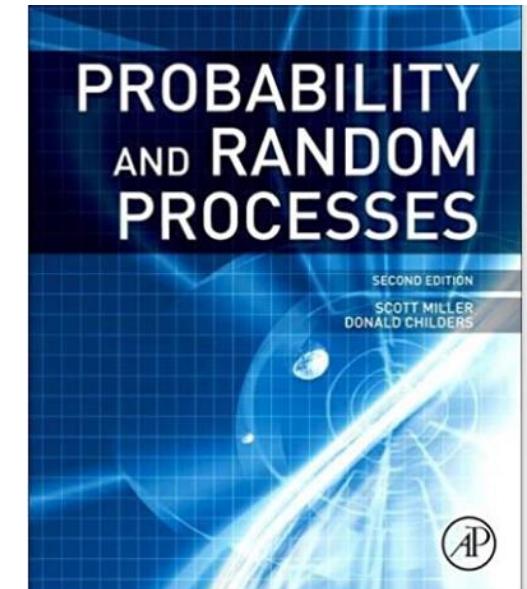
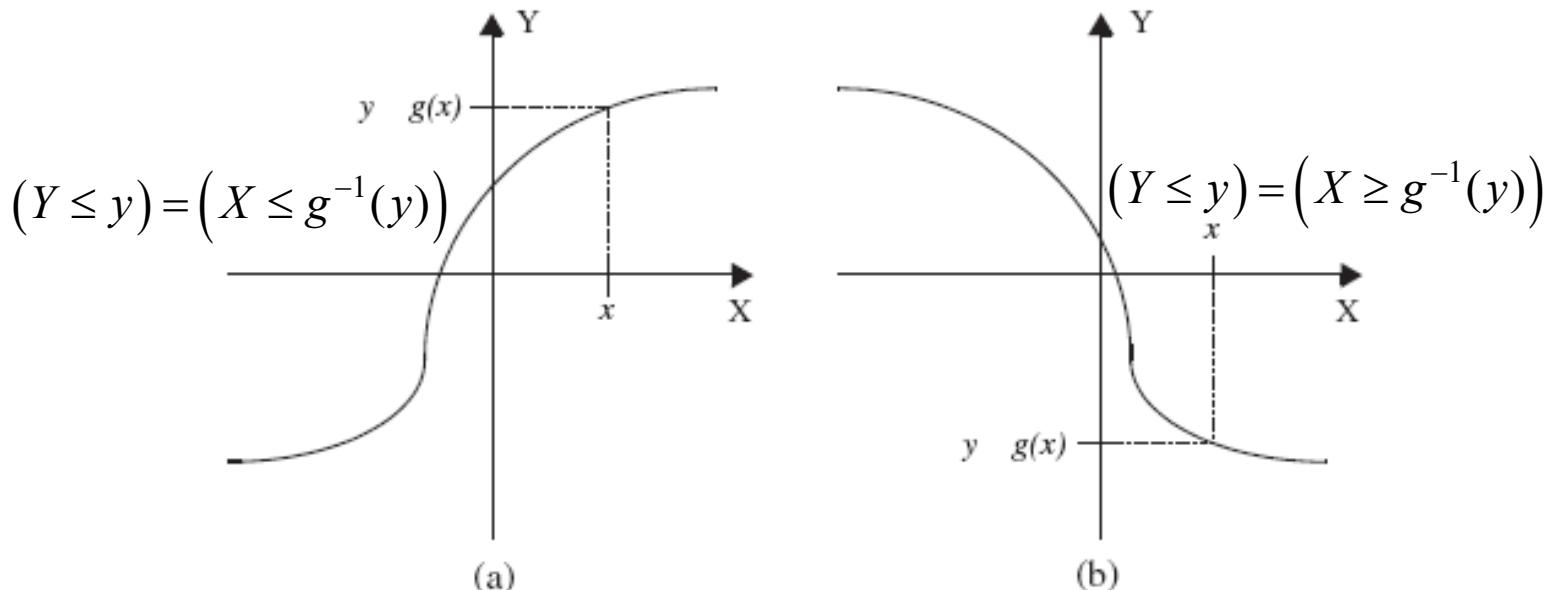


Figure 4.1 A monotonic increasing function (a) and a monotonic decreasing function (b).

Per funzioni **monotone crescenti** ottengo la densità di una variabile trasformata con le regole di derivazione delle funzioni composte

Differentiating Equation 4.20 with respect to y produces

densità di probabilità della variabile trasformata Y $\rightarrow f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} = f_X(x) \frac{dx}{dy} \Big|_{x=g^{-1}(y)}$,

La densità di probabilità di Y in un punto y è pari alla densità di X in un punto x (4.22)

while differentiating Equation 4.21 with respect to x gives

$$f_X(x) = f_Y(g(x)) \frac{dy}{dx} \Rightarrow f_Y(y) = \frac{f_X(x)}{\frac{dy}{dx}} \Big|_{x=g^{-1}(y)} \quad \leftarrow \text{come ottenere la densità di } Y \text{ a partire dalla densità di } X \quad (4.23)$$

Per funzioni **monotone decrescenti** si parte dalla distribuzione

La direzione della diseguaglianza nella probabilità arancio a differenza della funzione crescente si inverte per la natura decrescente della trasformazione

Differentiating with respect to y gives

$$f_Y(y) = -f_X(x) \frac{dx}{dy} \Big|_{x=g^{-1}(y)} \quad \begin{array}{l} \text{usa la complementare} \\ \text{segno negativo perché la} \\ \text{funzione è decrescente!} \end{array} \quad (4.25)$$

Similarly, writing $F_Y(g(x)) = 1 - F_X(x)$ and differentiating with respect to x results in

$$f_Y(y) = -\frac{f_X(x)}{\frac{dy}{dx}} \Big|_{x=g^{-1}(y)} . \quad (4.26)$$

Spiegazione evidenziato arancione:

la probabilità che Y sia $\leq y$ corrisponde alla probabilità che X è $\geq g^{-1}(y)$ perché la funzione è decrescente, quindi i valori più grandi di y corrispondono a valori più piccoli di x

Dal libro del prof. Gelli, "Probabilità e informazione"

► Esempio 4.7. Consideriamo nuovamente la trasformazione lineare

$$Y = aX + b, \quad \text{ricordando che } X \text{ è la variabile originale e } Y \text{ è la variabile ricavata}$$

Qualunque sia $y \in \mathbb{R}$, e per ogni $a \neq 0$, l'equazione $y = g(x) = ax + b$ ammette l'unica soluzione

$$x = \frac{y - b}{a}, \quad \text{ricavata tramite differenziabilità}$$

ed inoltre risulta

$$|g'(x)| = |a|, \quad \text{per cui: } f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right) \quad \text{permette di ricavare la non standard a partire dalla standard}$$

Si può procedere anche diversamente, ottenendo la distribuzione prima e derivando poi:

Nel caso $a > 0$, si ha:

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y - b}{a}\right) = F_X\left(\frac{y - b}{a}\right) \quad \text{usando la 4.22}$$

Per $a < 0$, il verso della diseguaglianza si inverte,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(aX + b \leq y) = P\left(X \geq \frac{y - b}{a}\right) \\ &= 1 - P\left(X < \frac{y - b}{a}\right) = 1 - F_X\left(\frac{y - b}{a}\right). \end{aligned}$$

la probabilità che Y sia \leq di un valore y corrisponde a calcolare la probabilità che X sia maggiore uguale al valore x

otteniamo lo stesso risultato per $a > 0$ e $a < 0$

$$\frac{d}{dy} F_Y(y) = \frac{1}{|a|} \cdot f_X\left(\frac{y - b}{a}\right)$$

Giustifico la forma della densità Gaussiana non standard a partire dalla standard

Valore atteso della funzione di una variabile aleatoria (non negativa)

ricorda $g(x) = y = ax + b$

Il valore atteso di Y $E[Y] \hat{=} \int_{y=0}^{\infty} y \cdot f_Y(y) dy$, con $Y \hat{=} a \cdot X$, $a > 0$ (esempio)
quindi $b=0$

$$E[aX] \hat{=} \int_{ax=0}^{\infty} ax \cdot f(ax) d(ax)$$

utilizzando $\frac{d}{dy} F_Y(y) = \frac{1}{|a|} \cdot f_X\left(\frac{y-b}{a}\right)$ con $b=0$

$$= \int_{x: ax=0}^{\infty} ax \cdot (1/a) f(y/a) \cdot d(ax) = \int_{x: ax=0}^{\infty} ax \cdot (1/a) f(ax/a) \cdot adx$$

$$= \int_{x=0}^{\infty} ax \cdot f(x) \cdot dx$$

Generalizzazione, con $g(x)$ idonea:

per idonea si intende monotona non decrescente e differenziabile

$$E[g(X)] \hat{=} \int_{x: g(x)=0}^{\infty} g(x) \cdot f(x) dx, \quad g(x) \geq 0$$

è esattamente la definizione del valore atteso di X per la costante a

Theorem: pdf for a transformed RV

Sia X : variabile aleatoria continua con densità f_X che non è 0 su un sottoinsieme I di numeri reali [i.e., $f_X(x) > 0, x \in I$ and $f_X(x) = 0, x \notin I$]. I può essere un punto!

Sia g : funzione monotona differenziabile con dominio I e immagine l'insieme dei reali.

Allora $Y = g(X)$: variabile aleatoria continua con densità f_Y definita come :

Densità di X valutata rispetto alla trasformazione inversa g nel punto y

$$f_Y(y) = \begin{cases} f_X[g^{-1}(y)][|(g^{-1})'(y)|], & y \in g(I) \\ 0, & \text{altrimenti} \end{cases}$$

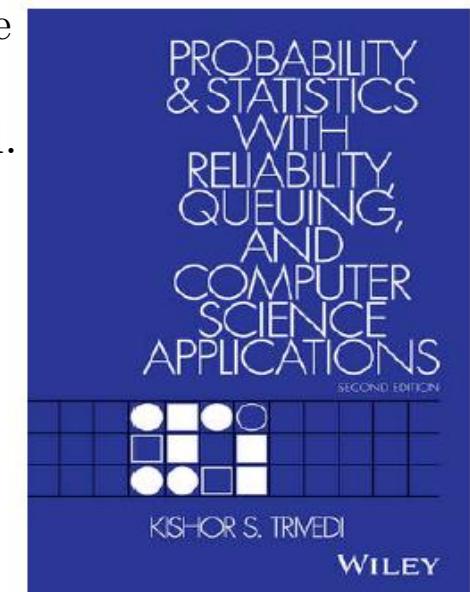
la derivata è la pendenza della funzione. Il valore assoluto mi dice di quanto sto "stirando" o "comprimendo" la distribuzione di probabilità

Prova:

Derivando ed utilizzando la regola di derivazione delle funzioni composte si ottiene che

$$F_Y(y) = P(Y \leq y) = P[g(X) \leq y] = P[X \leq g^{-1}(y)] = F_X[g^{-1}(y)]$$

Conosco la distribuzione di X e
 la funzione che trasforma X in Y ,
 quindi calcolo la distribuzione in Y



Consideriamo adesso la funzione quadratica: (che è monotona crescente strettamente)

Example 3.8

<https://ece.duke.edu/faculty/kishor-trivedi>

Distribuzione per $Y = g(X) = X^2$ Ricorda che
 $F_Y(y) = P(Y \leq y)$

$F_Y(y) = 0$, per $y \leq 0$ $P(Y \leq y)$ è zero! Essendo $Y = X^2 \implies X^2 \leq y$ ovvero $X^2 \leq$
di una quantità \leq di zero si risolve appunto solo
invece nel caso $y > 0$: per $X=0$

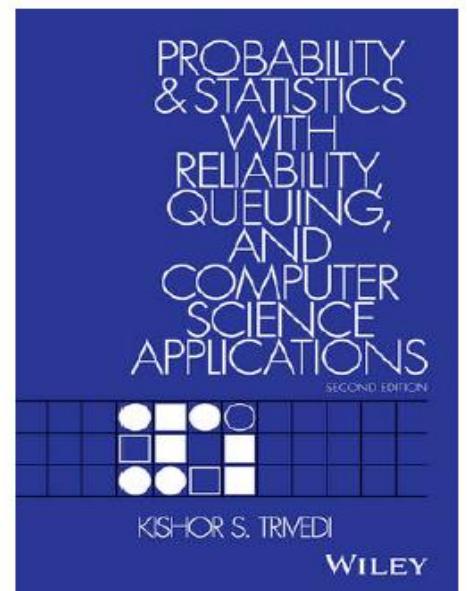
$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

Esprimo $X \in [-\sqrt{y}, \sqrt{y}]$ come differenza

La densità di Y , f_Y si ottiene con la regola di derivazione per funzioni composte,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})], & y > 0, \\ 0, & \text{altrimenti.} \end{cases}$$

Quando elevo al quadrato una variabile aleatoria sto "comprimendo" la parte negativa della distribuzione e sto "stirando" quella positiva



Per chiarezza, vi anticipo la densità della V.A. gamma, che sarà dimostrata più avanti nel corso come generalizzazione del modello di Erlang:

https://en.wikipedia.org/wiki/Agner_Krarup_Erlang

Variabile aleatoria Gamma

A partire dal secondo integrale di Eulero si definisce la funzione
portando fuori $\Gamma(\alpha)$ di densità Gamma
perchè costante
ottengo $\frac{1}{\Gamma(\alpha)} \cdot \Gamma(\alpha) = 1$

$$f(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}, \quad \alpha > 0, \quad t > 0$$

$$\int_0^\infty \frac{\hat{f}(t)}{\Gamma(\alpha)} dt = 1$$

appunto l'integrale
di una densità deve
essere pari a 1

- Show the recurrence for the gamma function:
 $\Gamma(\alpha) = (\alpha-1) \Gamma(\alpha-1)$; and show that $\Gamma(1/2) = \sqrt{\pi}$
- Because $\Gamma(1) = 1$, it follows that for an integer r , $\Gamma(r) = (r-1) \Gamma(r-1) = \dots = (r-1)!$
- So gamma with a positive integer valued shape parameter is the Erlang random variable

La variabile aleatoria Gamma con parametro di forma $\alpha = \frac{n}{2}$,

parametro di scala $\lambda = \frac{1}{2}$ è nota come variabile aleatoria "chi-square", (chi-quadrato) con n gradi di libertà

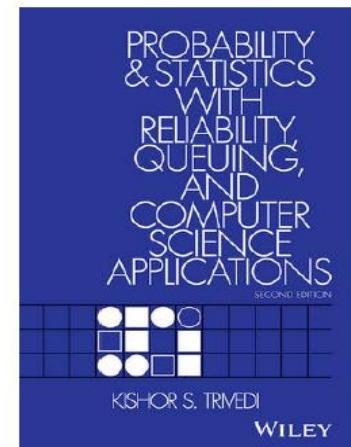
Se integro $\hat{f}(t)$
ottengo

$$\Gamma(\alpha) = \int_{t=0}^{\infty} (\lambda t)^{\alpha-1} \cdot e^{-\lambda t} d(\lambda t), \quad \lambda > 0$$

$$x \triangleq \lambda t \Rightarrow dx = \lambda dt$$

$$\Rightarrow x^{\alpha-1} \cdot dx = \lambda \cdot (\lambda t)^{\alpha-1}$$

Al crescere di alfa,
la distribuzione è
più simmetrica a destra
Al crescere di lambda
la distribuzione è
concentrata attorno a 0



IMPORTANTE
NOTIZIA

Caso particolare famoso (della funzione quadratica):

Quadrato della Gaussiana standard $N(0,1)$ Come calcoliamo le caratteristiche di densità e distribuzione del quadrato? Applicando il risultato del esempio 3.8 evidenziato in verde!

Example 3.9

In Example 3.8, assume X to be $N(0,1)$:

quindi si sa che la densità è $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $-\infty < x < \infty$.

- Using result from Example 3.8:

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} \left(\frac{1}{\sqrt{2\pi}}e^{-y/2} + \frac{1}{\sqrt{2\pi}}e^{-y/2} \right), & y > 0, \\ 0, & y \leq 0, \end{cases}$$

or, $f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}}e^{-y/2}, & y > 0, \\ 0, & y \leq 0. \end{cases}$

i.e.,

$$\Gamma(\alpha) = \int_{t=0}^{\infty} (\lambda t)^{\alpha-1} \cdot e^{-\lambda t} d(\lambda t), \quad \lambda > 0$$

Passaggi algebrici

$$\begin{aligned} f(y) &= \frac{1}{\Gamma(\frac{1}{2})} \cdot \frac{1}{2} \cdot \left(\frac{y}{2}\right)^{-\frac{1}{2}} \cdot e^{-\frac{y}{2}} \\ &= \frac{1}{\Gamma(\frac{1}{2})} \cdot \frac{1}{\sqrt{2} \cdot \sqrt{2}} \cdot \frac{\sqrt{2}}{\sqrt{y}} \cdot e^{-\frac{y}{2}} \\ &= \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{y}} \cdot e^{-\frac{y}{2}} \end{aligned}$$

Il quadrato di una standard porta ad una funzione gamma particolare con alfa e lambda particolari detta chi-quadrato che servirà per la statistica della varianza

Y has a *gamma* distribution with $\alpha = 1/2$ and $\lambda = 1/2$ e $y \equiv t$

- Which is also known as chi-square distribution with 1 degree of freedom

Una funzione logaritmo (ad hoc) su una var. al. X, UNIFORME in (0,1):

FINO A QUI DOMANDE DI ESAME!

Logaritmo è sempre
monotona e differenziabile come
vuole l'ipotesi ipotesi (slide 1)

$$Y \doteq -\lambda^{-1} \cdot \ln(1-X)$$

X distribuito uniformemente da 0 a 1 è un
generatore di numeri casuali

Example 3.10

- Let X be uniformly distributed, $\text{Unif}(0,1)$
- Then, $Y = -\lambda^{-1} \ln(1-X)$ is $\text{EXP}(\lambda)$.

$$\text{for } y \leq 0, F_Y(y) = 0$$

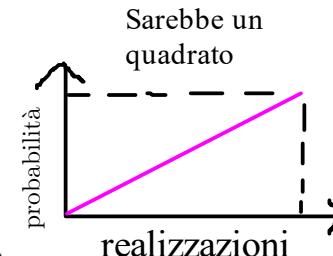
for $y > 0$,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P[-\lambda^{-1} \ln(1-X) \leq y] \\ &= P[\ln(1-X) \geq -\lambda y] \\ &= P[(1-X) \geq e^{-\lambda y}] \quad (\text{since } e^x \text{ is an increasing function of } x,) \\ &= P(X \leq 1 - e^{-\lambda y}) \\ &= F_X(1 - e^{-\lambda y}). \end{aligned}$$

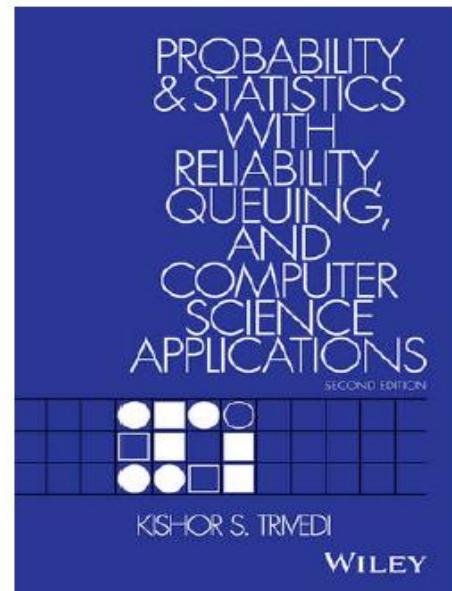
Se X è uniforme distribuito tra 0 e 1
la densità sarà costante dunque
la distribuzione che è la funzione l'integrale
della densità è lineare. Appunto integrale di una
costante funzione lineare da analisi 1

Since X is $\text{U}(0,1)$, $F_X(x) = x, 0 \leq x \leq 1$. Therefore,

$$F_Y(y) = 1 - e^{-\lambda y} \Rightarrow Y \text{ is } \text{EXP}(\lambda)$$



la realizzazione è
uguale alla probabilità
con cui si osserva la
realizzazione. Quindi
la distribuzione è la
diagonale del quadrato
vedi Monte Carlo



IN PARTICOLARE
La funzione (inversa) quantile è la
trasformazione che permette di
produrre realizzazioni a partire da
una funzione di distribuzione
invertibile (Esponenziale, Weibull ecc)

IDEA: avendo a disposizione un generatore di numeri casuali uniformemente distribuiti in (0,1) potremmo generare realizzazioni della V.A. esponenziale di parametro λ sarà il laboratorio di Excel

Example 8.1.2 (Log-Normal PDF). Let $X \sim \mathcal{N}(0, 1)$, $Y = e^X$. In Chapter 6 we named the distribution of Y the Log-Normal, and we found all of its moments using the MGF of the Normal distribution. Now we can use the change of variables formula to find the PDF of Y , since $g(x) = e^x$ is strictly increasing. Let $y = e^x$, so $x = \log y$ and $dy/dx = e^x$. Then

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = \varphi(x) \frac{1}{e^x} = \varphi(\log y) \frac{1}{y}, \quad y > 0.$$

Note that after applying the change of variables formula, we write everything on the right-hand side in terms of y , and we specify the support of the distribution. To determine the support, we just observe that as x ranges from $-\infty$ to ∞ , e^x ranges from 0 to ∞ .

We can get the same result by working from the definition of the CDF, translating the event $Y \leq y$ into an equivalent event involving X . For $y > 0$,

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \log y) = \Phi(\log y),$$

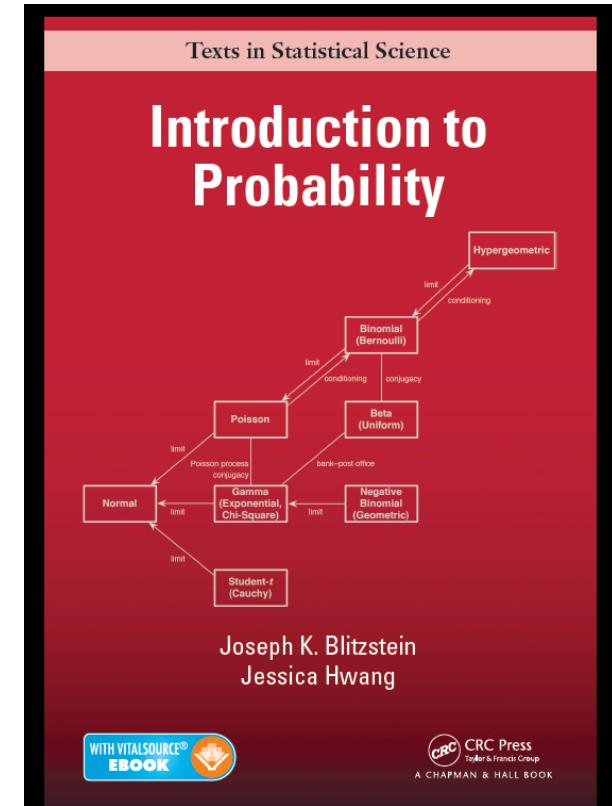
so the PDF is again

$$f_Y(y) = \frac{d}{dy} \Phi(\log y) = \varphi(\log y) \frac{1}{y}, \quad y > 0.$$

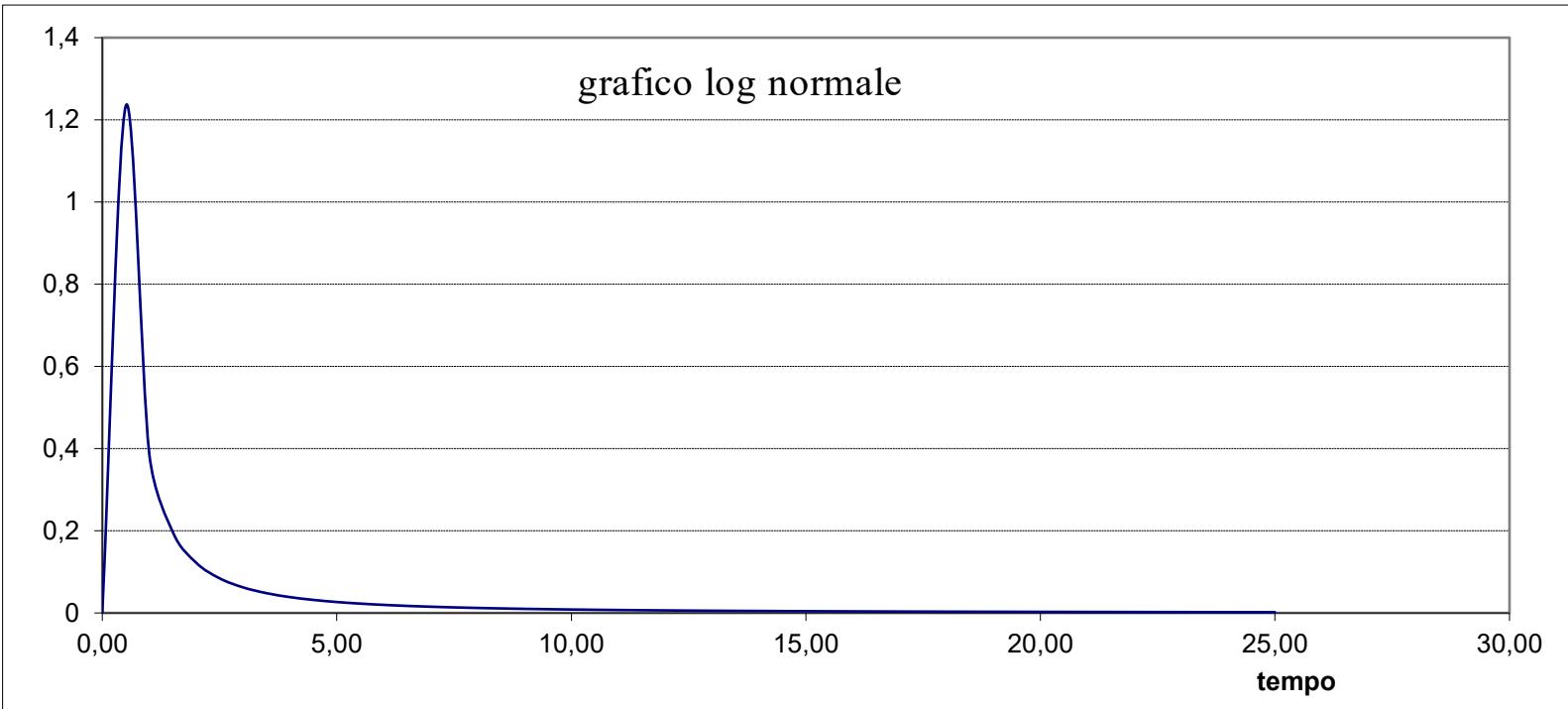
log normale *monotone* *gaussiana* *log normale*

distribuzione gaussiana

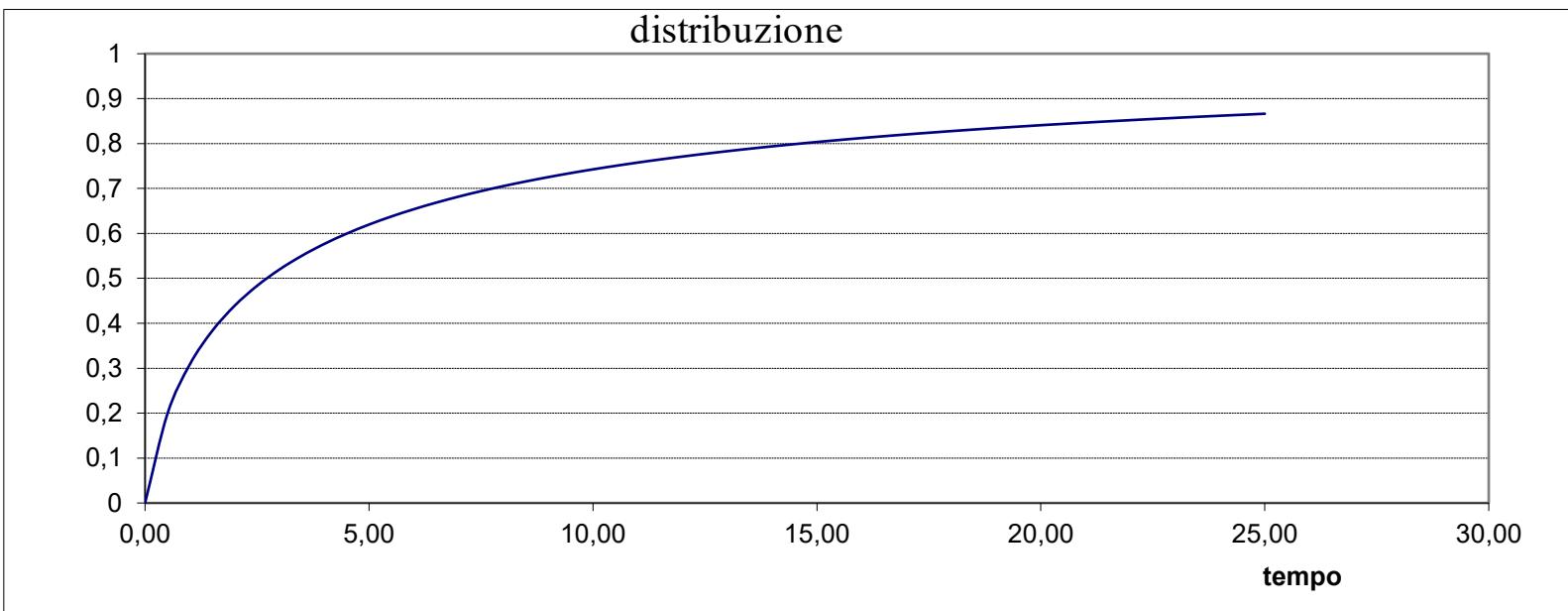
Φ è densità Gauss
Φ è distrib. Gauss



Take a look at my excel file «Modello lognormale (Legato)», please!



simile alla Weibull
ma qui la discesa/decrescita
è molto più ripida/veloce



From WIKIPEDIA

inverte la x con y

X=lognorm Y=norm

Notation	$\text{Lognormal}(\mu, \sigma^2)$
Parameters	$\mu \in (-\infty, +\infty)$, $\sigma > 0$
Support	$x \in (0, +\infty)$
PDF	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
CDF	$\frac{1}{2} + \frac{1}{2} \operatorname{erf}\left[\frac{\ln x - \mu}{\sqrt{2}\sigma}\right]$
Quantile	$\exp(\mu + \sqrt{2\sigma^2} \operatorname{erf}^{-1}(2F - 1))$
Mean	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$
Mode	$\exp(\mu - \sigma^2)$
Variance	$[\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2)$
Skewness	$(e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$
Ex. kurtosis	$\exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6$

$\Phi \triangleq \text{GAUSS DISTRIBUTION}$

$$x = e^y \implies y = \ln x$$

dimostrazione (non richiesta)

$$\begin{aligned} f_X(x) &= \frac{d}{dx} \Pr(X \leq x) = \frac{d}{dx} \Pr(\ln X \leq \ln x) \\ &= \frac{d}{dx} \Phi\left(\frac{\ln x - \mu}{\sigma}\right) \\ &= \varphi\left(\frac{\ln x - \mu}{\sigma}\right) \frac{d}{dx} \left(\frac{\ln x - \mu}{\sigma}\right) \\ &= \varphi\left(\frac{\ln x - \mu}{\sigma}\right) \frac{1}{\sigma x} \\ &= \frac{1}{x} \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right). \end{aligned}$$

normale
standardizzata

$$F_X(x) = \Phi\left(\frac{(\ln x) - \mu}{\sigma}\right)$$

$$\frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right) \right] = \frac{1}{2} \operatorname{erfc}\left(-\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)$$

where $\operatorname{erf}(z) \triangleq \frac{2}{\sqrt{\pi}} \int_{t=0}^z e^{-t^2} dt$

and $\operatorname{erfc}(z) \triangleq 1 - \operatorname{erf}(z)$

EXTRA

Appendice I

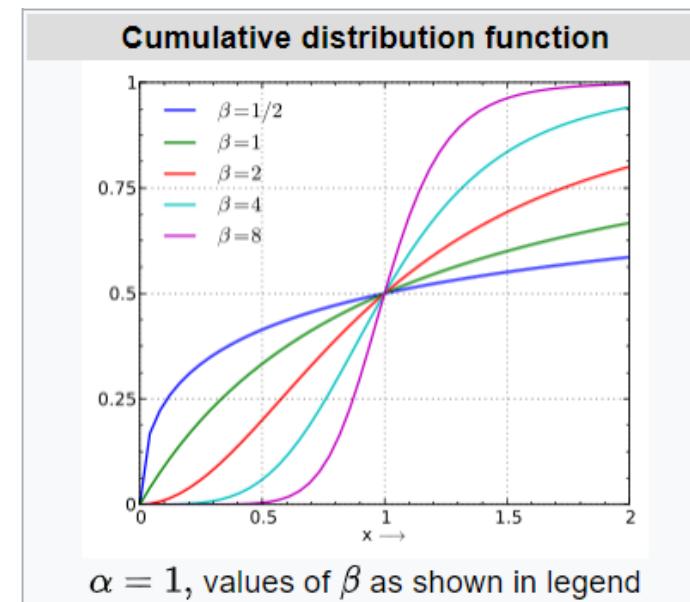
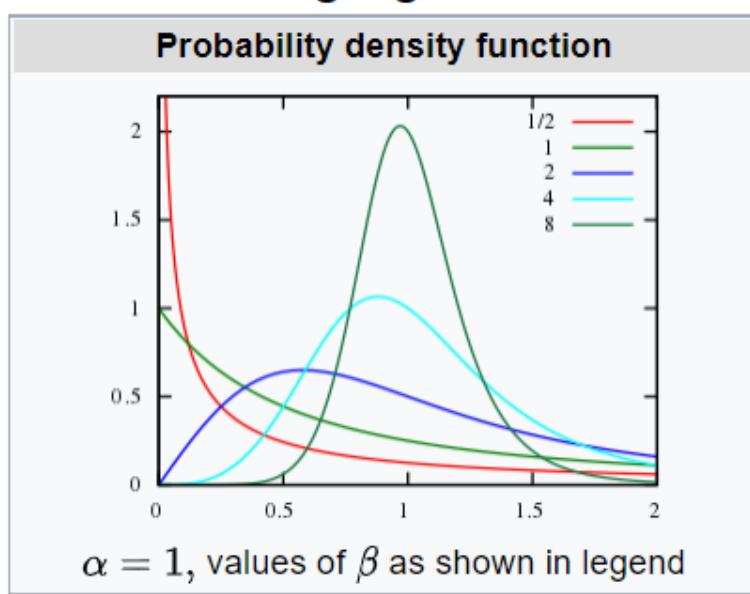
La funzione di distribuzione
Log-Logistic

Log-logistic distribution

From Wikipedia, the free encyclopedia

In probability and statistics, the **log-logistic distribution** (known as the **Fisk distribution** in economics) is a continuous probability distribution for a non-negative random variable. It is used in **survival analysis** as a parametric model for events whose rate increases initially and decreases later, as, for example, mortality rate from cancer following diagnosis or treatment. It has also been used in **hydrology** to model stream flow and **precipitation**, in **economics** as a simple model of the **distribution of wealth or income**, and in **networking** to model the transmission times of data considering both the network and the software.

The log-logistic distribution is the probability distribution of a random variable whose **logarithm** has a **logistic distribution**. It is similar in shape to the **log-normal distribution** but has **heavier tails**. Unlike the log-normal, its **cumulative distribution function** can be written in **closed form**.



There are several different parameterizations of the distribution in use. The one shown here gives reasonably interpretable parameters and a simple form for the cumulative distribution function.^{[3][4]} The parameter $\alpha > 0$ is a scale parameter and is also the median of the distribution. The parameter $\beta > 0$ is a shape parameter. The distribution is unimodal when $\beta > 1$ and its dispersion decreases as β increases.

The cumulative distribution function is

$$\begin{aligned} F(x; \alpha, \beta) &= \frac{1}{1 + (x/\alpha)^{-\beta}} \\ &= \frac{(x/\alpha)^\beta}{1 + (x/\alpha)^\beta} \\ &= \frac{x^\beta}{\alpha^\beta + x^\beta} \end{aligned}$$

where $x > 0, \alpha > 0, \beta > 0$.

The probability density function is

$$f(x; \alpha, \beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1 + (x/\alpha)^\beta)^2}$$

Attenzione:
questa volta α è usato per indicare
il fattore di scala e β il fattore di
forma!

Quantiles [edit]

The quantile function (inverse cumulative distribution function) is :

$$F^{-1}(p; \alpha, \beta) = \alpha \left(\frac{p}{1-p} \right)^{1/\beta}.$$

Può essere usata per generare realizzazioni!

It follows that the median is α , the lower quartile is $3^{-1/\beta}\alpha$ and the upper quartile is $3^{1/\beta}\alpha$.

Se $F(\bar{x}_p) = p$, $p \in (0,1)$
allora \bar{x}_p è detto
 p -simo quantile della F
 $p \doteq 0.5$ è detto mediana

Moments [edit]

The k th raw moment exists only when $k < \beta$, when it is given by^{[5][6]}

$$E(X^k) = \alpha^k B(1 - k/\beta, 1 + k/\beta)$$

$$= \alpha^k \frac{k\pi/\beta}{\sin(k\pi/\beta)}$$

where B is the beta function. Expressions for the mean, variance, skewness and kurtosis can be derived from this. Writing $b = \pi/\beta$ for convenience, the mean is

$$E(X) = \alpha b / \sin b, \quad \beta > 1,$$

and the variance is

$$\text{Var}(X) = \alpha^2 (2b / \sin 2b - b^2 / \sin^2 b), \quad \beta > 2.$$

Explicit expressions for the skewness and kurtosis are lengthy.^[7] As β tends to infinity the mean tends to α , the variance and skewness tend to zero and the excess kurtosis tends to 6/5 (see also related distributions below).

Per esercizio, potete graficare densità e distribuzione del modello "Log-logistic"!

Survival analysis [edit]

(RELIABILITY)

The log-logistic distribution provides one parametric model for survival analysis. Unlike the more commonly used Weibull distribution, it can have a non-monotonic hazard function: when $\beta > 1$, the hazard function is unimodal (when $\beta \leq 1$, the hazard decreases monotonically). The fact that the cumulative distribution function can be written in closed form is particularly useful for analysis of survival data with censoring.^[8] The log-logistic distribution can be used as the basis of an accelerated failure time model by allowing α to differ between groups, or more generally by introducing covariates that affect α but not β by modelling $\log(\alpha)$ as a linear function of the covariates.^[9]

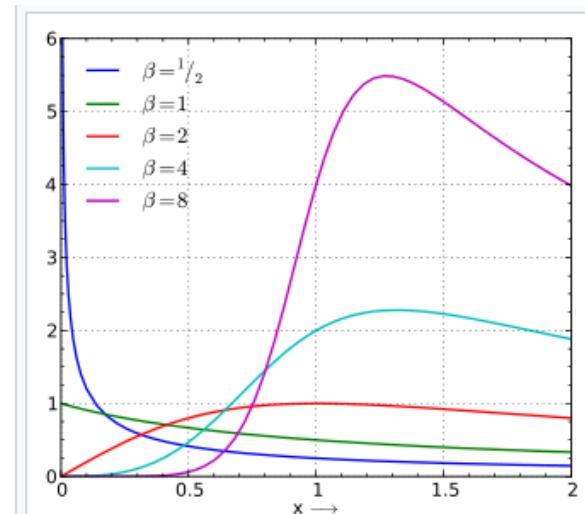
The survival function is

$$S(t) = 1 - F(t) = [1 + (t/\alpha)^\beta]^{-1},$$

and so the hazard function is

$$h(t) = \frac{f(t)}{S(t)} = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1 + (t/\alpha)^\beta}.$$

A anche questi grafici
potete riprodurre
con Excel !



Hazard function. $\alpha = 1$, values of β as shown in legend

EXTRA

Appendice II

Un teorema per:
**FUNZIONI DI VARIABILI ALEATORIE
strettamente monotone a tratti ...**

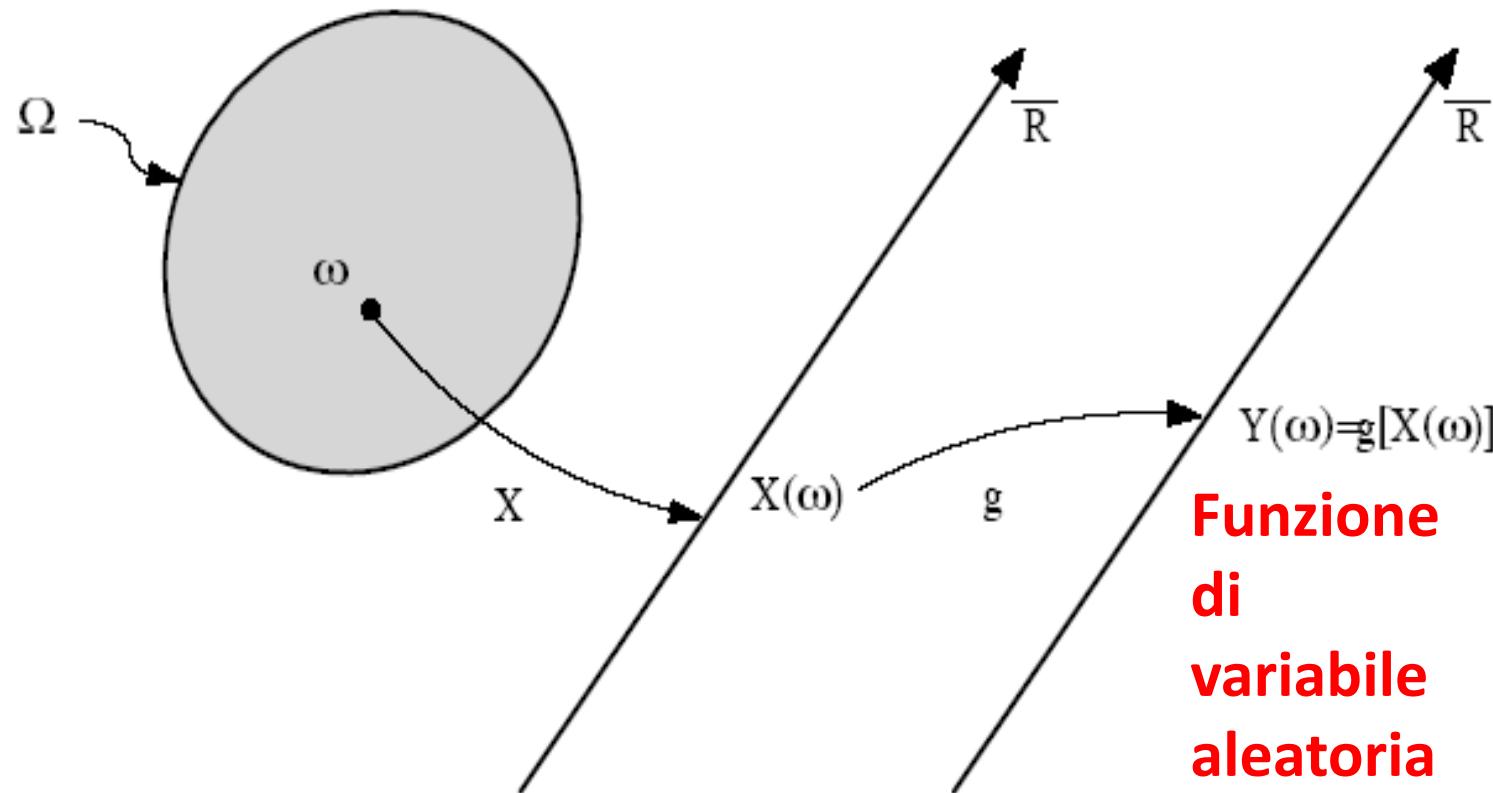
Un teorema per: FUNZIONI DI VARIABILI ALEATORIE

Dal libro del prof. Gelli, "Probabilità e informazione"

86

Trasformazioni di una variabile aleatoria

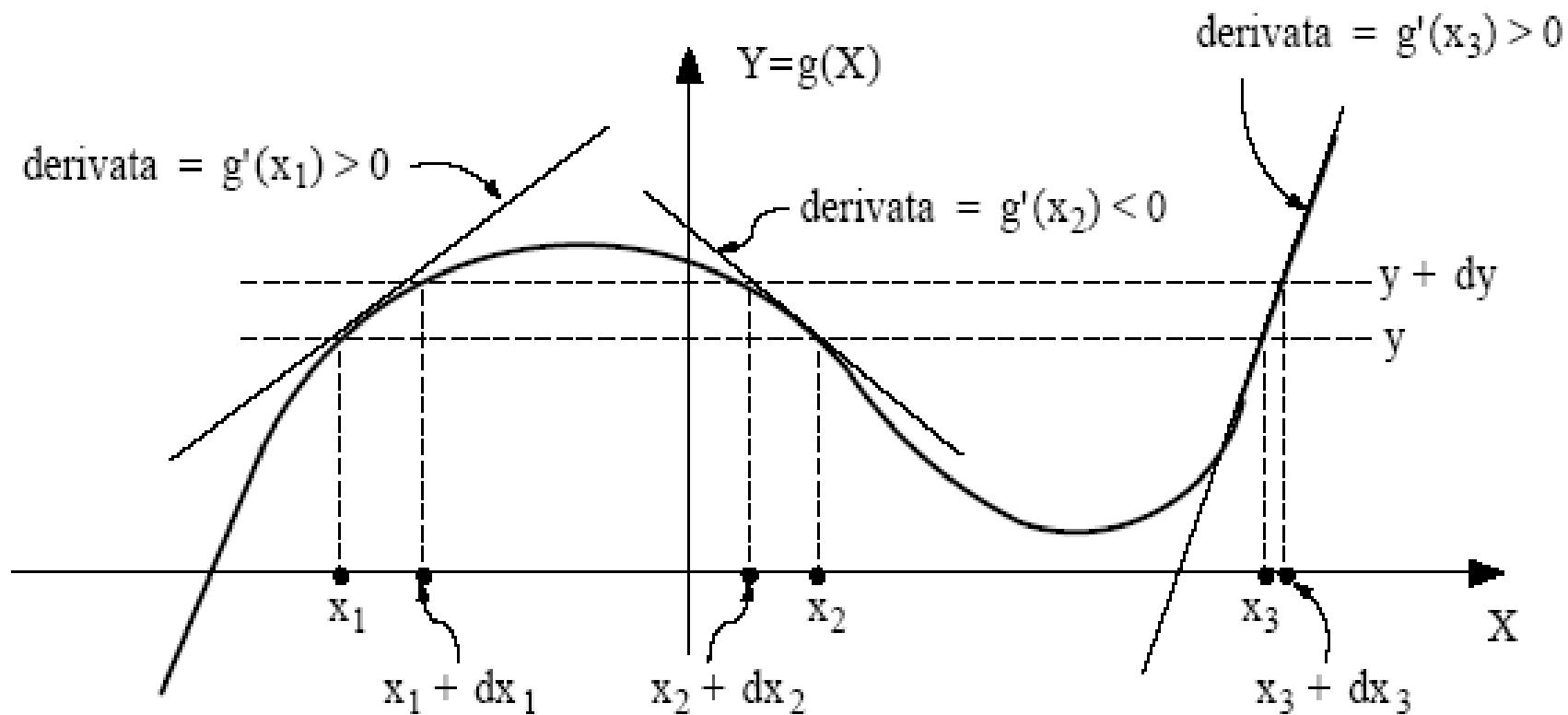
"Trasformazione" = "funzione"



Trattiamo il caso
«moderatamente
difficile»:

G è funzione
differenziabile e
«strettamente
monotona a tratti»,
con punti a derivata
nulla, che separano i
tratti,

ma senza interi
intervalli a derivata
nulla!



Definizione (trasformazione di una variabile aleatoria). Sia X una variabile aleatoria definita sullo spazio di probabilità (Ω, \mathcal{S}, P) , e $g(x)$ una funzione definita in \mathbb{R} e a valori in \mathbb{R} , tale che l'insieme di definizione di $g(x)$ contenga il codominio \mathcal{X} della funzione $X(\omega)$. La trasformazione $Y = g(X)$ definisce una nuova variabile aleatoria ottenuta associando a $\omega \in \Omega$ il valore $Y(\omega) = g[X(\omega)] \in \mathbb{R}$.

4.2.3 Calcolo della pdf di $Y = g(X)$ **$g(x)$ differenziabile e non monotona**

Affrontiamo adesso il problema di determinare la pdf di $Y = g(X)$ in funzione della pdf di X . Di importanza fondamentale è il seguente teorema, nel quale $g'(x)$ indica la derivata prima di $g(x)$:

Teorema 4.1 (teorema fondamentale sulle trasformazioni di variabili aleatorie). Sia X una variabile aleatoria avente pdf $f_X(x)$, e si consideri la trasformazione $Y = g(X)$; la pdf di Y è data da:

$$f_Y(y) = \begin{cases} 0, & \text{se l'equazione } y = g(x) \text{ non ammette soluzioni;} \\ \sum_I \frac{f_X(x_I)}{|g'(x_I)|}, & \text{dove } x_I \text{ è una soluzione dell'equazione } y = g(x). \end{cases}$$

Corollario: Se $g(x)$ è strettamente monotona, allora esiste una soluzione unica di $y=g(x)$

Caso di particolare interesse

$$Y = aX + b, \quad x = \frac{y - b}{a}, \quad |g'(x)| = |a|, \quad \text{per cui:} \quad f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right)$$

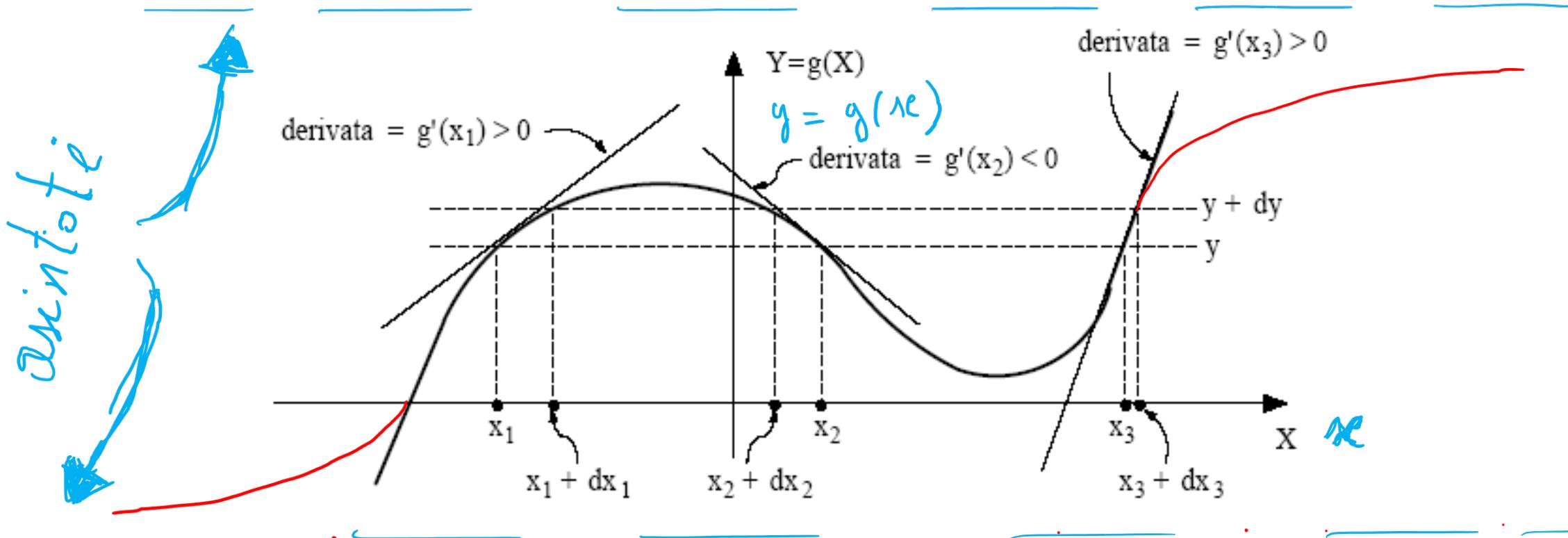


Fig. 4.14. Dimostrazione del teorema fondamentale sulle trasformazioni di variabili aleatorie. Le soluzioni dell'equazione $y = g(x)$ sono x_1 , x_2 , ed x_3 .

Se y è un valore tale che l'equazione $g(x) = y$ non ammette soluzioni, allora $f_Y(y) = 0$. Infatti, se y non appartiene alla frontiera del codominio di $g(x)$, è possibile scegliere dy sufficientemente piccolo tale che

$$\{y < g(x) \leq y + dy\} = \emptyset \Rightarrow f_Y(y) = 0.$$

Se invece y appartiene alla frontiera del codominio di $g(x)$, posso comunque porre $f_Y(y) = 0$.

Viceversa, si consideri il caso in cui y appartenga al codominio di $g(x)$,

$$f_Y(y) dy = P(y < Y \leq y + dy) = P(x_1 < X \leq x_1 + dx_1) + P(x_2 + dx_2 < X \leq x_2) + P(x_3 < X \leq x_3 + dx_3),$$

dove $dx_1 > 0$, $dx_2 < 0$, $dx_3 > 0$. (Fig. 4.14) e, poiché dy è infinitesimo, i tre insiemi cui appartiene X sono mutuamente esclusivi. Poichè:

$$P\{x_1 < X \leq x_1 + dx_1\} = f_X(x_1) dx_1; \quad dx_1 = dy/g'(x_1);$$

$$P\{x_2 + dx_2 < X \leq x_2\} = f_X(x_2) |dx_2|; \quad \text{ed inoltre} \quad dx_2 = dy/g'(x_2);$$

$$P\{x_3 < X \leq x_3 + dx_3\} = f_X(x_3) dx_3; \quad dx_3 = dy/g'(x_3);$$

dove (Fig. 4.14) $g'(x_1) > 0$, $g'(x_2) < 0$, e $g'(x_3) > 0$, risulta

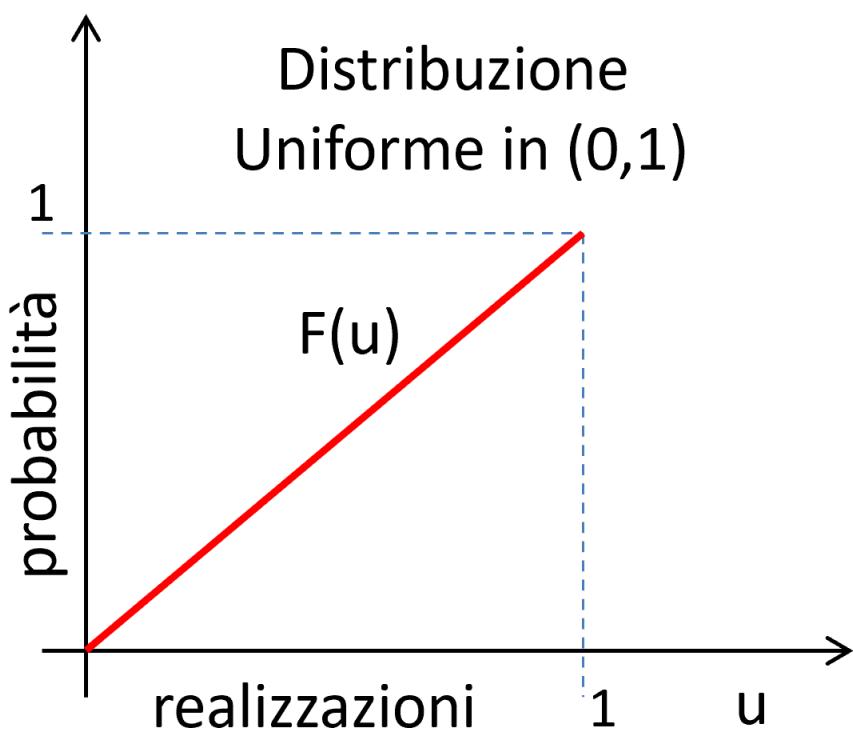
$$f_Y(y) dy = \frac{f_X(x_1)}{g'(x_1)} dy + \frac{f_X(x_2)}{|g'(x_2)|} dy + \frac{f_X(x_3)}{g'(x_3)} dy,$$

ed eliminando dy , si ha l'asserto.

Prob. di unione di eventi disgiunti

eventi

Come si è visto precedentemente, la funzione (inversa) quantile è la trasformazione che permette di produrre realizzazioni a partire da una funzione di distribuzione invertibile (Esponziale, Weibull ecc)



In generale dunque si ha che:

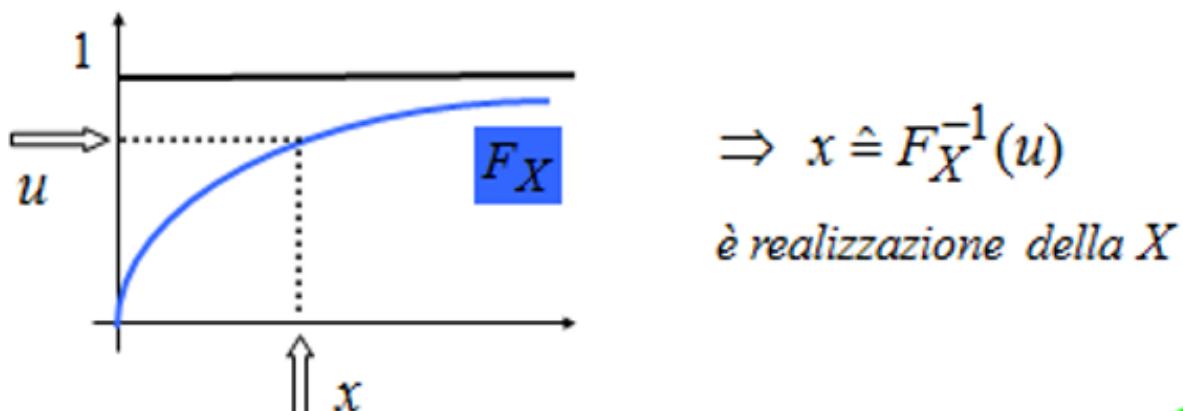
Dai numeri casuali alle realizzazioni di v.a.

$\{u_1, u_2, \dots, u_n, \dots\}$ realizzazioni indipendenti di $U\{0,1\}$

Generazione di realizzazioni delle Var. Al. d'ingresso

Metodo della trasformazione inversa
dimostrazione

$$\Pr\{X \leq x\} \hat{=} F_X(x) = F_X(F_X^{-1}(u)) = u = \Pr\{U \leq u\}$$



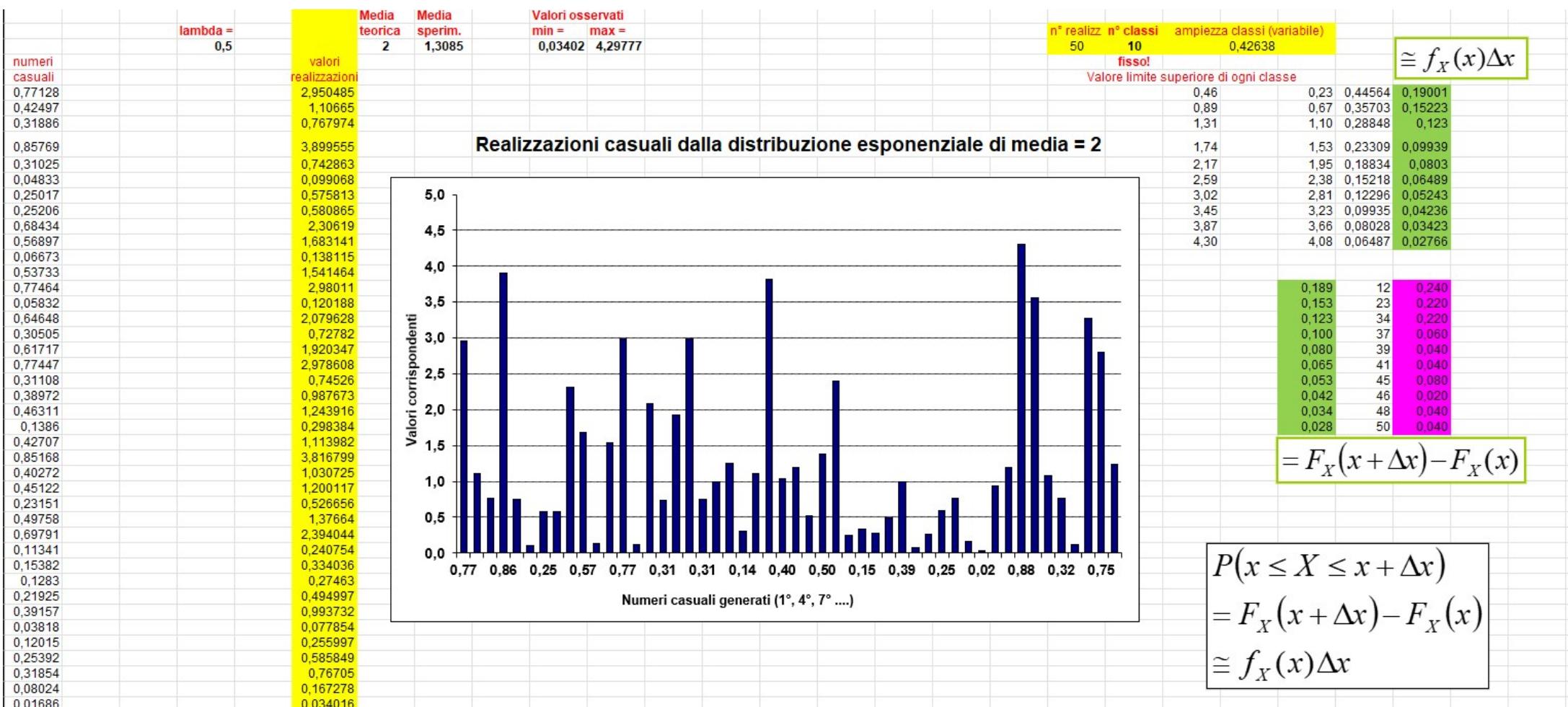
A parole: presa una funzione di distribuzione invertibile $F_x(x)$ allora si ha che

$F_X(F_X^{-1}(u)) = u$ dunque la funzione di distribuzione nell'inversa della distribuzione in u ritorna u. Esempio esponenziale con $\lambda = 1 \implies F_X(x) = 1 - e^{-x} \implies F_X^{-1}(x) = -\ln(1 - x)$

$F_X(F_X^{-1}(u)) = 1 - e^{-\ln(1+u)} = 1 - (1 - u) = u$ Quindi come abbiamo visto in precedenza la realizzazione è uguale alla probabilità con cui si osserva la realizzazione

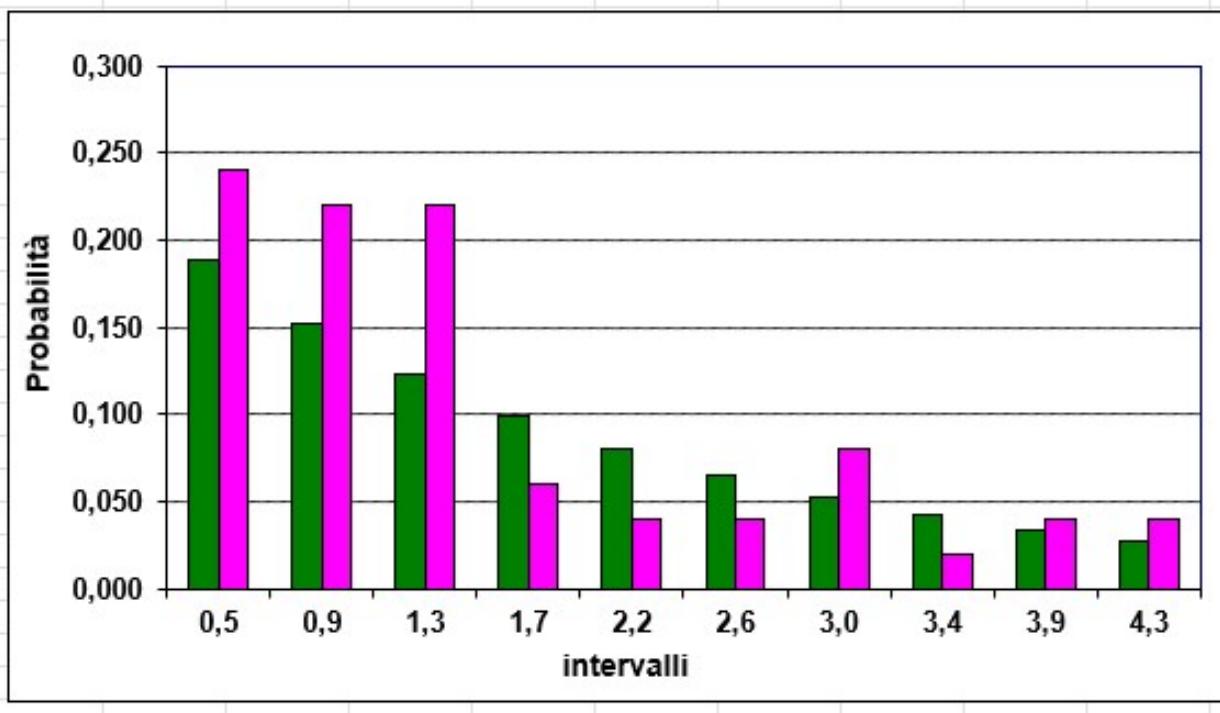
L'unica distribuzione che ha ascissa pari a ordinata è $F_X(x) = x$ appunto la diagonale del quadrato.

Ci mettiamo nell'ipotesi di modelli in cui la distribuzione è invertibile.



Grazie al metodo della trasformazione inversa si sono generate molte realizzazioni, che appunto nella cella mostrano la formula inversa. La media sperimentale sarebbe la classica media aritmetica e mostra nella simulazione che facendo più prove, tende al valore della media teorica. La media teorica sarebbe il valore atteso appunto. Quindi si capisce e vedremo in futuro che per n tendente all'infinito la media aritmetica tenda al valore atteso. Facendo simulazioni si ottengono valori che sono più piccoli o più grandi del valore atteso. L'oscillazione dei risultati a parità di dimensione del campione, è dovuta alla varianza! Non solo per la casualità. (Non si sa ancora di quanto si sbaglia) Si hanno poi i valori min e max, che mostrano quanto ad esempio un tempo di esecuzione può essere piccolissimo oppure molto grande. Il grafico di potrebbe ad esempio mostrare il fenomeno di burst all'interno di una coda. Ad esempio vado dall'elettrauto e devo cambiare la lampadine e sto 0.3 secondi, mentre può capitare che arriva quello che deve cambiare la batteria ed i cavi annessi e sta 4.3 secondi (ovviamente non impiega secondi un elettrauto). Riordinare i valori non basta per ricostruire la legge esponenziale di media. Oltre ai valori di min e max occorre fissare un numero di classi, ovvero un numero di raggruppamenti dove i valori possono finire, un numero di realizzazione da prendere in esame e l'ampiezza delle classi, che sarebbe la differenza tra max e min diviso il numero di classi fissato.

Ricostruzione legge esponenziale di media = 2



lizz	n° classi	ampiezza classi (variabile)	
	10	0,42638	
fisso!			
Valore limite superiore di ogni classe			
2			
	0,46	0,23	0,44564
	0,89	0,67	0,35703
	1,31	1,10	0,28848
	1,74	1,53	0,23309
	2,17	1,95	0,18834
	2,59	2,38	0,15218
	3,02	2,81	0,12296
	3,45	3,23	0,09935
	3,87	3,66	0,08028
	4,30	4,08	0,06487
			0,19001
			0,15223
			0,123
			0,09939
			0,0803
			0,06489
			0,05243
			0,04236
			0,03423
			0,02766

$$\cong f_X(x)\Delta x$$

Il limite superiore di ogni classe si ottiene partendo da 0 e sommando l'ampiezza della classe al valore minimo. L'ultimo limite sarà il valore di max.

$f_X(x)\Delta x$ corrisponde alla percentuale di valori che ci si aspettano in una classe. (19% nello screen)
Ma tale percentuale si può anche con la distribuzione

$$\cong f_X(x)\Delta x$$

classe

0,23	0,44564	0,19001
0,67	0,35703	0,15223
1,10	0,28848	0,123
1,53	0,23309	0,09939
1,95	0,18834	0,0803
2,38	0,15218	0,06489
2,81	0,12296	0,05243
3,23	0,09935	0,04236
3,66	0,08028	0,03423
4,08	0,06487	0,02766

0,189	12	0,240
0,153	23	0,220
0,123	34	0,220
0,100	37	0,060
0,080	39	0,040
0,065	41	0,040
0,053	45	0,080
0,042	46	0,020
0,034	48	0,040
0,028	50	0,040

$$= F_X(x + \Delta x) - F_X(x)$$

Guardando il grafico con 15 classi si nota come il modello teorico approssima ancora meglio la legge esponenziale, mentre la simulazione sballa di molto! Da notare che la media sperimentale nel esempio è di 1.9835! MOLTO VICINA A 2. Quindi la casualità anche se c'è non porta solo lei a sbagliare! Quanto si peggiora per caso dipende dalla varianza

$$\text{In particolare si sa che: } F_X(x + \Delta x) - F_X(x) \cong f_X(x)\Delta x$$

per il concetto di limite di rapporto incrementale. Dunque guardando le due colonne in verde le percentuali a meno di qualche decimale si approssima abbastanza bene. Ovviamente per applicazioni critiche ad esempio: percentuale di sopravvivenza di un uomo, i decimali (delta x) fanno la differenza. In conclusione l'istogramma è diviso in due: la parte verde è il modello teorico, la parte fucsia è il modello ricavato dalla simulazione. Il numero di classi ovviamente ha un ruolo nel mostrare il modello. Magari più classi mostrano meglio o magari no se la varianza è molto alta. Sotto un esempio con 15 classi:

1° classi ampiezza classi (variabile)

15 0,41535

fisso!

re limite superiore di ogni classe

0,42	0,21	0,45039	0,18707
0,83	0,63	0,36568	0,15189
1,25	1,04	0,29711	0,1234
1,66	1,46	0,24139	0,10026
2,08	1,87	0,19612	0,08146
2,49	2,29	0,15935	0,06618
2,91	2,70	0,12946	0,05377
3,33	3,12	0,10519	0,04369
3,74	3,53	0,08546	0,0355
4,16	3,95	0,06943	0,02884
4,57	4,36	0,05641	0,02343
4,99	4,78	0,04583	0,01904
5,40	5,19	0,03724	0,01547
5,82	5,61	0,03026	0,01257
6,23288	6,03	0,02458	0,01021

$$\cong f_X(x)\Delta x$$

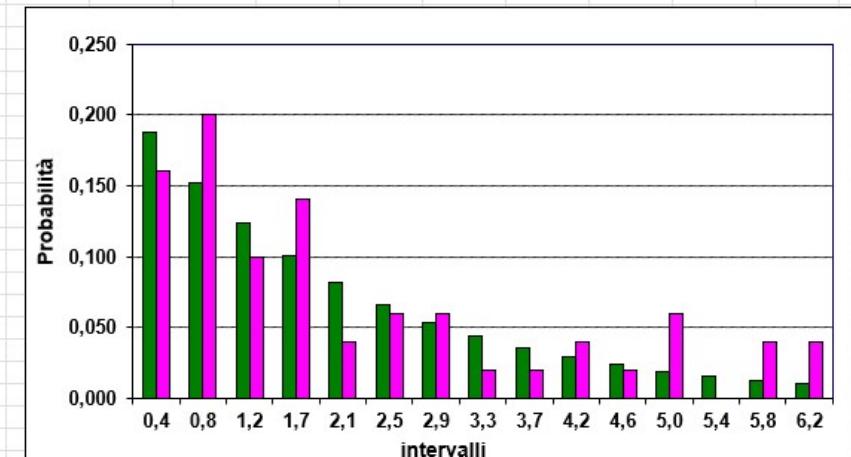
$$P(x \leq X \leq x + \Delta x)$$

$$= F_X(x + \Delta x) - F_X(x)$$

$$\cong f_X(x)\Delta x$$

$$= F_X(x + \Delta x) - F_X(x)$$

Ricostruzione legge esponenziale di media = 2

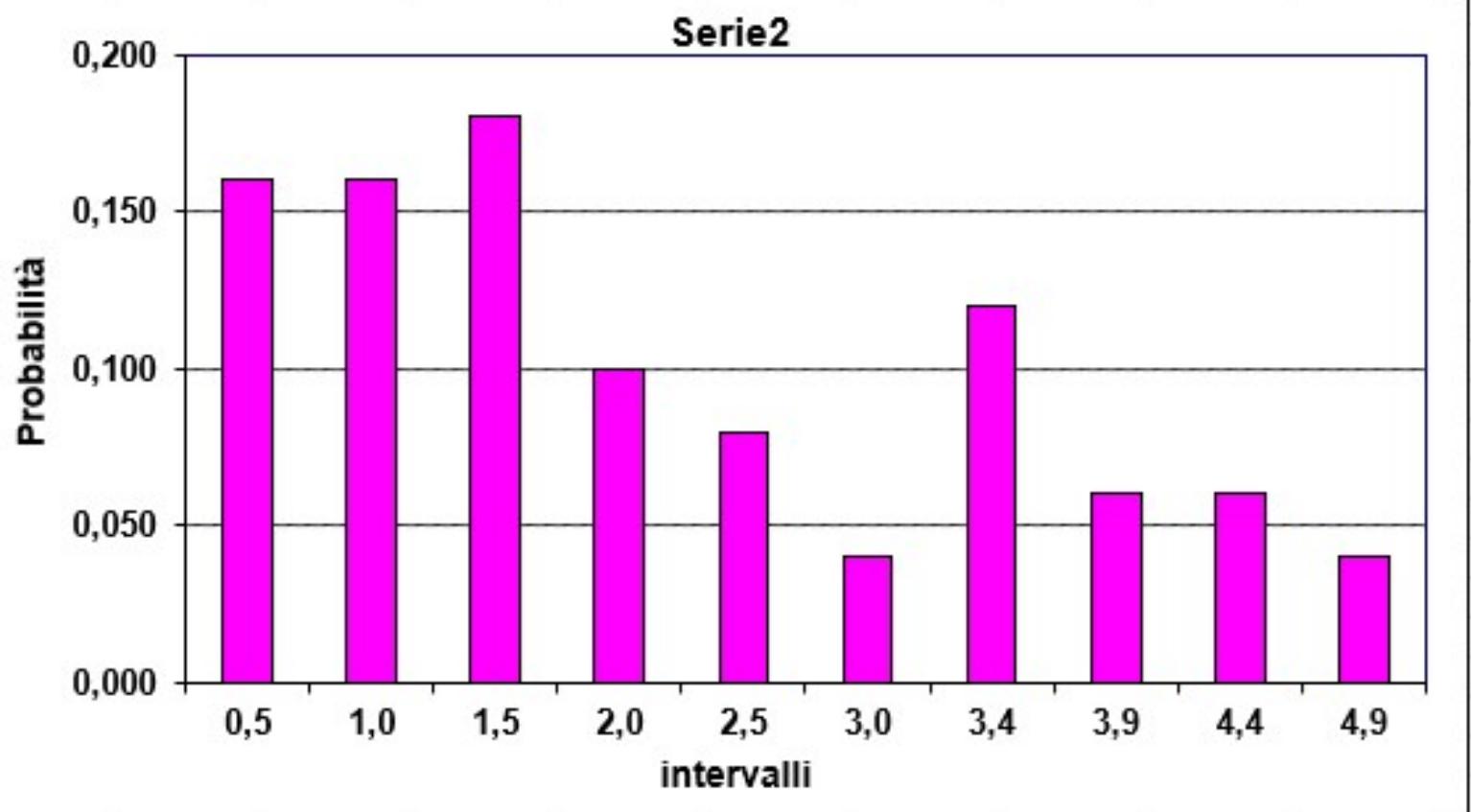


0,187	8	0,160
0,152	18	0,200
0,124	23	0,100
0,100	30	0,140
0,082	32	0,040
0,066	35	0,060
0,054	38	0,060
0,044	39	0,020
0,036	40	0,020
0,029	42	0,040
0,023	43	0,020
0,019	46	0,060
0,015	46	0,000
0,013	48	0,040
0,010	50	0,040

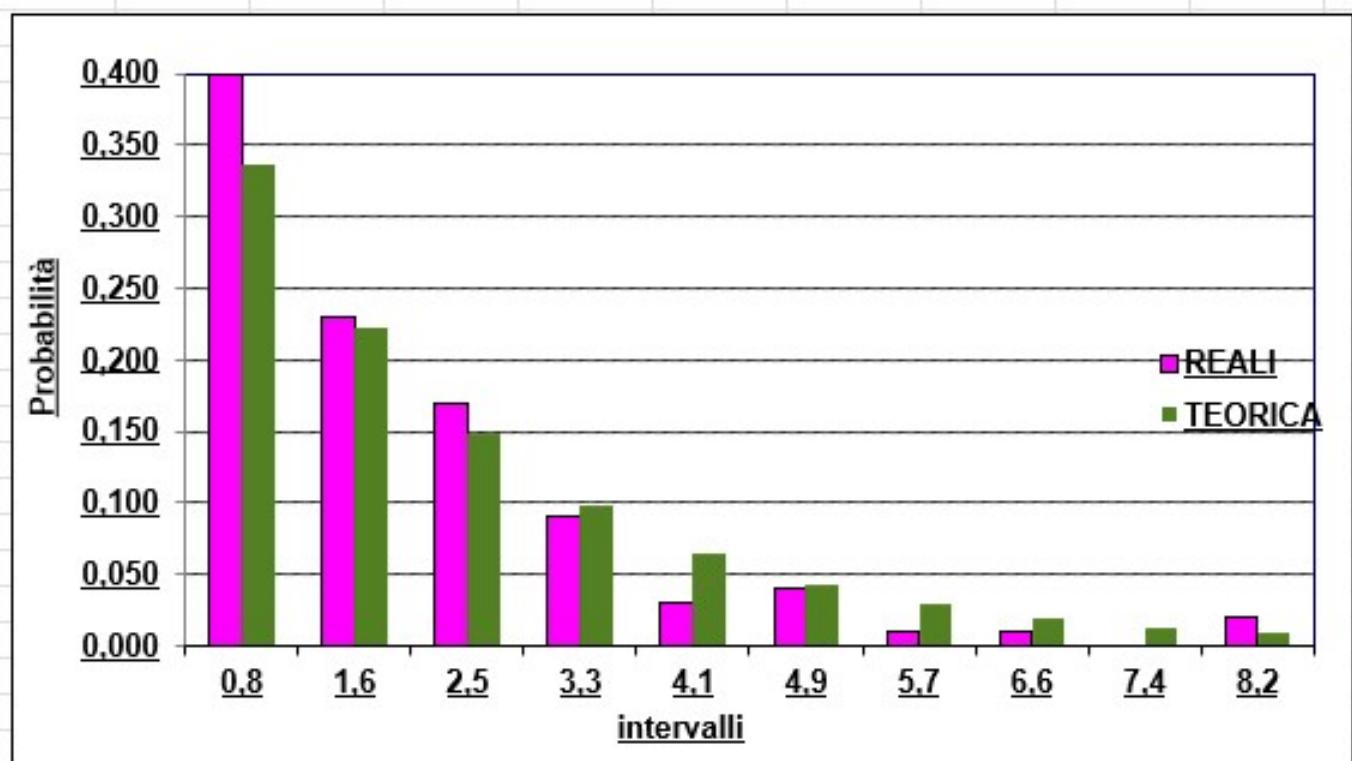
Noi faremo statistica inferenziale. Piccolo appunti

La statistica descrittiva consiste nel dire quali parametri ci danno una caratteristica sintetica sulla forma: valore atteso, varianza, momento del secondo ordine, simmetria ecc.

L'inferenza (deduzione vedi dizionario) sono i metodi per dire quanto sarà il valore atteso. Quindi prendo un oggetto/un fenomeno, di cui scelgo quanti e quali attributi e ne provo a dedurre la forma. Quindi in questo caso l'inferenza completa che noi facciamo è la forma della legge. Prendiamo ad esempio il seguente istogramma:



Non assomiglia per niente alla curva della legge esponenziale. La statistica consiste nel ricavare con pochi dati o meno dati possibili una risposta accettando il rischio di sbagliare. Quindi per prima cosa interessa il numero di dati. Poi il numero di classi.

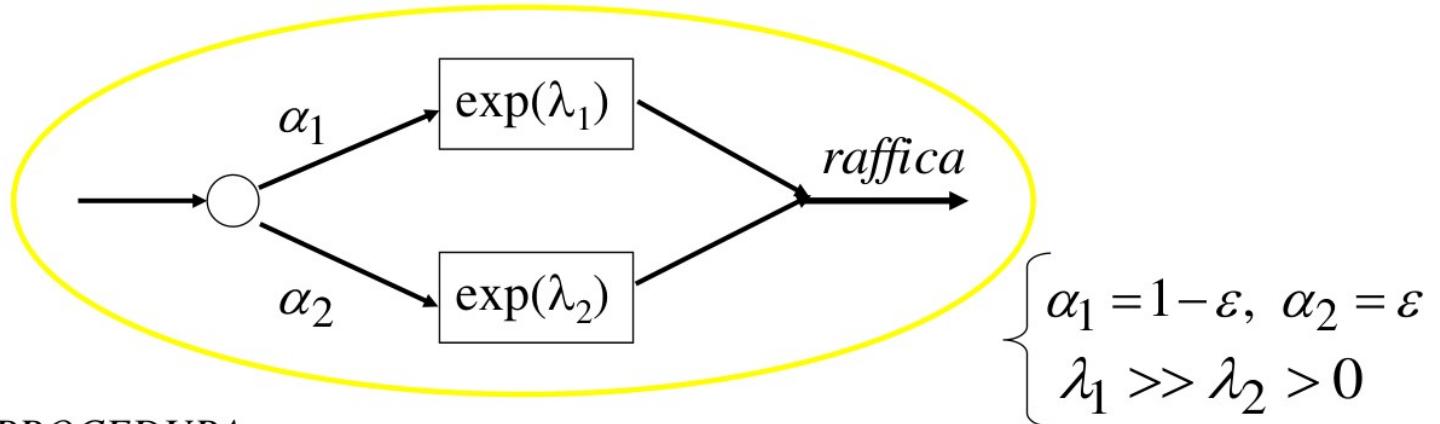


Nel grafico precedente si vede che anche se con media sperimentale 1.62(!), se uso 100 realizzazione quindi il doppio dei dati, per 10 classi la curva della legge esponenziale è abbastanza simile alla teorica. Prendiamo adesso la distribuzione iperesponenziale:

Trasformazione inversa per la distribuzione iperesponenziale

$$F_Y(y) \hat{=} \sum_{i=1}^n \alpha_i (1 - \exp\{-\lambda_i y\}), \quad \lambda_1 \neq \lambda_2 \neq \dots \neq \lambda_n$$

$$\text{con: } \sum_{i=1}^n \alpha_i = 1, \quad \alpha_i > 0 \quad i = 1, \dots, n$$



PROCEDURA:

genera $u \in U\{0,1\}$

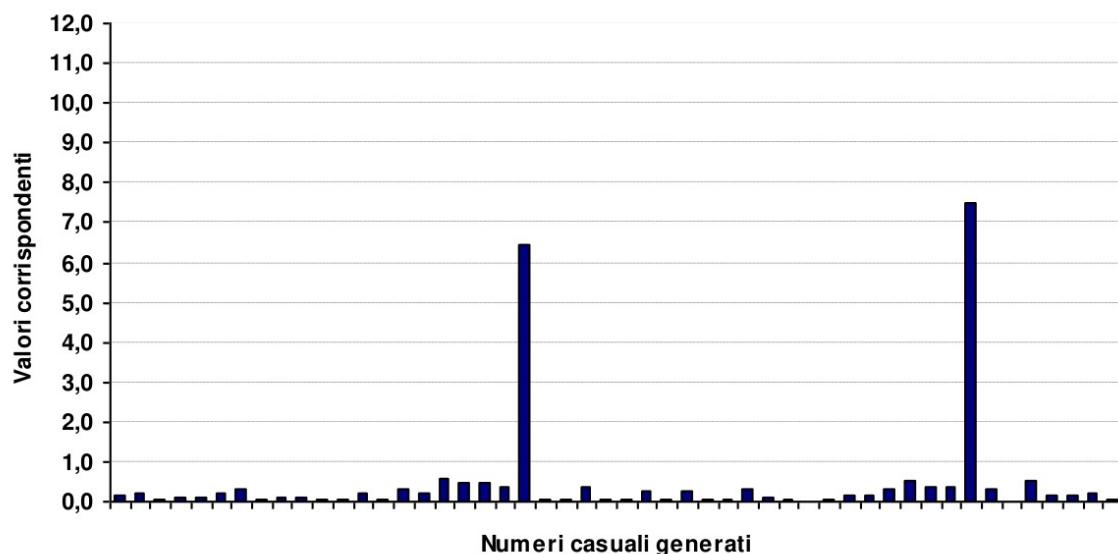
if $u \leq \alpha_1$ *then* *genera*($y \approx \text{EXP}(\lambda_1)$)

else *genera*($y \approx \text{EXP}(\lambda_2)$)

La trasformazione inversa del iperesponenziale è una combinazione convessa. Tramite l'iperesponenziale si può studiare l'effetto di bursting. In particolare come mostra l'immagine si ha che λ_1 è molto grande, quindi l'esponenziale molto veloce e tempi brevi λ_2 è piccolissimo quindi esponenziale lento e tempi lunghissimi. Quindi generato un numero casuale u tra 0 e 1, si ha che $u \leq \alpha_1$, $\alpha_1 = 0.95$ genera una realizzazione di e^{λ_1} altrimenti di e^{λ_2} . I numeri "passano" uno alla volta ecco perchè si osserva un effetto raffica. Il compito della statistica sarà capire i valori di alpha e lambda, a partire da dei dati. Vediamo di seguito la generazione monte carlo di realizzazioni iperesponenziali

Realizzazioni casuali dalla legge IPEResponenziale

(alfa = 0,95 - lambda_1 = 5 - lambda_2 = 0,5)



STATISTICHE E STIMATORI

statistica utilizzata per stimare un parametro incognito nel campione di dati considerato

Una **statistica** è una variabile aleatoria funzione di un numero fissato ($n \geq 1$) di altre variabili aleatorie, ma che non contiene alcun parametro incognito.

La realizzazione della statistica è una quantità numerica calcolata a partire da un campione di dati (variabile aleatoria funzione di un numero fissato di altre variabili aleatorie). Proprio perché parte da un campione di dati, non dipende da parametri incogniti

Sia T una statistica per le variabili aleatorie i.i.d. $X_1 \dots X_n$ e sia θ un parametro incognito di queste ultime; allora $T(\mathbf{X})$ è detto **stimatore corretto di θ** se risulta $E[T(\mathbf{X})] = \theta$.

- popolazione: tutti gli studenti di ingegneria
- campione: gruppo di n studenti presi casualmente dalla facoltà di ingegneria

CARATTERISTICHE e valutazione di uno stimatore

($E[T(\mathbf{X})] - \theta$) è detto **errore sistematico (bias)**

è la differenza tra il valore atteso dello stimatore ed il vero valore del parametro. Uno stimatore corretto perciò ha un bias nullo

$MSE \hat{=} E[(T(X_1, \dots, X_n) - \theta)^2]$ è l'**errore quadratico medio**

Misura la variabilità dello stimatore attorno al vero valore del parametro. Un MSE basso indica uno stimatore più preciso

Stimatore consistente se:

capacità di avvicinarsi sempre più al vero valore del parametro che si vuole stimare all'aumentare della dimensione del campione

$\lim_{n \rightarrow \infty} \Pr(|T(X_1, \dots, X_n) - \theta| \geq \varepsilon) = 0$

stimatore
vero valore del parametro
valore molto piccolo
la stima diventa sempre più precisa man mano che si raccolgono più dati
"all'aumentare del campione"
probabilità che la stima T si discosti dal vero valore, diventa sempre più piccola

Valutare tutte le suddette caratteristiche con riferimento ad uno stimatore trovato e usato dagli statistici non è cosa facile, in generale!

Anticipazione di due risultati di analisi probabilistica utili adesso

Dimostreremo in una lezione futura che:

1. il valore atteso di una somma di variabili aleatorie è pari alla somma dei valori attesi di ciascuna, senza bisogno di assumere che quelle variabili siano indipendenti.
2. La varianza di una somma di variabili aleatorie è pari alla somma delle varianze di ciascuna se assumiamo che quelle variabili siano indipendenti.

La statistica “media campionaria”

E' detta media campionaria la var. al seguente :

$$\bar{X}(n) \hat{=} \frac{1}{n} \sum_{i=1}^n X_i$$

X_1 pari alla realizzazione
del primo run
della monte carlo
↓
singole osservazioni
del campione
→ dimensioni
del campione

Se le X_i sono identicamente distribuite e con lo stesso valore atteso, $E[X_i] \hat{=} \mu_i = \mu, i = 1, 2, \dots, n.$

Valore atteso
media campionaria

$$E[\bar{X}(n)] = \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{n \cdot \mu}{n} = \mu$$

⇒ $\bar{X}(n)$ e' uno stimatore corretto di μ

se calcolassi la media di tutte le possibili medie campionarie che potrei ottenere da tutti i possibili campioni, otterrei esattamente il valore della media della popolazione.

media della
popolazione

poiché le osservazioni
sono indipendenti e
identicamente distribuite il
valore atteso della somma
è uguale alla somma dei
valori attesi

il valore atteso di ogni
singola osservazione
è pari alla media
della popolazione

$$E[\bar{X}(n)] = \frac{1}{n} (\mu + \mu + \dots + \mu) = \frac{1}{n} \cdot n \cdot \mu$$

Passando alla varianza della media campionaria e assumendo che le X_i siano INDIPENDENTI e con la stessa varianza: $\sigma_i^2 = \sigma^2 \forall i$

indica quanto le medie campionarie di diversi campioni estratti dalla stessa popolazione variano tra loro

$$\begin{aligned} Var[\bar{X}(n)] &= \frac{1}{n^2} E\left[\left(\sum_{i=1}^n X_i - E\left[\sum_{i=1}^n X_i\right]\right)^2\right] = \\ &\quad \text{quadrato della dimensione del campione} \end{aligned}$$

varianza "bassa" allora
stima più precisa

I passaggi algebrici
saranno mostrati in
una futura lezione!

= ...

$$= \frac{\sigma^2}{n}$$

⇒ $\bar{X}(n)$ stimatore di bontà crescente con "n"
la varianza della media campionaria

La statistica “varianza campionaria” 1(2)

Siano X_1, X_2, \dots, X_n I.I.D. valori individuali di un campione di dimensione "n"
 $E[X_i]$ è il valore atteso della popolazione da cui è estratto il campione

$$\text{con: } E[X_i] \hat{=} \mu, \quad \text{Var}[X_i] \hat{=} \sigma^2, \quad i = 1, \dots, n$$

$$DA: \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \quad (\bar{X} \hat{=} \frac{1}{n} \sum_{i=1}^n X_i)$$

$$SI OTTIENE: E \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] = (n-1)\sigma^2$$

se non $E[Y] \hat{=} \int_{-\infty}^{\infty} y \cdot f_Y(y) dy$, con $Y \hat{=} a \cdot X$, $a > 0$ (esempio)

$$\begin{aligned} E[aX] &\hat{=} \int_{-\infty}^{\infty} ax \cdot f(x) dx \\ &= \int_{-\infty}^{\infty} ax \cdot (1/a)f(y/a) dy \\ &= \int_{-\infty}^{\infty} x \cdot f(y/a) dy \\ &= a \cdot E[X] \end{aligned}$$

Generalizzazione: con $g(x)$ idonea:

$$E[g(X)] \hat{=} \int_{-\infty}^{\infty} g(x) \cdot f(x) dx, \quad g(x) \geq 0$$

I passaggi algebrici sono stati indicati nella dispensa del corso, parte II, “Analisi Statistica”, pag. 24 e 25, ma NON fanno parte del programma di esame!

Si userà come stima della varianza S al posto di sigma

IN CONCLUSIONE: $E \left[\frac{1}{(n-1)} \sum_{k=1}^n (X_k - \bar{X})^2 \right] = \sigma^2$

“n-1” e non “n” perchè per ottenere una stima più accurata della varianza usiamo un “aggiustamento statistico”

NON SI TRATTA DI GAUSS

se calcolo la varianza campionaria su molti campioni diversi le media di questa varianza campionaria si avvicinerebbe sempre di più alla vera varianza della popolazione

Osservazione: La varianza campionaria è la stima puntuale della varianza.

Nella parentesi quadrata c’è uno stimatore corretto del parametro varianza, cioè la statistica “varianza campionaria”!

Formula alternativa usata in Excel per esprimere lo stimatore della varianza

https://www2.isye.gatech.edu/people/faculty/David_Goldsman/

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \\ &= \frac{\sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2)}{n - 1} \\ &= \frac{\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2}{n - 1} \\ &= \frac{\sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2}{n - 1} \\ &= \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n - 1} \end{aligned}$$

La statistica “varianza campionaria” 2(2)

In base al precedente risultato, la variabile aleatoria:

$$S^2 \triangleq \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1}$$

è detta “varianza campionaria” e viene usata per “stimare” “puntualmente” il parametro σ^2 grazie alla seguente (che non dimostreremo):

$$Var[S^2] = \frac{2\sigma^4}{n-1} \rightarrow 0 \quad per \quad n \rightarrow \infty$$

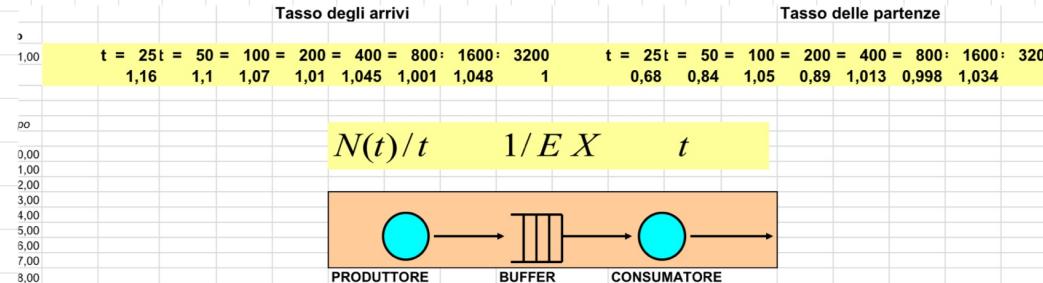
Osservazione:

Lo stimatore corretto del parametro varianza migliora al crescere di “n” (perché si riduce la sua varianza)

Però, purtroppo, questo stimatore ha una varianza sua che è pari al quadrato della varianza che deve stimare!

MODELLO PRODUTTORE-CONSUMATORE

Utilizzo il metodo Montecarlo per generare realizzazioni e modellare il modello client-server seguente:

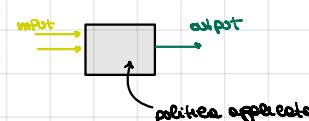


IN PIÙ...

Se gli interarrivi sono esponenziali, anche i tempi di servizio dunque gli input sono di carattere esponenziale. L'output invece non sarà di carattere esponenziale.

Dal punto di vista della simulazione monte carlo, posso riprodurre la storia degli stati di questo sistema nel tempo? Analizzo i due input del sistema: **flusso di arrivo** e la **caratteristica dei servizi**.

L'output è il tempo di risposta/attesa del sistema.
(variabile aleatoria, di cui serve il valore atteso e varianza)



Contatore	Tasso Interarrivi			Tasso di Servizio			Tempi di Soggiorno		
	$\lambda = 1$			$\mu = 1,1$			$E[S] = 1/\mu(1-p) = 10$		
	Valore Atteso in Input	Valore Atteso in Output	Ordine	Tempi di Servizio (1/ μ)	Media su 3200 Osservazioni	Tempi di Servizio	Valore Atteso Calcolato	E[S] = 1/ $\mu(1-p)$	Media su 3200 Osservazioni
1	Interarrivi (1/t)	1	FIFO	Tempi di Servizio (1/ μ)	0,909090909	Tempi di Servizio	0,907979337	10	Media su 3200 Osservazioni
2	Media su 3200 Osservazioni	Interarrivi	Interarrivi	Media su 3200 Osservazioni	0,958853087	Tempi di Soggiorno	13,92191877	10	Tempi di Soggiorno
3	0,958853087	0,958853087	0,958853087	0,958853087	0,958853087	0,958853087	0,958853087	0,958853087	0,958853087
4	Numeri Casuali	Tempi di Interarrivo	Istanti di Arrivo	Numeri Casuali	Tempi di Servizio	Istanti di Partenza	-0,760521178		
5	0,514899008	0,663784499	0,663784499	0,385433153	0,866715914	1,15760293			
6	0,887865793	0,118934682	0,782719181	0,319521399	1,037210029	2,567710442			
7	0,887994107	0,118790172	0,901509353	0,202910298	1,449992071	4,017702513			
8	0,706357031	0,34763446	1,249143813	0,275110822	1,173255702	5,199058215			
9	0,998920204	0,001080379	1,250224192	0,816787918	0,183978004	5,374936219			
10	0,285359542	1,254005343	2,504229535	0,293072189	1,115760293	6,490696512			
11	0,71201352	0,339658379	2,843887914	0,137924153	1,800955786	8,291652298			
12	0,695067289	0,363746619	3,207634533	0,177181899	1,573253088	9,864905386			
13	0,383131575	0,959376812	4,16701344	0,490920213	0,646794239	10,51169962			
14	10*	0,759480021	0,275121262	4,442132607	0,442506263	0,741182419	11,2528204		
15	11*	0,698349439	0,359035673	4,80116828	0,060268553	2,553586197	13,80646824		
16	12*	0,516391639	0,660889814	5,462058091	0,695429355	0,330205315	14,13667356		
17	13*	0,084551549	2,470393884	7,932451976	0,672166384	0,361135794	14,49780935		
18	14*	0,922743416	0,080404072	8,012856048	0,107281114	2,02936605	16,5271754		
19	15*	0,536793467	0,622141683	8,634997911	0,234827967	1,317183715	17,84435912		
20	16*	0,683086114	0,381134349	9,016132256	0,522356167	0,590368738	18,43472785		
21	17*	0,734663742	0,308342379	9,324474635	0,00711789	4,49558384	22,93031324		
22	18*	0,056506035	2,724087589	12,04856222	0,021809752	3,477634616	26,40794785		
23	19*	0,178758964	1,721716952	13,77027918	0,135614602	1,8161307476	28,22425533		
24	20*	0,538244808	0,619441788	14,38972096	0,296575003	1,104959209	29,32921454		
25	21*	0,92428958	0,078145797	14,46786676	0,230974006	1,33227366	30,6614419		
26	22*	0,571644757	0,559237535	15,0271043	0,897498275	0,988312801	30,75975471		
27	23*	0,057839343	2,849315255	17,87641955	0,513670597	0,605611892	31,3653666		
28	24*	0,750335687	0,287234591	18,16365414	0,361302241	0,925491307	32,2908579		
29	25*	0,496853655	0,699459753	18,86311389	0,767112548	0,241019774	36,83157768		
30	26*	0,090907199	2,397916082	21,26102998	0,803401428	0,199000709	32,73087839		
31	27*	0,131174381	2,03122769	23,29225767	0,998592447	0,001280495	32,7321588		
32	28*	0,851756029	0,160455144	23,45271281	0,302429418	1,087188507	33,81934739		
33	29*	0,538789346	0,618430607	24,07114342	0,288407013	1,130347779	34,94969517		

Parto dal generare numeri casuali per generare l'interarrivo. Per passare dal primo al secondo uso la trasformazione inversa e segno l'istante d'arrivo.

* Tramite un altro numero casuale genero un tempo di servizio, dove il primo tempo di servizio ottenuto è proprio quello della prima persona.

Utilizzo il generatore di numeri casuali **PERCHÉ VOGLIO GARANTIRE L'INDIPENDENZA**.

Gli istanti di partenza sono dati dalla somma dell'istante di arrivo più il servizio. Posso ricavarli anche ricorsivamente. Mi concentro sul processo delle partenze che è simmetrico rispetto al processo degli arrivi

NB d'immagine riportata è solo una parte di un unico file excel che io "spazzetto" per riportare meglio i concetti. "Sopra" il file da simula verso destra, dunque l'immagine successiva non sarà sotto che lo stesso file excel traslato più a destra

Tempo di Attesa in Coda		Modello analitico di riferimento		passo
Valore Atteso Calcolato	E[W]=p/[μ*(1-p)]=	9.090909091		
Media su 3200 Osservazioni				
Tempo di Attesa in Coda	13,01393943	13,01393943		
Tempi di Soggiorno	Tempi di Attesa in Coda	Equazione di Lindley W _i		
0,866715914	0	0	Pearson	
1,784991261	0,747781	0,747781232	Correlazione	
3,11619316	1,666201	1,666201089	-0,152677476	1,00
3,941814402	2,768559	2,7685587	-0,026784313	
4,124712027	3,940734	3,940734023	0,01868795	
3,986466977	2,870707	2,870706684	-0,026082237	
5,447764384	3,646809	3,646808598	0,002299292	
6,657270853	5,084018	5,084017765	-0,016934534	
6,344688281	5,697894	5,697894041	-0,027880259	
6,810749438	6,069567	6,069567018	-0,03519187	
9,005299962	6,451714	6,451713765		
8,674615465	8,34441	8,34441015		
6,565357375	6,204222	6,204221581		
8,514319352	6,484953	6,484953303		
9,209361204	7,892177	7,892177489		
9,418595597	8,828227	8,8282226859		
13,6058386	9,110253	9,110253218		
14,35938563	10,881751	10,88175101		
14,45397615	12,637669	12,63766868		
14,93949357	13,834534	13,83453437		
16,19357514	14,861348	14,86134778		
15,73265041	15,634338	15,63433761		
13,48894705	12,883335	12,88333515		
14,12720376	13,201712	13,20171246		
13,66876378	13,427744	13,42774401		
11,46984841	11,270848	11,2708477		
9,439901216	9,438621	9,438620721		
10,36663458	9,279446	9,279446072		
10,87855175	9,748204	9,748203972		
11,22180258	7,919041	7,919040612		
12,07879412	10,84444	10,84443974		
12,31984311	11,66901	11,66900972		
4,14173461	10,70210068	10,70210068		

Lindley: relazione ricorsiva tra il tempo di attesa del n-esimo e il tempo del (n-i)-esimo. Ad esempio relazione tra tempo di attesa della 10° persona (ultima) che attende il servizio e della 9°

Calcolo i tempi di attesa mediante l'equazione di Lindley

I tempi di soggiorno sono variabili aleatorie, il primo valore coincide con il servizio. Immagino al sequenza temporale di soggiorno (X₁, X₂...X_n), le X sarebbero le realizzazioni dei tempi di soggiorno a partire dal primo. Esempio: ogni mattina vado alla posta a vedere quanto tempo aspetta la terza persona in fila. Il terzo di ogni mattina rappresenta un campione, dunque soggiorno della terza persona.

→ se sei già qui clicco più volte sulla cornice delle soggiorno della 3° persona, otengo tante realizzazioni indipendenti (indipendenti perché al generatore di numeri casuali) allo stesso valore aleatorio "soggiorno" della terza persona (S₃)

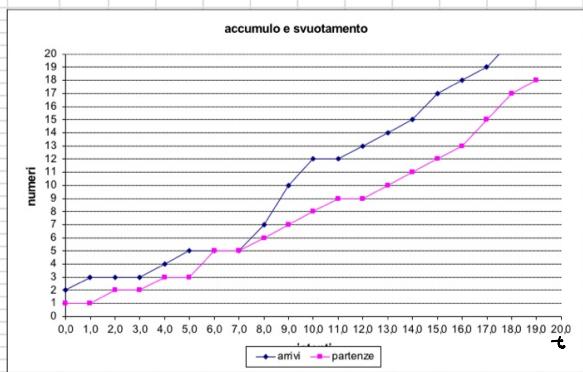
- Potrebbe servire la statistica? La realizzazione di S₃ rispetto a quella di S₅ è indipendente? O meglio, la sequenza dell'uscita del sistema di osservazioni è indipendente? Gli input per ipotesi sono indipendenti. La realizzazione di S₅ è correlata a quall di S₃. Le due realizzazioni sono identicamente distribuite? Concetto di stazionarietà (un qualcosa poi non cambia più)

se dovessi studiare questo sistema, spero che per $n \rightarrow \infty$ ci sarà una sola variabile aleatoria, ossia "soggiorno n-esimo", dopo 1000 o più realizzazioni (a lungo termine) diventerà una sola realizzazione. "Ne studio uno ed è come se li studiassi tutti"

- Come stimo la varianza? Se implemento la varianza campionaria sorge il problema della indipendenza e della correlazione (che si vedrà in futuro)

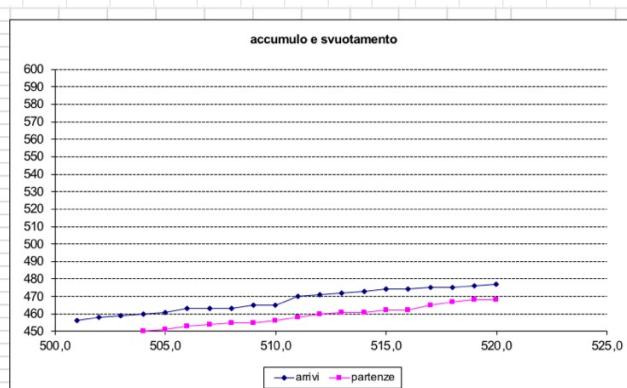
- Mi focalizzo sul tempo di attesa in coda (è una media aritmetica). Se clicco sulla sua casella mi rendo conto che salto da circa 10 a circa 5 con una "pigiata", quindi che varianza c'è? Come la stimo? Che "fiducia" posso dare a quel valore? Nascono le stime intervallari (intervallo di confidenza). Esse vengono utilizzate per stimare delle prestazioni. Per esempio ad un indice di prestazione (variabile aleatoria) come il tempo di risposta, quale valore atteso minimo "posso chiedere"?

Simulazione Monte Carlo: riproduco la storia degli stati. Di segui le traiettorie che rappresentano arrivi e partenze:



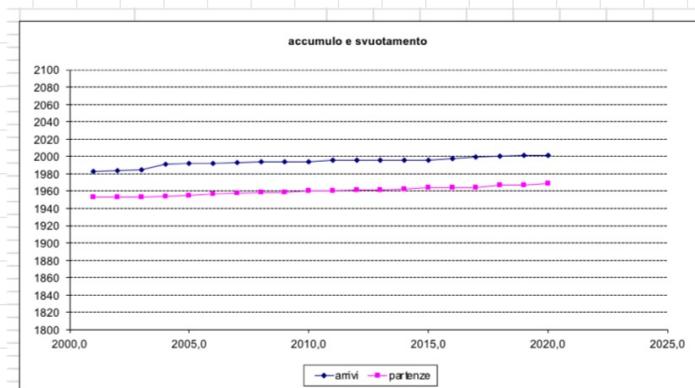
$0 \leq t \leq 20$ (inizio)

Simulo la traiettoria del processo stocastico degli arrivi e quello delle partenze. Inizialmente, non posso sapere se il sistema è congestionato o no



$500 \leq t \leq 525$ (metà)

Osservo il sistema a lungo termine

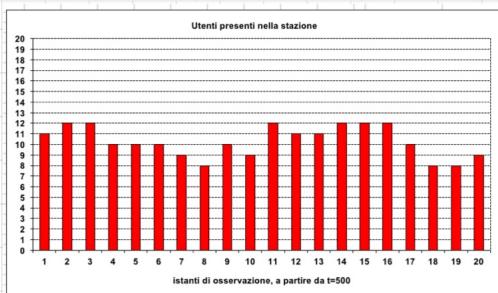
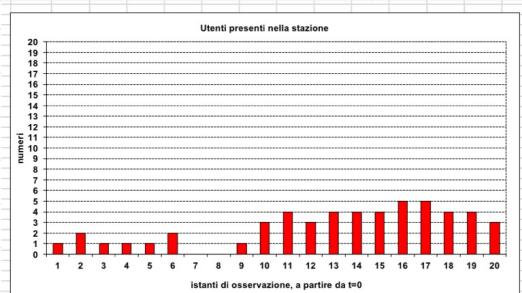


$2000 \leq t \leq 2025$ (verso la fine)

Quando il sistema arriva a regime il numero di arrivi e partenze tende a uguagliarsi

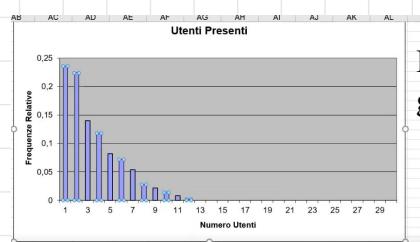
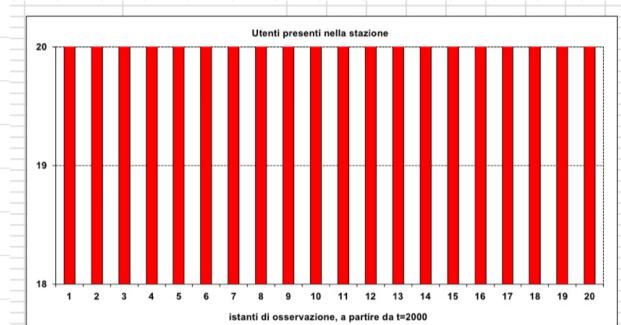
Arrivi (x_0, y_0) – Partenze(x_0, y_0) = numero di persone coinvolte

Tasso Interarrivi	1	Tasso di Servizio	1,1
Valore Atteso in Input		Valore Atteso in Input	
Interarrivi ($1/\lambda$)	1	Ordine	Tempo di Servizio ($1/\mu$)
Media su 3200 Osservazioni		FIFO	0,909090909
Interarrivi	0,958853087	Media su 3200 Osservazioni	
		Tempo di Servizio	0,907979337



In questo caso, $\text{TASSO DI SERVIZIO} > \text{TASSO INTERARRIVI}$
→ 10% più veloce

Cosa accade se cambio ciò?



Il grafico rappresenta proprio il modello geometrico di occupazione del buffer

PRODURRE STIME INTERVALLARI

casuale	LAMBDA	ESPOENZIALE	scarti	scarti^2		stima della media e della varianza	n	n-1	alfa	alfa/2	1-alfa/2
0,907059	0,25	B - tempo di soggiorno	A	A		A	100	99	0,05	0,025	0,975
0,071457	MEDIA	5,649488	31,91671								
0,475876	3,853689	-3,55713	12,6532								
0,620852		2,58411	-1,26958	1,61183							
0,881526	SOM SCA	3,879311	0,025624	21,88896							
0,242467	2152,032	8,532253	1,110751	-2,74294	7,523707						
0,62759		3,951039	0,097349	0,009477							
0,94516		11,61336	7,759661	60,21245							
0,210836		0,947125	-2,90565	8,448114							
0,315209		1,514568	-2,33912	5,471486							
0,750039		5,545806	1,692117	2,863259							
0,293832		1,39161	-2,46208	6,061832							
0,202667		0,905933	-2,94776	8,689265							
0,146378		0,633068	-3,22062	10,3724							
0,844105		7,434278	3,580589	12,82061							
0,571085		3,385983	-0,46771	0,218749							
0,3715		1,857675	-1,99601	3,984071							
0,997946		24,75135	20,89767	436,7124							
0,815052		6,750719	2,89703	8,392782							
0,010066		0,040468	-3,81322	14,54065							
0,256869		1,187533	-2,66616	7,10839							
0,464483		2,49809	-1,3556	1,83765							

LIVELLO DI "CONFIDENZA" o "FIDUCIA" (DALL'INGLESE "CONFIDENCE")

Il valore atteso del tempo di soggiorno, in precedenza era pari a: 11.055934. Potrebbe cambiare rendendo più complicato il modello se vengono applicate politiche differenti (round-robin, ecc.) In questi casi occorre fare delle simulazioni

In questo caso specifico, non sono sicuro di trovarmi di fronte ad un errore quindi faccio una stima intervallare.

Se volessi cambiare i dati, l'unico valore che potrei variare è n altrimenti cambierei il problema.

$z_{\alpha/2}$ valori della Gaussiana in corrispondenza del quantile $\frac{\alpha}{2}$
1.96 è l'ascissa entro cui realizzo $\frac{\alpha}{2}$

Non so come si produce una realizzazione da una normale perché non vale il metodo delle trasformazioni inverse. Utilizzo il metodo implementato da Excel

aumento la varianza da 10 a 20 per aggiornare i valori



Devo dimostrare che con probabilità 0,95 ottengo l'intervallo che contiene il valore VERO. Tale intervallo garantisce che al 95% (livello di confidenza) contiene il valore vero. Dunque il valore vero è equiprobabile e non si trova solo al centro.

$$\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2}$$

SOTTO QUALI
IPOTESI L'INTERVALLO
È CORRETTO?

LEZIONE
16

Stima centrale che oscilla tra il suo valore meno qualcosa ed il suo valore sommato a qualcosa

Gli intervalli di confidenza NON SONO simmetrici ma quello della media lo è. Ha come variabile la gaussiana Z

Quando si valuta il grado di confidenza entra in gioco una statistica a prescindere dalle realizzazioni con cui lavoro. Per il valore atteso è la gaussiana, per la varianza è la chi-square.

L'ampiezza dell'intervallo è proporzionale a "S". Se ho un'ampiezza "larga" e voglio diminuirla posso lavorare solo su "n" ossia sul carico di dati. Data la radice quadrata il numero di dati deve aumentare sostanzialmente

\bar{X} è la media campionaria

$\frac{S}{\sqrt{n}}$ è la deviazione standard della varianza campionaria

$\bar{X} \pm \frac{S}{\sqrt{n}} z_{\alpha/2}$ è l'intervallo aleatorio

Il TLC per la “media campionaria”

Nell’analisi statistica del modello PRODUTTORE - CONSUMATORE, è assai utile poter assumere che una realizzazione della media campionaria corrisponda alla media aritmetica di una sequenza di realizzazioni indipendenti della stessa variabile aleatoria, X , magari rilevate con osservazioni sperimentali, indipendenti (runs del Metodo Monte Carlo).

Infatti, particolarizzando il teorema limite centrale a questo caso, risulta:

in giallo sarebbe la standardizzazione della media campionaria $\tilde{Z}_n = \frac{\bar{X}(n) - E[\bar{X}(n)]}{\sqrt{VAR[\bar{X}(n)]}} = \frac{\bar{X}(n) - \mu}{\sqrt{\sigma^2 / n}}$ ricordando che il TCL riguarda la somma di variabili aleatorie.

che tende alla normale standard per $n \rightarrow \infty$ e, di conseguenza:

$$\bar{X}(n) = \frac{\sigma}{\sqrt{n}} \tilde{Z}_n + \mu \quad \text{tende ad essere distribuita come una legge normale di media } \mu \text{ e varianza } \sigma^2/n \text{ al crescere di } n$$

deviazione standard

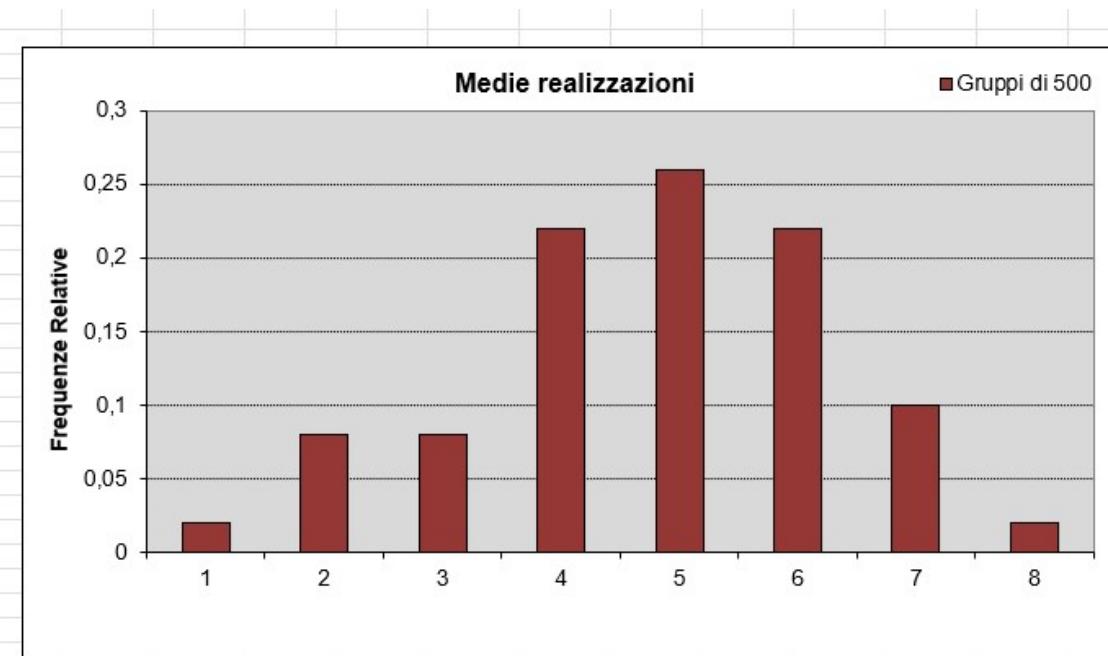
In particolare, si verifica sperimentalmente che la forma della $\bar{X}(n)$ diventa approssimabile ai fini pratici con la forma normale per $n > 30$

Per il modello produttore consumatore

Sperimentalmente aprodo il foglio excel: normalità della media campionaria si verifica il teorema del limite centrale

TESI:
La media campionaria tende a essere distribuita come una gaussiana centrata sul valore medio della legge esponenziale

Attraverso le generazioni Montecarlo, l'obiettivo di tale esperimento è stato quello di cercare di convalidare il teorema del limite centrale. Nella pratica, aumentando le n (dimensione del campione) ho cercato di approssimare fedelmente la "campana di Gauss". Con un numero di medie campionarie pari a 200 e poi ancora più con un numero di campioni pari 500, riesco a verificare sperimentalmente il suddetto teorema, ottenendo nelle varie simulazioni, sempre risultati più o meno attendibili. Ovviamente per campioni sempre maggiori l'attendibilità del teorema risulterà ancora più evidente.



Appunto più sale il gruppo di medie più si tende alla gaussiana. Oppure si poteva salire con il numero di realizzazioni!

La media di medie è detta grande media, il cui valore atteso della grande media è pari alla somma tra il valore atteso delle singole medie.

L'intervallo di confidenza per la media (valore atteso)

Siano X_1, X_2, \dots, X_n variabili aleatorie tutte indipendenti e identicamente distribuite

$\approx N(\mu, \sigma^2)$, allora:

$$\bar{X}(n) \hat{=} \frac{\sum_{i=1}^n X_i}{n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Ricorda che il valore atteso non è una media di perse ma perchè la media aritmetica lo stima!

per il “teorema di riproducibilità” della normale.

Rammentando che: (grazie all'indipendenza)

$$Z \hat{=} \frac{\bar{X}(n) - \mu}{\sigma / \sqrt{n}} \approx N(0,1) \quad \leftarrow$$

se inserisco S (deviazione della var. camp.) al posto della deviazione standard σ ottengo un rapporto tra variabili aleatorie e perciò una nuova variabile aleatoria

e che:

$$\Pr_{\text{quantile}}[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1 - \alpha \quad \leftarrow$$

dove:

$$\alpha/2 = \int_{-\infty}^{-z_{\alpha/2}} f_Z(z) dz = \int_{z_{\alpha/2}}^{+\infty} f_Z(z) dz$$

si ottiene: (isolando μ)

$$\Pr\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right] = 1 - \alpha$$

significa:
probabilità che la realizzazione di Z sia compresa tra i quantili è abbastanza vicino a 1

In pratica, è stato stabilito che l'intervallo aleatorio:

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right] \quad (\text{IC})$$

contiene, con probabilità $1 - \alpha$, il valore (incognito) del parametro μ .

Ogni realizzazione (intervallo numerico) di (IC) può essere considerata una stima di μ , ma non si può dire che essa contiene μ e tanto meno che il valore corrispondente a μ è il centro dell'intervallo: ogni punto dell'intervallo ha la stessa probabilità di essere μ .

(IC) è detto intervallo di confidenza al $100(1 - \alpha)\%$ (livello di confidenza).

Infine, posto $d \hat{=} z_{\alpha/2} \sigma / \sqrt{n}$, si tenga presente che la numerosità (n) del campione richiesto per stimare μ con un intervallo di ampiezza $2d$ risulta:

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{d^2}$$

Per progettare un intervallo di confidenza serve un parametro ed una statistica

OSSERVAZIONE: Per ottenere una realizzazione dell'intervallo di confidenza sulla media del processo, occorre una stima della varianza del processo!

Importanza dell'intervallo di confidenza:

- introduce l'idea di stimare un parametro non più attraverso la realizzazione di una variabile aleatoria, bensì attraverso la realizzazione di un intervallo aleatorio che contiene, con la probabilità voluta, il parametro stesso.

Qualità dell' intervallo:

- fissato il livello di confidenza, $100(1-\alpha)\%$, e con una certa deviazione standard (σ), intervalli migliori (cioè più ristretti) possono essere ottenuti solo aumentando di parecchio la numerosità (n) del campione.

Problemi d'uso:

- La stima per intervallo di un parametro del 1° ordine, quale la “media”, richiederebbe la conoscenza di un parametro del 2° ordine, quale la “deviazione standard”.
- Sembrerebbe valido solo nell'ipotesi che il parametro-media da stimare sia quello di una legge normale.

Soluzione dei problemi:

- Si può usare la varianza campionaria al posto della varianza vera e, sfruttando il teorema limite centrale, si può dimostrare che:

$$\Pr\left[\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2} \leq \bar{X} \leq \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2}\right] \cong 1 - \alpha \quad \text{per } n \rightarrow \infty$$

quando manca l'ipotesi di normalità delle variabili i.i.d. X_1, X_2, \dots, X_n .

L'intervallo: $\left[\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2}\right]$ è considerato una buona approssimazione dell'intervallo vero già con $n \geq 30$.

NOTA:

Nella pratica si usa fissare il livello di confidenza al 90% o 95% a ciò corrisponde: $\alpha = 0.10$ $\alpha = 0.05$ $z_{\frac{\alpha}{2}} = 1.645$ $z_{\frac{\alpha}{2}} = 1.960$

Approfondimento - varianza campionaria

Di seguito saranno dimostrati i seguenti risultati:

$$1. \quad \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

$$2. \quad E \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] = (n-1)\sigma^2$$

Risultato 1

Aggiungiamo e sottraiamo μ al primo membro; dall'identità:

$$X_i - \bar{X} = X_i - \bar{X} + \bar{X} - \mu = (X_i - \mu) - (\bar{X} - \mu)$$

si ottiene:

$$(*) \quad = \sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2 = \sum_{i=1}^n [(X_i - \mu)^2 - 2(\bar{X} - \mu)(X_i - \mu) + (\bar{X} - \mu)^2]$$

Possiamo distribuire la sommatoria sui diversi termini. Allora:

$$\sum_{i=1}^n (\bar{X} - \mu)^2 = n(\bar{X} - \mu)^2$$

poiché non dipende da i ed è sommato a se stesso n volte. Otteniamo:

$$(*) \quad = \sum_{i=1}^n (X_i - \mu)^2 - 2 \sum_{i=1}^n (\bar{X} - \mu)(X_i - \mu) + n(\bar{X} - \mu)^2$$

Ancora una volta $(\bar{X} - \mu)$ non dipende dalla sommatoria, dunque:

$$(*) \quad = \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + n(\bar{X} - \mu)^2$$

Poiché si era posto : $\bar{X} \hat{=} \frac{1}{n} \sum_{i=1}^n X_i \Rightarrow \sum_{i=1}^n X_i = n\bar{X}$, dunque:

$$\sum_{i=1}^n (X_i - \mu) = n\bar{X} - n$$

poiché μ non dipende da i . La (*) perciò diventa:

$$(*) \quad = \sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu)(n\bar{X} - n) + n(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^n (X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^n (X_i - \bar{X})^2 - n(\bar{X} - \bar{X})^2$$

come volevasi dimostrare.

Risultato 2

Si vuole dimostrare ora che:

$$E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = (n-1)\sigma^2 \quad (**)$$

Il valore atteso della somma è la somma dei valori attesi:

$$E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = E\left[\sum_{i=1}^n (X_i - \bar{X})^2 - n(\bar{X} - \bar{X})^2\right] = \sum_{i=1}^n E[(X_i - \bar{X})^2] - E[n(\bar{X} - \bar{X})^2]$$

Ricordiamo ora la formula della varianza di una variabile aleatoria:

$$VAR[Y] = E[(Y - E[Y])^2]$$

Allora, poiché le X_i sono tutte i.i.d. con:

$$E[X_i] \hat{=} \quad \text{e} \quad VAR[X_i] \hat{=} \sigma^2$$

la quantità: $E[(X_i - \bar{X})^2]$ è proprio la varianza delle X_i . Perciò:

$$\sum_{i=1}^n E[(X_i - \bar{X})^2] = n\sigma^2$$

Ma, poiché anche \bar{X} (la media campionaria) è una variabile aleatoria con:

$$E[Y] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \quad \text{e} \quad VAR[Y] = E\left[\left(\sum_{i=1}^n \frac{X_i}{n} - \bar{X}\right)^2\right] = \frac{\sigma^2}{n}$$

(con X_i variabili aleatorie i.i.d.). Di conseguenza:

$$E[n(\bar{X} - \bar{X})^2] = -n \cdot E[(\bar{X} - \bar{X})^2] = -n \cdot E\left[\left(\sum_{i=1}^n \frac{X_i}{n} - \bar{X}\right)^2\right] = -n \cdot \frac{\sigma^2}{n} = -\sigma^2$$

Perciò otteniamo:

$$(**) \quad E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

come volevasi dimostrare.

La distribuzione chi-quadrato

è una particolare Gamma ricordalo!

Come si è già detto, la distribuzione gamma, ha dato vita ad una particolare distribuzioni, dette chi-quadrato. Andiamo adesso a caratterizzarne la densità della variabile aleatoria chi-square. Considerando una variabile aleatoria X_γ^2 , detta chi-square la densità della chi-square è pari a:

$$f_X(x) = \frac{1}{\Gamma(\gamma/2)2^{\gamma/2}} x^{\frac{\gamma}{2}-1} e^{-\frac{x}{2}} \text{ con } x \geq 0$$

con $X \equiv X_\gamma^2 \equiv x_\gamma^2$ con γ "gradi di libertà". Su Excel la densità gamma è pari a:

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \text{ con } \beta = \frac{1}{\lambda} \quad \text{dunque è una gamma con: } \beta = 2 \alpha = \frac{\gamma}{2} \gamma \in [1, 2 \dots n]$$

x_γ^2 individua il punto della semiretta reale a partire dal quale l'area sottesa della densità è proprio γ , ovvero $Pr[X_\gamma^2 \geq x_\gamma^2] = \gamma$ (N.B. sta definizione nel powerpoint non c'è)

Si ricava che la media è la varianza sono pari a rispettivamente a: $E[X] = \gamma$ $Var[X] = 2 \cdot \gamma$
In particolare i gradi di libertà coincidono con il valore atteso. Si vuole dimostrare il seguente teorema: la variabile aleatoria "chi-quadrato" con 1 grado di libertà corrisponde al quadrato della "normale standard".

$$F_{Z^2}(z) \hat{=} \Pr\{Z^2 \leq z\} = \Pr\{-\sqrt{z} \leq Z \leq \sqrt{z}\} \quad \text{Sarebbe la dimostrazione di Trivedi}$$

$$= F_Z(\sqrt{z}) - F_Z(-\sqrt{z}) = 2 \int_0^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} \cdot \exp(-u^2/2) \cdot du$$

$$= 2 \int_0^{\sqrt{z}} \left(1/\Gamma(1/2)\sqrt{2}\right) \cdot \exp(-u^2/2) \cdot du$$

$$\left(\text{con } u \hat{=} \sqrt{v} \Rightarrow du = v^{-1/2} \cdot dv = v^{1/2-1} \cdot dv \right)$$

$$= \int_0^z \left(1/\Gamma(1/2)2^{1/2}\right) \cdot v^{1/2-1} \cdot \exp(-v/2) \cdot dv \quad \text{Ricordando che:}$$

$$f_{x_\gamma^2}(x) = \left(\frac{1}{\Gamma(\frac{\gamma}{2})}2^{\frac{\gamma}{2}}\right) \cdot x^{\frac{\gamma}{2}-1} \cdot e^{-\frac{x}{2}} \implies F_{x_\gamma^2}(z) = F_{Z^2}(z)$$

Teorema di riproducibilità della legge chi-quadrato (non dimostrato)

Siano: $X_{\gamma_1}^2, X_{\gamma_2}^2 \dots X_{\gamma_n}^2$ indipendenti, con i rispettivi gradi di libertà: $\gamma_1^2, \gamma_2^2 \dots \gamma_n^2$ allora:

$X_{\gamma_1}^2, X_{\gamma_2}^2 \dots X_{\gamma_n}^2$ è ancora una chi-quadrato, con gradi di libertà pari a: $\gamma_1, \gamma_2 \dots \gamma_n$

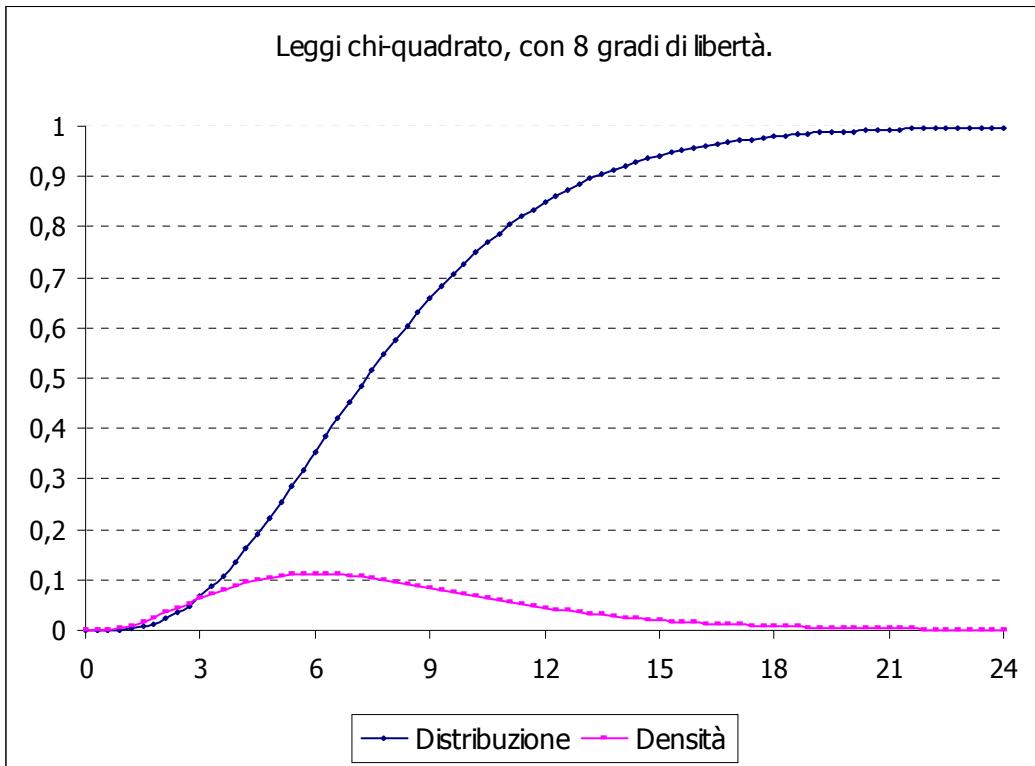
Attraverso le trasformate di Laplace ed al prodotto di convoluzione si prova il teorema.

IMPORTANZA: per "stimare" la varianza e la forma di una distribuzione, a partire da un insieme di realizzazioni sperimentali.

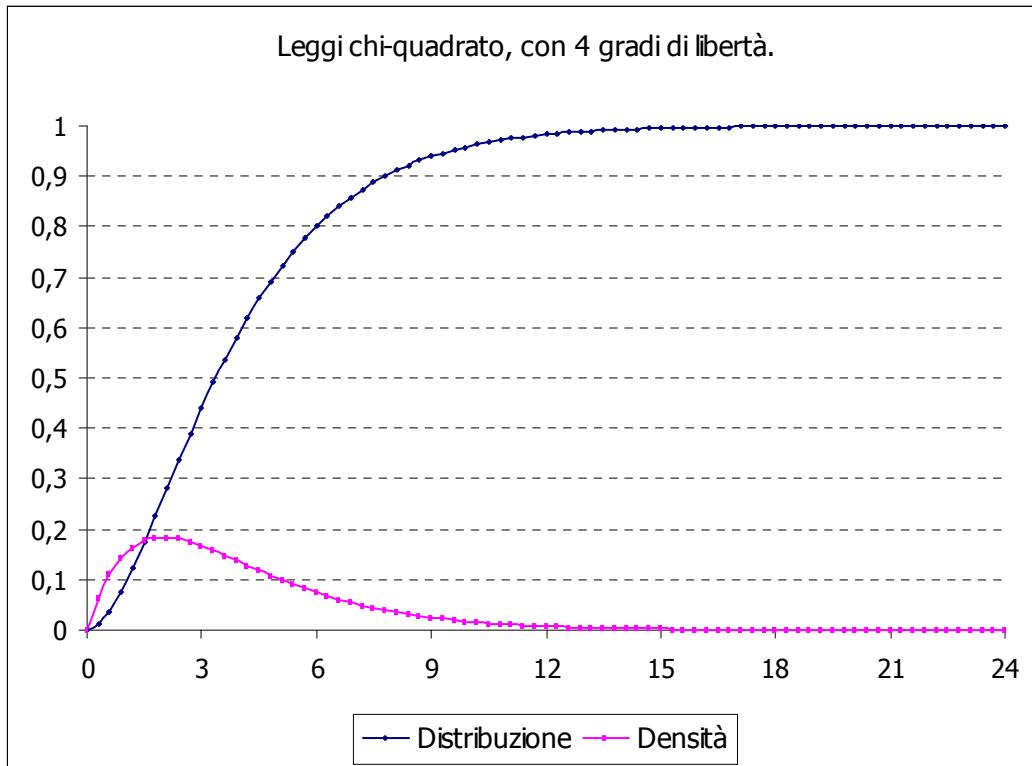
In parole povere il teorema afferma che la somma di chi-quadrato è ancora una chi-quadrato dove i gradi di libertà si sommano. Come detto è importante per "stimare" la varianza e la bontà della forma. (Si vedrà in futuro)

Bisognerebbe separare il grafico,
ad esempio il punto di intersezione
non si capisce ed una curva "sovrastra"
l'altra

Rappresentazione della chi-quadrato



da notare anche che non c'è nessun valore negativo. la chi-square è solo positiva!

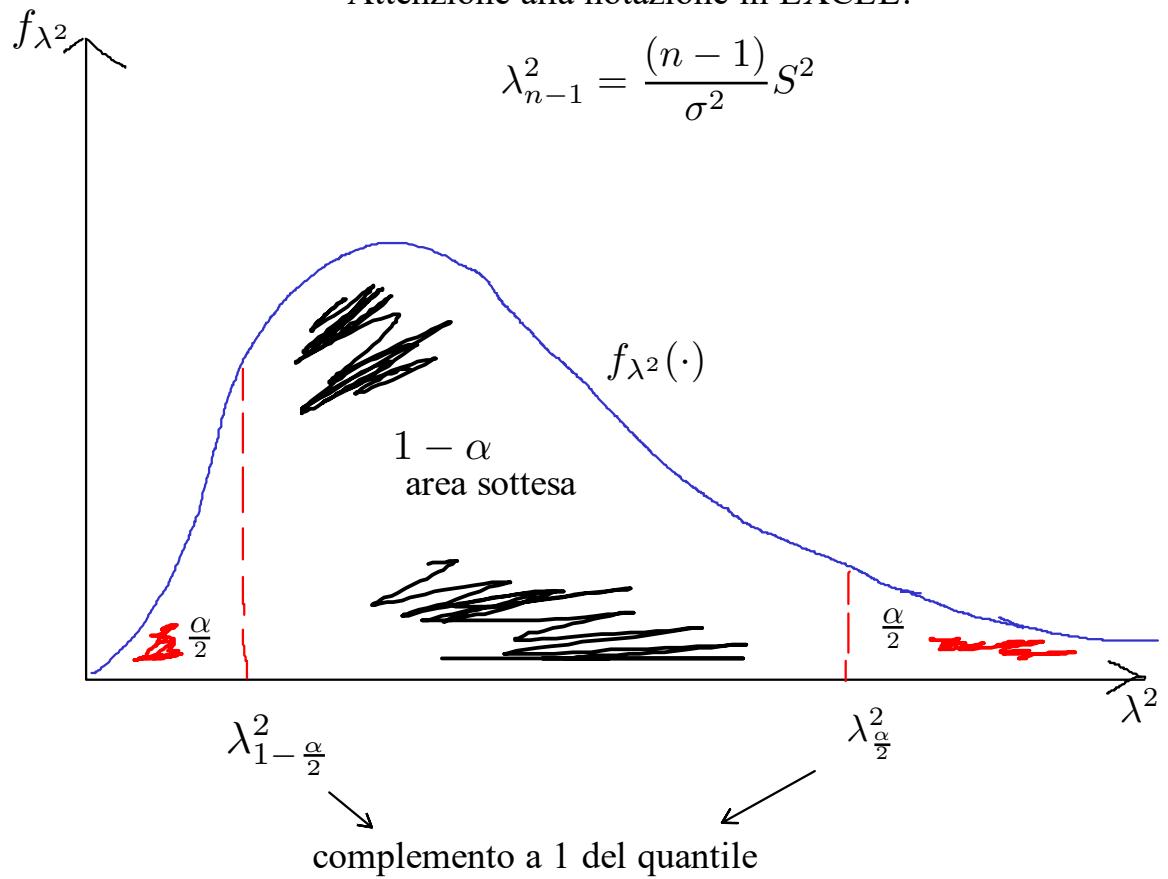


N.B. più aumentano i gradi di libertà più la chi-quadrato tende in forma ad una normale.
Vedi EXCEL

Attenzione alla notazione in EXCEL:

$$\lambda_{n-1}^2 = \frac{(n-1)}{\sigma^2} S^2$$

In rosso si ha l'area pari a $\frac{\alpha}{2}$



$\lambda_{\frac{\alpha}{2}}^2$ sarebbe l'area che si lascia a destra
ovvero il complemento del quantile

L'intervallo di confidenza per la varianza

Ripensando al ragionamento che ha condotto a “scoprire” un intervallo stimatore (detto, poi, intervallo di confidenza) per il parametro σ^2 , si dovrebbe riconoscere che è stata fondamentale la disponibilità di una variabile aleatoria di distribuzione nota che conteneva quel parametro:

$$Z \triangleq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$$

Dunque, volendo insistere su una strada analoga per individuare un intervallo di confidenza per un altro parametro importante quale la varianza, si deve cercare un'altra variabile aleatoria di distribuzione nota che contenga il parametro varianza e, se possibile, priva di ulteriori parametri che potrebbero creare complicazioni nell'uso pratico dell'intervallo, perché incogniti. Nel linguaggio degli statistici, si direbbe che si sta cercando una “statistica”, cioè una variabile aleatoria capace di produrre una stima.

Col seguente teorema si trova la statistica per la varianza:

Teorema (della chi-quadrato) Stessa ipotesi della media campionaria

Siano X_1, X_2, \dots, X_n indipendenti e identicamente distribuite $\approx N(\mu, \sigma^2)$, allora:

$$\text{statistica} \rightarrow \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 = \chi^2_{\gamma=n-1} \quad (\text{Tesi}) \quad \begin{aligned} &\text{ricordando la varianza campionaria} \\ &S^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2 \end{aligned}$$

La statistica è: $(n-1)S^2 / \sigma^2$. Infatti, grazie al teorema, si può scrivere:

$$\Pr[\chi^2_{1-\alpha/2} \leq (n-1)S^2 / \sigma^2 \leq \chi^2_{\alpha/2}] = 1 - \alpha$$

da cui: probabilità che la realizzazione della statistica cada nell'intervallo

$$\Pr\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \frac{\sigma^2}{S^2} \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1 - \alpha$$

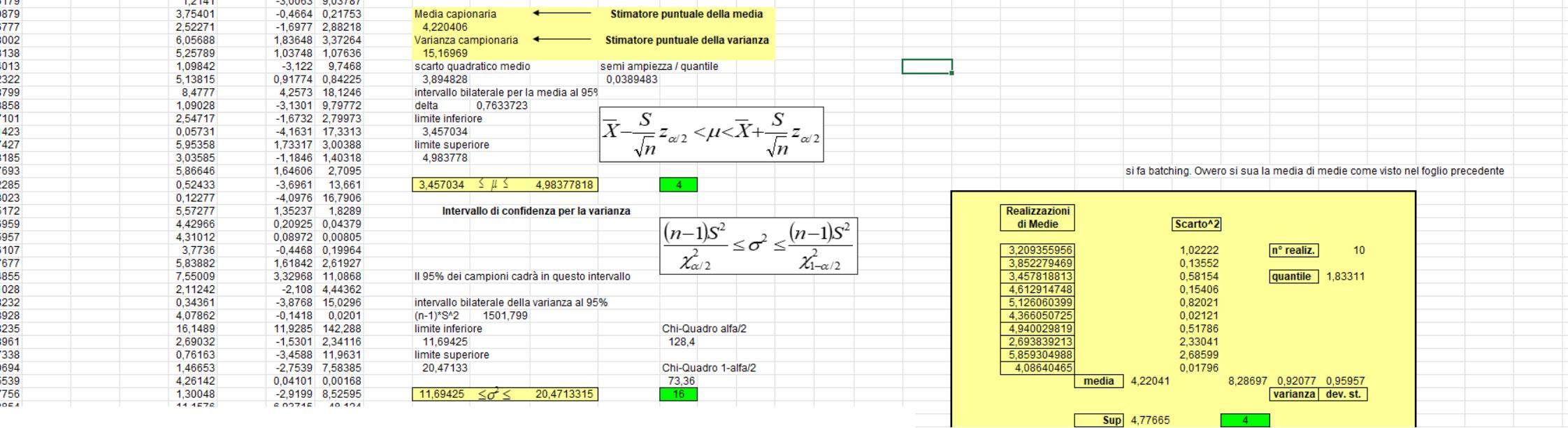
Quindi,

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \right]$$

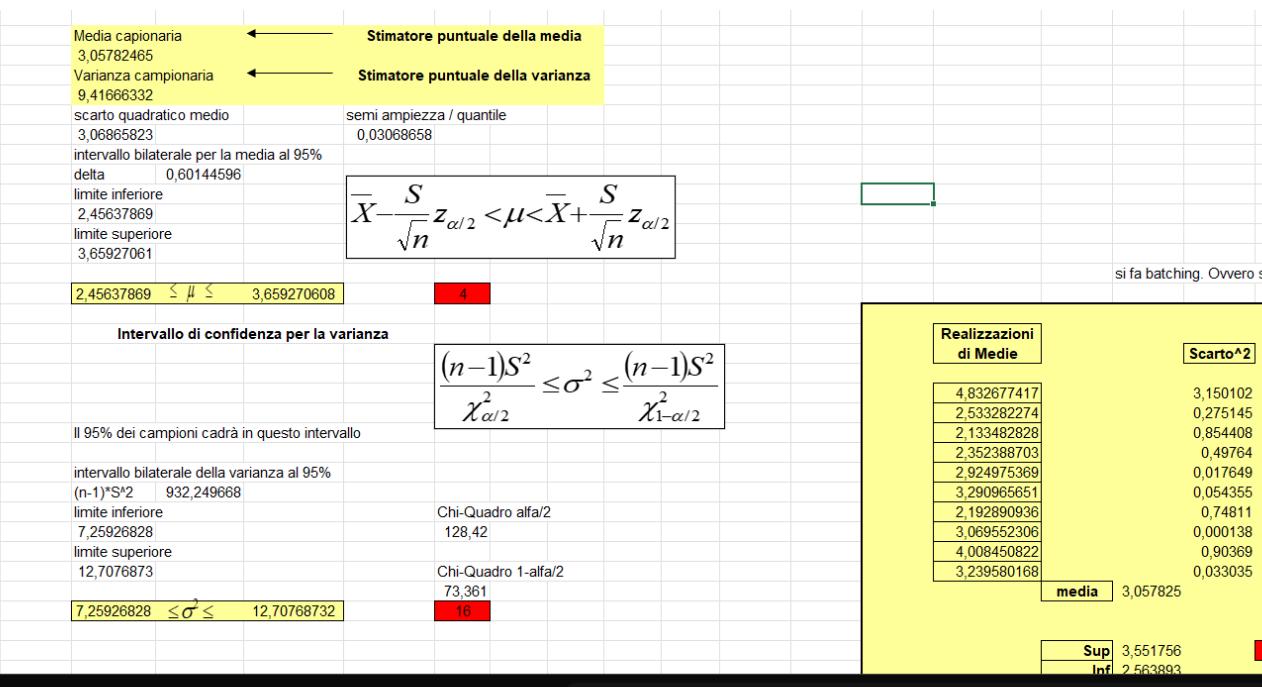
è l'intervallo di confidenza al $100(1-\alpha)\%$ per σ^2 .

Da notare che in particolare gli intervalli di confidenza sulla varianza a differenza di quelli sulla media campionaria, escono "larghi" già in partenza. Si possono stringere quanto si vuole ma resteranno comunque "larghi". Vedi file Excel

INOLTRE NON SONO SIMMETRICI COME NEL CASO DELLA MEDIA CAMPIONARIA
basta guarda i due denominatori dell'intervallo.



Nello screenshot si nota che gli intervalli sia per media che per varianza contengono il valore vero rispettivo. In particolare l'intervalllo per la media è sufficientemente largo, ma non simmetrico(!) dato il caso esponenziale. Mentre il caso della varianza è molto largo. Le stime sono uscite anche relativamente buone. Con un 4.22 a fronte di 4 e 15.19 a fronte di 16.



In questo altro caso invece, si hanno intervalli errati! Per la media si ha una stima di 3.05 a fronte di 4. ed un intervallo "stretto" di a 1.2 e spostato a sinistra del valore reale. Per la varianza si ha una stima di 9.4(!) a fronte di 16. L'intervalllo è largo 5.5 e spostato a destra del valore reale.

La diseguaglianza (o lemma) di Chebyshev

Sia Y la var. al. media campionaria di “n” variabili aleatorie indipendenti e identicamente distribuite, con un valore atteso comune (μ) e una varianza comune (σ^2).

$$\text{Dunque: } Y \hat{=} \frac{1}{n} \sum_{i=1}^n X_i \quad \text{con} \quad E[Y] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \mu \quad \text{e} \quad \text{VAR}[Y] = \frac{\sigma^2}{n} \quad (++)$$

Si è indotti a ritenere che la crescita di “n”, cioè della “numerosità del campione” dovrebbe produrre anche un effetto di crescente concentrazione delle realizzazioni della Y attorno alla sua media (μ), fino a far degenerare la forma a “campana” in una sorta di funzione limite del tutto speciale: con un solo valore all’infinito e nello stesso tempo di area sottesa unitaria!

Sorge, altresì, spontanea la seguente domanda: in che misura è possibile, al crescere di n, escludere il verificarsi di realizzazioni della Y che siano significativamente distanti dalla media?

La risposta alla domanda è fornita dal seguente lemma dovuto a *Chebyshev*, che vale per qualunque variabile aleatoria e sarà dimostrato facendo riferimento ad una Y continua tra $-\infty$ e $+\infty$ e arbitrariamente distribuita, purché di valore atteso $E[Y] \hat{=} \mu$ e varianza $Var[Y]$. La varianza è solitamente indicata come σ^2 , ma non ora per evitare confusione con la (++) .

Lemma di Chebyshev :

$$\Pr\left\{|Y - \mu| \geq k \cdot \sqrt{Var[Y]}\right\} \leq \left(\frac{1}{k^2}\right), \quad k \mid (1/k^2) < 1 \quad (\text{da fissare})$$

$$\text{Ovvero} \quad \Pr\left\{|Y - \mu| \geq h\right\} \leq \left(Var[Y]/h^2\right), \quad k \cdot \sqrt{Var[Y]} \hat{=} h \quad (\#)$$

Si ribadisce che il Lemma è riferito a variabili aleatorie continue e con realizzazioni comprese tra $-\infty$ e $+\infty$ (come la gaussiana, ad esempio, ma non necessariamente!)

La particolarizzazione dello stesso Lemma a variabili aleatorie continue e non negative (come la esponenziale, la Weibull e altre) non è banale, tant’è che è dovuta al matematico siciliano Francesco Paolo Cantelli!

$$\text{Lemma di Cantelli: } Y \in [0, +\infty) \Rightarrow \Pr\left\{Y \leq \mu + k \sqrt{Var[Y]}\right\} \leq 1 - (1/(1+k^2))$$

La qualità pratica del Lemma di Chebyshev, cioè la sua capacità di offrire un limite superiore abbastanza stretto è di solito insoddisfacente, ma la sua importanza teorica è notevole perché essa è alla base del concetto di “convergenza in probabilità” e, da qui, alla base della “legge debole dei grandi numeri” e del “teorema di Bernoulli”. La qualità pratica viene investigata nel file Excel associato a questa lezione. Ad esempio, con riferimento ad una gaussiana di media nulla, la probabilità vera che le realizzazioni siano contenute ad una distanza superiore a 2 deviazioni standard a destra e a sinistra dello zero è 0.046, ma la disegualanza di Chebyshev stabilisce solo che ciò accade con probabilità ≤ 0.25 !

PROVA del solo Lemma di Chebyshev:

Partendo dalla definizione di varianza, $\text{Var}[Y] = E[(Y - \mu)^2]$, quindi:

$$\text{Var}[Y] \hat{=} \int_{-\infty}^{+\infty} (y - \mu)^2 f(y) dy$$

vale la seguente disegualanza:

$$\text{Var}[Y] \geq \int_{-\infty}^{-\mu-h} (y - \mu)^2 f(y) dy + \int_{+\mu+h}^{+\infty} (y - \mu)^2 f(y) dy$$

perché si sta escludendo dal calcolo dell'integrale la porzione di area sottesa alla $f(y)$ che risulta compresa nell'intervallo di integrazione $[\mu-h, \mu+h]$. E si tratta di una area che è sicuramente ≥ 0 perché $f(y)$ è non negativa e quindi $(y - \mu)^2 f(y)$ è pure ≥ 0 ;

adesso:

$$\text{Var}[Y] \geq h^2 \left(\int_{-\infty}^{-\mu-h} f(y) dy + \int_{+\mu+h}^{+\infty} f(y) dy \right) \quad (\$)$$

perché quando la variabile di integrazione (y) scorre tra $-\infty$ e $(-\mu-h)$ risulta $< \mu$ e, analogamente, quando scorre tra $(\mu+h)$ e $+\infty$ risulta $> \mu$;

infine:

$$\text{Var}[Y] \geq h^2 \left(\Pr \{ |Y - \mu| \geq h \} \right) \quad \text{ovvero}$$

$$\Pr \{ |Y - \mu| \geq h \} \leq \text{Var}[Y]/h^2$$

perché il primo integrale nella (\$) calcola, per definizione, la $\text{Prob}[Y < (-\mu-h)]$ e il secondo calcola la probabilità la $\text{Prob}[Y > (\mu+h)]$. FINE PROVA.

Osservazione conclusiva:

Ritornando alla domanda sulla media campionaria, adesso si può rispondere che, al

crescere della numerosità del campione (n), diventerà sempre più piccola la probabilità

N.B. All'aumentare dei gradi di libertà la chi-quadrato arriva ad assomigliare alla forma di una normale. Vedi EXCEL

di osservare una realizzazione della media campionaria che abbia una distanza apprezzabile dal valore atteso . Precisamente, considerando la Y del Lemma di Chebyshev come una media campionaria con $\text{Var}[Y] = (\sigma^2 / n)$, si può riprendere la seconda forma (#) della disuguaglianza di Chebyshev:

$$\Pr \{ |Y - \vartheta| \geq h \} \leq \left(\frac{\text{Var}[Y]}{h^2} \right) = \left(\frac{(\sigma^2 / n)}{h^2} \right)$$

e fissando $h=\sigma$ (cioè il parametro “h” pari proprio alla deviazione standard delle “n” variabili componenti la media campionaria) si ottiene la seguente indicazione: $\Pr \{ |Y - \vartheta| \geq \sigma \} \leq 1/n$, che fa capire l’importanza della dimensione del campione.

Convergenza in probabilità e legge dei grandi numeri

Un concetto nuovo di convergenza, di tipo probabilistico, è chiaramente affiorato nei ragionamenti appena sviluppati, grazie alla disuguaglianza di Chebyshev, sulla media campionaria. La formalizzazione di esso porta alla definizione di convergenza in probabilità, nell’analisi probabilistica.

Definizione di convergenza in probabilità:

Si dice che una generica sequenza di variabili aleatorie Y_1, Y_2, \dots, Y_n converge in probabilità ad un valore finito ϑ e si scrive:

$$\lim_{n \rightarrow \infty} \Pr \{ |Y_n - \vartheta| = 0 \} = 1$$

quando, per qualsivoglia $\varepsilon > 0$, scelto arbitrariamente piccolo, esistono sempre una quantità $\delta > 0$ e uno specifico valore di n sufficientemente grande (\tilde{n}) tale che risulti verificata la seguente:

$$\Pr \{ |Y_{\tilde{n}} - \vartheta| < \delta \} > 1 - \varepsilon$$

Si osservi che la convergenza in probabilità è riferita alla *singola* quantità aleatoria $|Y_{\tilde{n}} - \vartheta|$ e non può essere estesa alla congiunzione logica (AND) di tutte le seguenti quantità:

$$|Y_{\tilde{n}+1} - \vartheta| \cap |Y_{\tilde{n}+2} - \vartheta| \cap |Y_{\tilde{n}+3} - \vartheta| \cap \dots$$

Dunque, questo tipo di convergenza non può escludere, al crescere di n oltre il valore \tilde{n} il verificarsi di realizzazioni apprezzabilmente distanti dal valore atteso con probabilità

finita. Cosa che è invece esclusa dalla versione “forte” della LGN (che sarà enunciata più avanti).

Legge (debole) dei grandi numeri

Nella formulazione di interesse per questo corso, la legge dei grandi numeri esprime il risultato della convergenza in probabilità della media campionaria Y al valore atteso finito, , di ognuna delle variabili identicamente distribuite, X_1, X_2, \dots, X_n , ma non necessariamente indipendenti, quando la dimensione del campione (n) tende all’infinito. Tale risultato di convergenza è più generale della cosiddetta convergenza “in forma” sul teorema limite centrale, perché è valido senza bisogno che le variabili X_1, X_2, \dots, X_n siano indipendenti.

Per sottolineare l’importanza della dimensione del campione si userà la seguente notazione:

$$Y(n) \hat{=} \frac{1}{n} \sum_{i=1}^n X_i, \quad \text{con} \quad E[Y(n)] = \frac{1}{n} \sum_{i=1}^n E[X_i] =$$

e rinunciando all’ipotesi di indipendenza, si scriverà soltanto:

$$VAR[Y(n)] = \frac{1}{n^2} VAR\left[\sum_{i=1}^n X_i\right] \hat{=} v^2(n)$$

e si farà solo l’ipotesi (meno restrittiva) che risulti : $\lim_{n \rightarrow \infty} v^2(n) = 0$

A questo punto, riscrivendo la diseguaglianza di *Chebyshev* nella forma complementare a quella ricavata prima:

$$\Pr\{|Y(n) - | < k v(n)\} > 1 - (v^2(n)/h^2) = 1 - (1/k^2)$$

e ponendo $\delta = kv(n)$ ed $\varepsilon = (1/k^2)$ si vede subito che l’ipotesi $\lim_{n \rightarrow \infty} v^2(n) = 0$

garantisce il rispetto della definizione di convergenza in probabilità della media campionaria al valore atteso .

Per completezza, viene di seguito viene riportata pure la dimostrazione della LGN debole sotto la ipotesi più restrittiva, che le variabili aleatorie in gioco siano indipendenti.

La diseguaglianza di Chebyshev:

$$\Pr \{ |Y(n) - \mu| \geq h \} \leq \left(\text{Var}[Y(n)] / h^2 \right) \quad \text{con} \quad (h \hat{=} n\sqrt{\text{Var}[Y(n)]})$$

viene riscritta in forma complementare:

$$\Pr \{ |Y(n) - \mu| < h \} \geq 1 - \left(\text{Var}[Y(n)] / h^2 \right)$$

Introducendo l'ipotesi di varianze finite ($\text{Var}[X_k] = \sigma_k^2$, $k=1,2,\dots,n$) e limitate tutte dalla stessa costante, C , risulta (per l'additività della varianza sotto ipotesi di indipendenza!):

$$\text{Var}[Y(n)] = \sum_{k=1}^n \sigma_k^2 < nC$$

sfruttando la precedente, si ottiene:

$$\Pr \{ |Y(n) - \mu| < h \} > 1 - \left(\frac{C}{n \cdot h^2} \right) > 1 - \varepsilon$$

e si arriva alla tesi:

$$\lim_{n \rightarrow \infty} \Pr \{ |Y(n) - \mu| < h \} = 1$$

Un caso speciale della legge debole dei grandi numeri (teorema di Bernoulli)

Si ottiene applicando la diseguaglianza di Chebyshev ad una variabile discreta, B , con legge binomiale di parametri n e p e, dunque, di media: np e varianza: $np(1-p)$.

Si osservi che risulta:

$$B = \sum_{i=1}^n X_i,$$

con X_i indipendenti e tutte bernoulliane di parametro p , ovvero risulta che B/n corrisponde ad una media campionaria sulle n variabili aleatorie che rappresentano gli esiti di n prove di Bernoulli (precisamente è la proporzione di successi su n prove).

Dunque:

$$\Pr \{ |B - np| < k\sqrt{np(1-p)} \} > 1 - \left(1/k^2 \right)$$

o meglio

$$\Pr \left\{ \left| \frac{B}{n} - p \right| < k \sqrt{\frac{p(1-p)}{n}} \right\} > 1 - \frac{1}{k^2}$$

da cui:

$$\Pr \left\{ \left| \frac{B}{n} - p \right| < \delta \right\} > 1 - \varepsilon \quad (+) \quad \text{con} \quad \delta \triangleq \sqrt{\frac{p(1-p)}{n\varepsilon}}, \quad \forall \varepsilon > 0$$

La (+) implica la convergenza in probabilità di B/n a p , mostrando in che misura il “rapporto tra casi favorevoli e casi possibili” possa approssimare la probabilità vera su un numero infinitamente grande di esperimenti indipendenti. A titolo di esempio, lanciando 1,000 volte (n) una moneta ($p=0.5$ per una qualunque delle due facce) e scegliendo di fissare $\delta=0.1$ si trova $\varepsilon = [p(1-p)/(n\delta^2)] = 1/40$; dunque la probabilità di sbagliare la stima del valore vero (0.5) di meno del 10% risulterebbe $>1-\varepsilon=39/40$.

Per altro verso, la (+) potrebbe anche essere riscritta come segue:

$$\Pr \left\{ \left| \frac{B - np}{n} \right| < \delta \right\} > 1 - \varepsilon$$

e vista così rivela che la (+) non può escludere il caso in cui la differenza $|B-np|$ - ovvero la differenza fra il numero di successi realizzati e il numero di successi attesi - diverge con la velocità della radice di “ n ” mentre “ n ” stessa cresce linearmente all’infinito!

Per eliminare questa debolezza della (+), ovvero rendere forte la (+), bisogna ricorrere al concetto di convergenza “forte” e al seguente teorema di Borel, basato appunto sulla convergenza forte:

$$\Pr \left\{ \lim_{n \rightarrow \infty} \left| \frac{B}{n} - p \right| = 0 \right\} = 1, \quad \text{e si dice che } \frac{B}{n} \rightarrow p \quad \text{"con prob. 1", per } n \rightarrow \infty$$

intendendo quanto segue: $\forall \varepsilon > 0$ e quindi $\delta > 0 \quad \exists \tilde{n}, \tilde{n}+1, \tilde{n}+2, \dots$

$$Prob \left[|Y(n) - p| < \delta \cap |Y(\tilde{n}+1) - p| < \delta \cap |Y(\tilde{n}+2) - p| < \delta \cap \dots \right] < 1 - \varepsilon$$

con $\tilde{n} \gg 0$

Questo risultato forte della legge dei grandi numeri garantisce che a partire da un “certo n ” (\tilde{n}) in poi le realizzazioni della variabile aleatoria B/n (proporzione di successi) che risulteranno tanto vicine a p quanto più si vuole corrisponderanno ad un sottospazio di Ω che ha probabilità 1 (convergenza “con prob 1”). In questo senso, il verificarsi di realizzazioni apprezzabilmente distanti da p è un evento che ha probabilità $1-P(\Omega)$, dunque nulla.

Legge dei grandi numeri “debole” e legge “forte”

Per comodità, viene qui ripetuta la formulazione della **LGN debole** d’interesse per questo corso:

IPOTESI_1:

-) X_1, X_2, \dots var. al. indip. e id. distr. con varianze $\sigma_1^2, \sigma_2^2, \dots$

$$\text{tali che } \sigma_1^2 < C, \sigma_2^2 < C, \dots, \sigma_k^2 < C, \dots \text{ con } C < \infty$$

IPOTESI_2:

-) X_1, X_2, \dots var. al. id. distr. con $\frac{1}{n^2} \text{VAR} \left[\sum_{i=1}^n X_i \right] \hat{=} v^2(n)$
sotto ipotesi che $\lim_{n \rightarrow \infty} v^2(n) = 0$

TESI comune:

$$\lim_{n \rightarrow \infty} \text{Prob} [|Y(n) - \mu| = 0] = 1$$

e s’intende così: $\forall \varepsilon > 0$ e quindi $\delta > 0$ $\exists \tilde{n}$ (che dipende da δ , che dipende da ε)

per il quale risulta: $\text{Prob} [|Y(\tilde{n}) - \mu| < \delta] > 1 - \varepsilon$, con

$$\tilde{n} \gg 0$$

E qui di seguito, per comodità di confronto, ecco la **LGN forte**:

IPOTESI_1:

-) X_1, X_2, \dots v. a. indip. e id. distr., con $E[X] = \mu$ comune a tutte.

IPOTESI_2:

-) X_1, X_2, \dots v. a. indip. $\begin{cases} \text{di media } \mu_1, \mu_2, \dots & \text{con } \sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} < \infty \\ \text{di varianza } \sigma_1^2, \sigma_2^2, \dots \end{cases}$

TESI comune:

$$\text{Prob} \left[\lim_{n \rightarrow \infty} |Y(n) - \mu| = 0 \right] = 1$$

e s’intende così: $\forall \varepsilon > 0$ e quindi $\delta > 0$ $\exists \tilde{n}, \tilde{n} + 1, \tilde{n} + 2, \dots$

$$\text{Prob} [|Y(n) - \mu| < \delta \cap |Y(\tilde{n} + 1) - \mu| < \delta \cap |Y(\tilde{n} + 2) - \mu| < \delta \cap \dots] < 1 - \varepsilon$$

$$\text{con } \tilde{n} \gg 0.$$

In alternativa al teorema di Borel ma per arrivare sempre allo stesso risultato pratico

Sia "X" la variabile aleatoria bernoulliana (real. 0 e 1) che indica il verificarsi o meno di un evento "A" e sia "n" il numero di esperimenti indipendenti.

$$\begin{aligned} \text{Siccome: } E[X] &= 0 \cdot \Pr(0) + 1 \cdot \Pr(1) \\ \Rightarrow \text{Prob}(A) &\equiv E[X] \end{aligned}$$

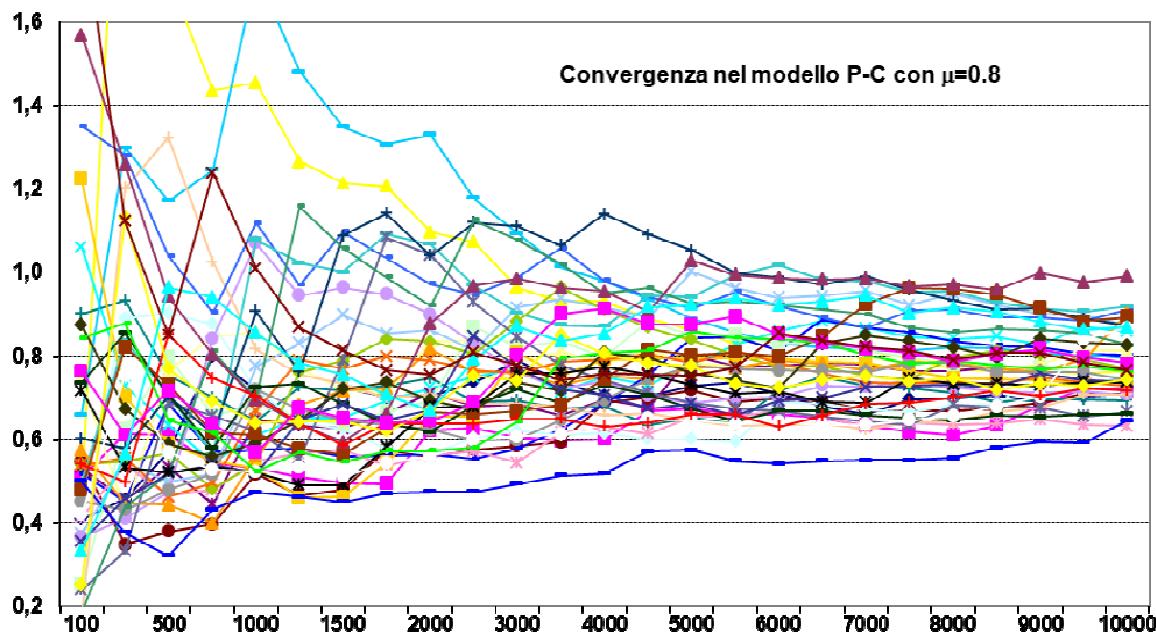
$$\text{allora: } \bar{X}(n) \hat{=} \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{\text{pr 1}} \text{Prob}(A), \quad n \rightarrow \infty$$

Il rapporto fra il numero di volte in cui si verifica l'evento A e il numero sufficientemente grande di esperimenti indipendenti effettuati può essere preso come la Prob(A).

La dimostrazione della LGN forte è oggetto di corsi di dottorato di ricerca.

Ma una illustrazione sperimentale della modalità di convergenza di una media campionaria si può proporre facendo riferimento al modello produttore-consamatore studiato e riprodotto in Excel (simulato) grazie al metodo Monte Carlo. La media campionaria di interesse è quella relativa alla sequenza di variabili aleatorie W_1, W_2, \dots, W_n che rappresentano, rispettivamente, il tempo di attesa del primo utente arrivato e servito, quello del secondo arrivato e servito e così via fino a $n=10,000$. Ripetendo 40 simulazioni in Excel, ognuna da 10,000 osservazioni, è possibile ottenere 400,000 numeri; questi possono essere organizzati in una matrice avente un numero di colonne pari alle 40 simulazioni e un numero di righe pari a 10,000. E s'intende che i 40 numeri che è possibile leggere su una riga comunque scelta sono proprio 40 realizzazioni (indipendenti) della variabile aleatoria W_k , se k è l'indice della riga scelta. Adesso, passando a ragionare su una colonna comunque scelta fra le 40 a disposizione, ad esempio la j -ma, si immagini di calcolare una sequenza di medie aritmetiche cosiddette "a cipolla" e siano esse $y_j(k)$. La media aritmetica a cipolla si può ottenere sommando i primi " k " numeri (ad esempio i primi 100) e poi i primi 500 e poi 1000: in

pratica ogni media contiene la precedente (come le foglie di una cipolla) e la espande fino a $k=10,000$. Si osservi che la scelta di passare dai primi 100 a 500 e poi 1,000 è fatta solo per arrivare a 10,000 senza avere troppi punti da rappresentare nella fase iniziale. A tale scopo si osservi il grafico riportato sotto, dove ogni media aritmetica a cipolla ha un colore diverso e tutte tendono a convergere al valore atteso vero (0.8 nel caso raffigurato). E tendono a convergere a dispetto del fatto che le medie aritmetiche a cipolla non siano realizzazioni indipendenti!



Certamente le medie a cipolla convergono in forma debole: a 10,000 nessuna dista più di $|0.2|$ da 0.8; ma allungando le simulazioni fino a 30,000 e oltre resta confermato che le traiettorie colorate si mantengono tutte dentro una distanza ancor più piccola e ancor meglio centrata attorno 0.8. Infine, si osservi che quelle raffigurate sono appunto 40 traiettorie del processo stocastico definito dalla sequenza di var. al. “Media camp. a cipolla su 100,..., Media camp. a cipolla su 10,000,”. Allora si può pure capire che con il termine “convergenza con probabilità 1 sull’insieme delle traiettorie possibili” si deve intendere che la modalità forte di convergenza attribuisce probabilità nulla ad una traiettoria che possa scappare via verso il basso o verso l’alto, lontano da 0.8, per $k > 10,000$.

Stime puntuale e stime intervallari

I metodi dell’analisi statistica prendono come riferimento i risultati empirici ottenuti da osservazioni scientifiche su sistemi reali, oppure – in fase di progettazione - da osservazioni riprodotte al computer (simulate) tramite l’approccio Monte Carlo. I metodi di analisi sono detti metodi di Inferenza Statistica e hanno il compito di ricavare (inferire) informazioni sulle leggi probabilistiche che governano la manifestazione/produzione di quelle osservazioni, reali o simulate che siano. È chiaro che le informazioni minimali riguardano i momenti del primo e del secondo ordine, ovvero la media e la varianza delle osservazioni attraverso le quali si manifesta il fenomeno d’interesse (ad esempio l’attesa per ottenere una risorsa). Più ambizioso è il compito di ricostruire la legge probabilistica vera e propria. Più ricorrente, non solo nell’ingegneria gestionale, è il compito di studiare due fenomeni attraverso il confronto fra medie e varianze delle osservazioni d’interesse. Per fissare le idee, basta pensare al modello “produttore – buffer – consumatore” e al compito di valutare il tempo medio d’attesa degli oggetti nel buffer, la probabilità di attendere troppo, la legge di occupazione del buffer e, ancora, di valutare le differenze (ad esempio in termini di media e di varianza sui tempi d’attesa) fra due politiche alternative di gestione delle operazioni di produzione e/o di consumo.

DEF.: Una **statistica** è una variabile aleatoria funzione di un numero fissato ($n \geq 1$) di altre variabili aleatorie, ma che non contiene alcun parametro incognito.

In pratica, “le altre variabili aleatorie” rappresentano osservazioni d’interesse e la statistica viene utilizzata per calcolare (stimare, nel linguaggio tecnico) un parametro incognito d’interesse riferito alle osservazioni d’interesse. Più precisamente, una realizzazione della statistica deve essere ottenuta immediatamente attraverso le realizzazioni delle osservazioni che la definiscono e la realizzazione media della statistica potrebbe essere usata per stimare il parametro incognito.

DEF.: Sia T una statistica per le variabili aleatorie i.i.d. $X_1 \dots X_n$ e sia θ un parametro incognito di queste ultime; allora $T(X)$ è detto **stimatore corretto di θ** se risulta $E[T(X)] = \theta$.

Quando risulta $E[T(X)] \neq \theta$, si è in presenza del cosiddetto **errore sistematico**, o **bias**, appunto pari alla differenza fra il valore atteso dello stimatore ed il parametro da stimare.

Se Y è la media campionaria, come già precedentemente definito, cioè

$$Y \triangleq \frac{1}{n} \sum_{i=1}^n X_i$$

e se le X_i sono tutte identicamente distribuite e indipendenti, con una media incognita
 $_i = \forall i$, allora:

$$E[Y] = \frac{1}{n} \sum_{i=1}^n E[X_i] =$$

Dunque: $T(X_1, \dots, X_n) \triangleq (1/n) \sum_{i=1}^n X_i$ è uno stimatore corretto per il parametro.

Allo stesso modo, passando alla varianza delle variabili aleatorie, $\sigma_i^2 = \sigma^2 \forall i$ e ricordando che risulta:

$$E\left[\frac{1}{(n-1)} \sum_{i=1}^n (X_i - Y)^2\right] = \sigma^2$$

si ottiene la seguente statistica S^2 , che è uno stimatore corretto per il parametro σ^2 :

$$S^2 \triangleq \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n-1}$$

DEF: Uno stimatore è **asintoticamente corretto**, se è corretto al tendere di n all'infinito.

DEF: L'**efficienza relativa** è il rapporto tra le varianze di due stimatori corretti.

DEF: Uno stimatore è detto **consistente** se converge in probabilità al parametro da stimare, ovvero: $\lim_{n \rightarrow \infty} \Pr(|T(X_1, \dots, X_n) - \theta| \geq \varepsilon) = 0$

La bontà di uno stimatore deve essere valutata anzitutto in termini della sua varianza: più bassa la varianza maggiore la qualità. Infine, si tenga presente che una statistica viene usata come stimatore di un parametro anche quando non risulta essere uno

stimatore corretto (se non si dispone di meglio). In tal caso assume importanza l'**errore quadratico medio**:

DEF: l'errore quadratico medio (MSE) di uno stimatore $T(X)$ di θ è il seguente:

$$MSE \hat{=} E[(T(X_1, \dots, X_n) - \theta)^2]$$

Più basso è questo valore maggiore è la qualità dello stimatore (comunque non corretto).

La legge forte dei grandi numeri e il metodo Monte Carlo

La legge forte dei grandi numeri esprime un risultato di convergenza matematica (quasi ovunque) della media aritmetica delle realizzazioni (indipendenti?) di una variabile aleatoria, X , al valore atteso della stessa variabile.

Formalmente:

$$Pr \left\{ \lim_{n \rightarrow \infty} \frac{1}{n} (x_1 + x_2 + \dots + x_n) = E[X] \right\} = 1$$

quando, fissato arbitrariamente $\varepsilon > 0$ e comunque piccolo, esiste sempre un valore \tilde{n} tale che:

$$\forall n > \tilde{n} \quad \text{risulti: } \left| \frac{x_1 + \dots + x_n}{n} - E[X] \right| < \varepsilon, \quad \text{"con prob. 1".}$$

Sul piano teorico, rappresentando con X (a valori 0 e 1) l'occorrenza o meno di un evento (A) e con "n" il numero di esperimenti indipendenti, è facile riconoscere che $P(A) = E[X]$ e quindi il rapporto tra il numero di volte ($x_1 + \dots + x_n$) in cui si osserva l'evento A ed il totale (n) degli esperimenti tende a stimare la $P(A)$ "con certezza".

Sul piano pratico, si vuole illustrare il metodo Monte Carlo per il calcolo approssimato del seguente integrale:

$$I \triangleq \int_a^b g(x) dx$$

A tal fine, s'introduce la variabile aleatoria "uniforme", $U[a,b]$, cioè definita da una densità uniforme nell'intervallo [a,b]:

$$f_U(u) \triangleq \begin{cases} (b-a)^{-1} & a \leq u \leq b \\ 0 & \text{altrimenti} \end{cases}$$

e si considera il seguente valore atteso, riferito non ad u ma a $g(u)$:

$$E[g(u)] \triangleq \int_a^b g(u) f_U(u) du$$

Osservando che risulta: $I = E[g(u)](b-a)$ e sfruttando la legge forte dei grandi numeri, si può stimare $E[g(u)]$ e quindi I con la seguente:

$$\frac{\sum_{i=1}^n g(u_i)}{n}$$

dove u_1, u_2, \dots, u_n è un campione sufficientemente grande di realizzazioni indipendenti della U .

Il metodo Monte Carlo rimanda al **problema di generare il campione** (u_1, u_2, \dots, u_n) di punti in corrispondenza dei quali calcolare i valori ($g(u_1), g(u_2), \dots, g(u_n)$) assunti dalla funzione integranda di I .

A tal fine si consideri il seguente ragionamento.

Sia $X\{0,\infty\}$ una generica variabile aleatoria a valori reali non negativi e con funzione di distribuzione $F_X(x)$ arbitraria, purché invertibile, e sia $U\{0,1\}$ una seconda variabile a valori reali compresi tra 0 e 1, con distribuzione uniforme. Si considerino adesso i valori reali ottenibili con la seguente operazione:

$$x \doteq F_X^{-1}(u), \text{ dove } u \text{ è una realizzazione della } U\{0,1\}$$

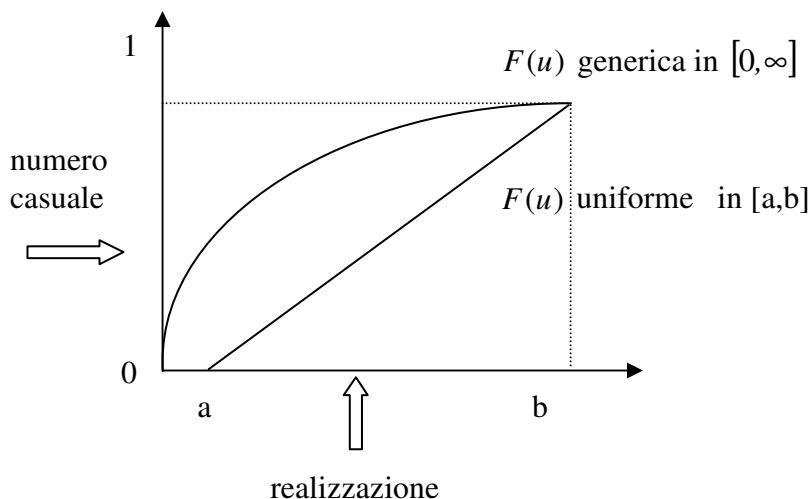
Allora risulta:

$$\Pr\{X \leq x\} \doteq F_X(x) = F_X(F_X^{-1}(u)) = u = \Pr\{U \leq u\}$$

e quindi anche la x ottenuta come $F_X^{-1}(u)$ è una realizzazione della $X\{0,\infty\}$.

Quello appena illustrato è il cosiddetto metodo Monte Carlo per la generazione di realizzazioni (indipendenti) di una variabile aleatoria, a partire da numeri (indipendenti) completamente casuali, compresi tra 0 e 1.

E' noto come metodo della "trasformazione inversa" ed è sintetizzato graficamente qui di seguito:



Il metodo della trasformazione inversa è fondato sulla disponibilità di un generatore di numeri casuali che determini il numero casuale compreso tra zero ed uno, con il quale posizionarsi sull'ordinata per poi leggere l'ascissa corrispondente.

Uso del metodo Monte Carlo

Generazione e ricostruzione della legge esponenziale

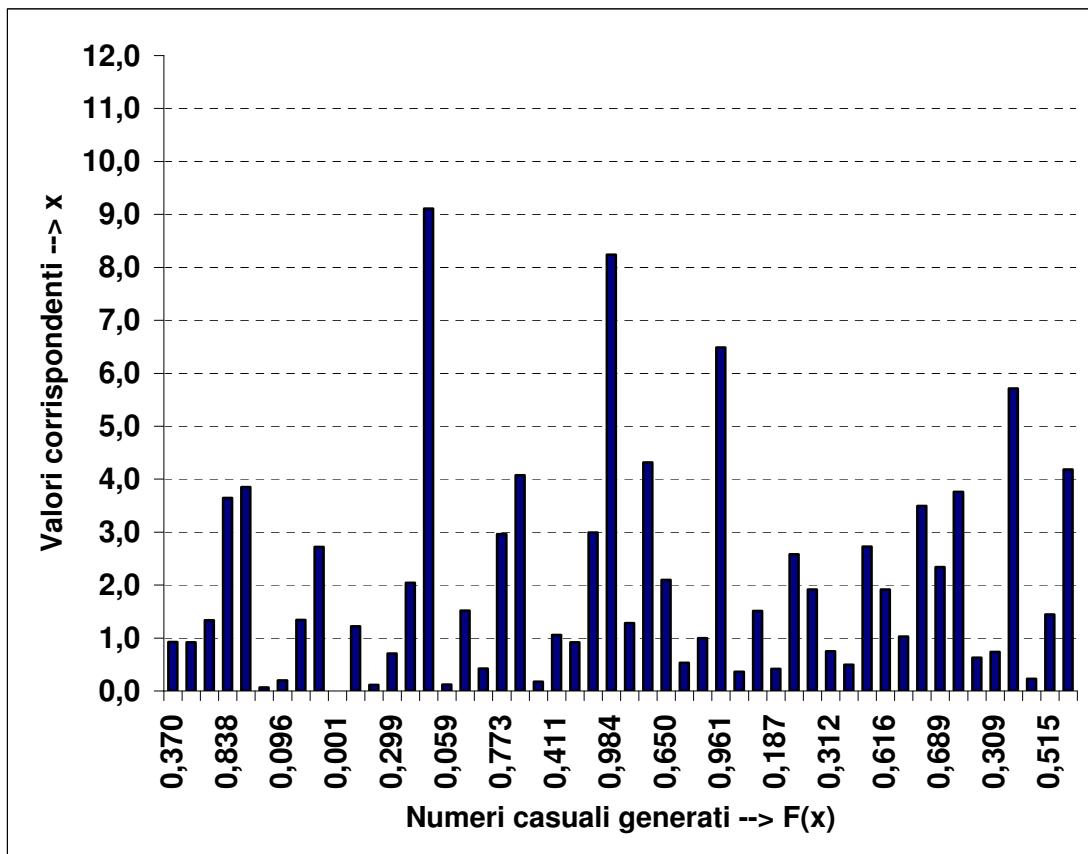
Sfruttiamo il metodo Monte Carlo per generare valori casuali della legge esponenziale:

$$F_X(x) = 1 - e^{-\lambda x}$$

Poiché essa è invertibile, possiamo avvalerci del metodo della trasformazione inversa per ottenere, dal valore $F_X(x)$, la realizzazione x della legge esponenziale:

$$x = -\frac{\ln(1 - F_X(x))}{\lambda} \quad \text{o ancora:} \quad x = -\frac{\ln(F_X(x))}{\lambda}$$

poiché se $F_X(x)$ è un numero casuale, anche $1 - F_X(x)$ lo sarà, e la nostra analisi non perde di generalità. Utilizzando la funzione *casuale()* di Excel e svolgendo le trasformazioni inverse ricaviamo 50 realizzazioni della legge esponenziale, considerando, ad esempio, un tasso $\lambda=0,5$.



Realizzazioni casuali dalla legge esponenziale di media=2.

Di tali realizzazioni si vuole ora ricostruire la legge (esponenziale) che le genera, illustrandola tramite un istogramma.

Un istogramma è una forma di aggregazione nel discreto e di rappresentazione grafica di un campione di realizzazioni sperimentali (osservate nella realtà o generate col metodo Monte Carlo) di una variabile aleatoria continua. Se si dispone anche di una funzione densità per la variabile aleatoria continua di interesse, si può trasformare in istogramma anche quest'ultima al fine di effettuare un confronto (di prima approssimazione) fra i due istogrammi. Ovviamente, la funzione di ripartizione delle probabilità di una variabile aleatoria discreta è già un istogramma.

Per realizzare un istogramma a partire da una funzione densità nota e supposta continua nel seguito, si divide il dominio delle realizzazioni della variabile aleatoria continua in k intervalli, adiacenti e tutti con la stessa ampiezza, Δb . Quindi, detta h_j la proporzione delle realizzazioni della variabile aleatoria X che risultano contenute nell'intervallo j , si definisce la seguente funzione $h(x)$:

$$h(x) \hat{=} \begin{cases} 0 & \text{se } x < 0 \\ h_j & \text{se } b_{j-1} < x \leq b_j \\ 0 & \text{se } x > b_k \end{cases} \quad \Delta b \hat{=} b_j - b_{j-1}$$

dove $h_j \hat{=} P(b_{j-1} < X \leq b_j)$, quale istogramma associato alla variabile aleatoria continua X .

Essendo, per ipotesi, nota la funzione densità $f(x)$, è chiaro che ognuno dei valori dell'insieme:

$$\{h_j, j = 1, \dots, k\}$$

si calcola con la seguente: $P(b_{j-1} < X \leq b_j) = \int_{b_{j-1}}^{b_j} f(x) dx$.

Il teorema del valore medio per funzioni continue in un intervallo chiuso garantisce che esiste un valore $\bar{x} \in (b_{j-1}, b_j]$ per il quale risulta:

$$\int_{b_{j-1}}^{b_j} f(x) dx = \Delta b \cdot f(\bar{x})$$

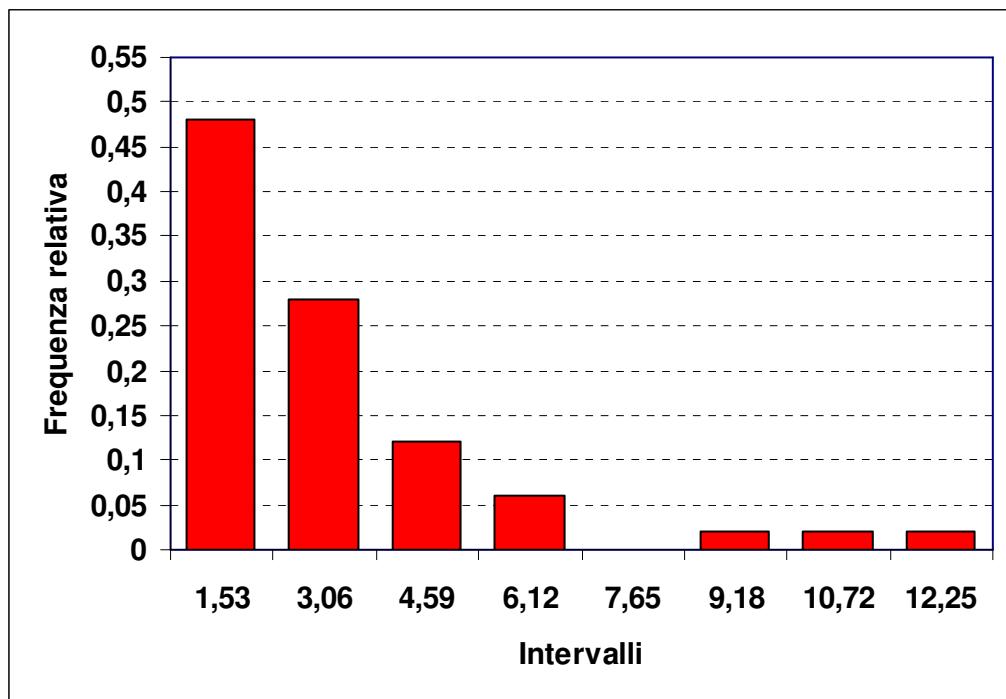
e in definitiva $h_j \approx \Delta b \cdot f(\bar{x})$, ovvero, la funzione costante a tratti riproduce la forma della funzione densità ed è detta istogramma.

Discende dal ragionamento svolto che quando l'istogramma è costruito con un campione di realizzazioni osservate l'insieme di valori $\{h_j, j = 1, \dots, k\}$, può essere usato per stabilire se quelle realizzazioni possono essere considerate "a vista" e in prima approssimazione come realizzazioni che potrebbero risultare in accordo con la densità $f(x)$, oppure devono essere considerate in evidente disaccordo.

La maggiore difficoltà nell'uso di questo metodo riguarda la scelta del numero k degli intervalli. La regola migliore conosciuta è quella di Sturges che suggerisce di scegliere k tramite la formula seguente: $k = \lfloor 1 + \log_2 n \rfloor$.

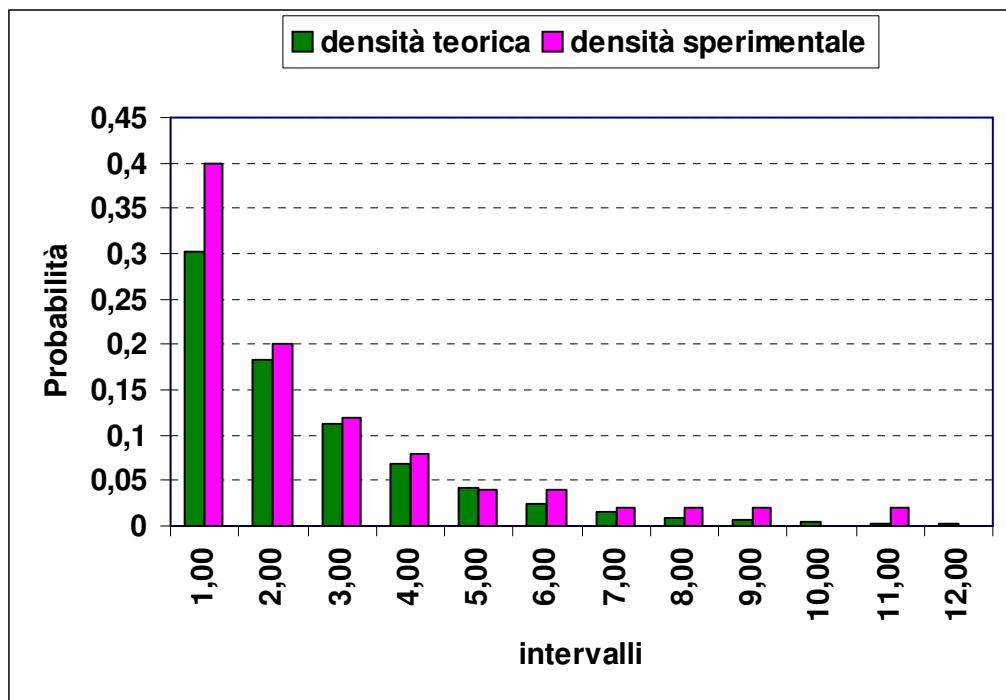
Spesso, però, conviene provare differenti valori per Δb , e scegliere quello che produce l'istogramma che maggiormente si avvicina a qualche distribuzione nota. La scelta di Δb può essere causa di errori. Infatti se Δb è scelto troppo ampio i dati possono essere aggregati in maniera grossolana perdendo molte informazioni, se invece Δb è troppo piccolo la varianza di h_j risulta eccessivamente alta.

La figura seguente mostra l'istogramma delle realizzazioni della legge esponenziale.



Ricostruzione legge esponenziale di media 2.

È interessante effettuare un confronto tra la densità teorica estratta dalla legge esponenziale e la densità sperimentale ricostruita a partire dalle realizzazioni casuali date dal metodo Monte Carlo, scelte entrambe con lo stesso parametro λ .



Confronto tra densità teorica e sperimentale ($\lambda = 0,5$)

Si noti che per la costruzione di questi istogrammi si è utilizzata una procedura di aggregazione analoga a quella appena descritta, ma più semplice, che fissa il numero di classi di aggregazione ed un'ampiezza arbitraria, cioè non calcolata a partire dal dominio delle realizzazioni che abbiamo ricavato.

Possiamo notare come i dati sperimentali non seguono perfettamente la densità teorica, sebbene provengano dalla stessa distribuzione. In casi meno fortunati la discrepanza è ancora più accentuata, al punto da generare il dubbio che si tratti di dati provenienti dalla stessa distribuzione. La teoria dei test di bontà della forma, che sarà trattata in seguito, si occupa di definire metodi formali per giudicarne, probabilisticamente, l'appartenenza.

Esistono, comunque, anche metodi euristici di confronto fra le distribuzioni di probabilità associabili a campioni di dati teorici ed empirici. Questi sono trattati in appendice.

Svolgimento esempio di applicazione del metodo Monte Carlo

Dati i seguenti numeri casuali:

0,03704	0,67611	0,26433	0,16358	0,75976
0,80568	0,82472	0,81805	0,31089	0,44881
0,00094	0,43072	0,35101	0,4359	0,14326
0,1244	0,89574	0,25408	0,57721	0,98016
0,34384	0,11242	0,32774	0,89906	0,90705
0,3327	0,70591	0,99781	0,70618	0,59835
0,6219	0,92303	0,32425	0,6313	0,18441
0,31356	0,69295	0,66009	0,91451	0,64431
0,99347	0,29734	0,38779	0,64861	0,25429
0,58823	0,24593	0,49337	0,56793	0,83636

- A) generare con il metodo Monte Carlo le realizzazioni di una v.a. distribuita con legge esponenziale di parametro 0,5;
- B) costruire un istogramma delle realizzazioni ottenute al punto precedente.

Soluzione

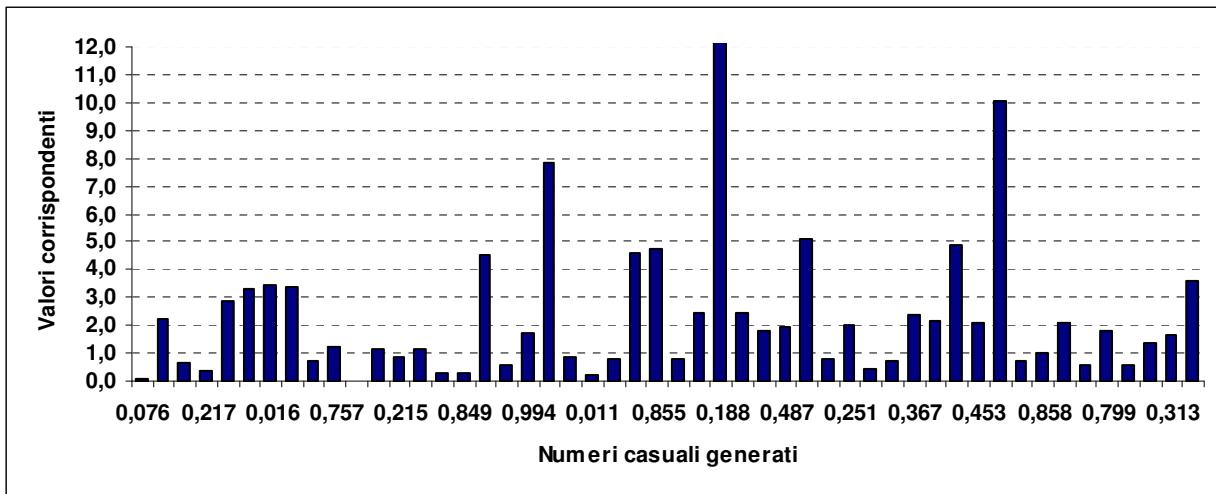
Applicando il metodo della trasformata inversa della funzione di distribuzione esponenziale:

$$x = -\frac{\ln(1-r)}{\lambda} \quad \text{con } r \in U(0,1)$$

si ottiene:

0,07549	2,25469	0,61394	0,35726	2,85223
3,27653	3,48278	3,40808	0,7447	1,19136
0,00188	1,12678	0,86466	1,14503	0,30923
0,26568	4,5217	0,58629	1,72175	7,83992
0,84272	0,23852	0,79421	4,58651	4,75135
0,80905	2,44777	12,248	2,44957	1,82436
1,94521	5,12877	0,78385	1,99553	0,40769
0,75248	2,36148	2,15812	4,91878	2,0674
10,0614	0,70576	0,98134	2,09174	0,58684
1,77459	0,56454	1,35994	1,67833	3,62023

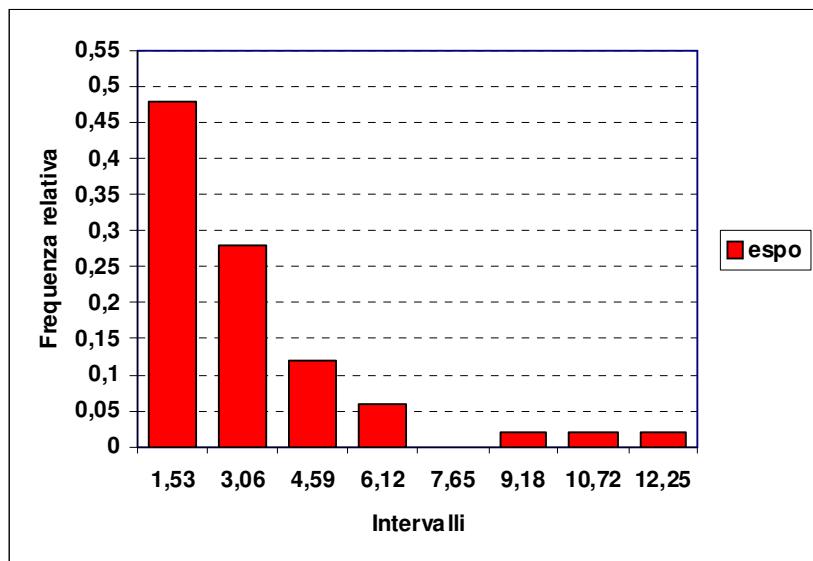
Segue il grafico delle realizzazioni.



Scegliendo di utilizzare 8 intervalli con il valore minimo = 0,001878 e il valore massimo = 12,24797, dalla formula $(\text{max}-\text{min})/8$ si ottiene un'ampiezza di classe pari a 1,530762. A partire da questo risultato, si ottiene il limite superiore (LS) di ciascuna classe e si possono calcolare le frequenze assolute (FA) e le frequenze relative (FR):

classe	LS	FA	FR
1	1,53076	24	0,48
2	3,06152	14	0,28
3	4,59229	6	0,12
4	6,12305	3	0,06
5	7,65381	0	0
6	9,18457	1	0,02
7	10,7153	1	0,02
8	12,248	1	0,02

Infine si disegna l'istogramma.



Metodo della trasformazione inversa per la legge di Erlang

La v.a. $X \hat{=} X_1 + \dots + X_n$ espressa come convoluzione di n esponenziali identiche e indipendenti di parametro λ

$$F_{X_i}(x_i) = 1 - e^{-\lambda x_i} \quad i = 1..n$$

si distribuisce in accordo alla legge di Erlang:

$$F_X(x) = 1 - e^{-\lambda x} \sum_{i=0}^{n-1} \frac{(\lambda x)^i}{i!} \quad \lambda > 0.$$

Richiamando che per generare una realizzazione della funzione esponenziale occorre prima generare un numero casuale compreso nell'intervallo $[0,1]$ e poi ricavare il valore della realizzazione mediante l'opportuna formula inversa ovvero:

$$u_i = 1 - e^{-\lambda x_i} \quad \rightarrow \quad x_i = -\frac{1}{\lambda} \ln(1 - u_i)$$

analogamente, per la legge di Erlang

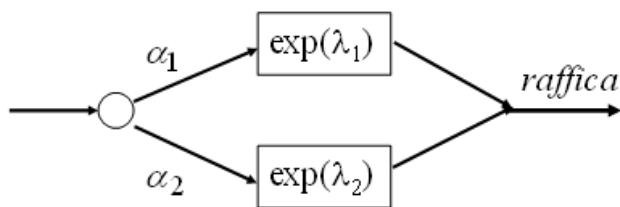
$$x = \sum_{i=1}^n x_i = \sum_{i=1}^n -\frac{1}{\lambda} \ln(1 - u_i) = -\frac{1}{\lambda} \ln \left(\prod_{i=1}^n (1 - u_i) \right).$$

Metodo della trasformazione inversa per la legge Iperesponenziale

Una legge iperesponenziale è una combinazione convessa di n funzioni esponenziali diverse di parametro $\lambda_i \quad i = 1..n$ ovvero

$$F_Y(y) = \sum_{i=1}^n \alpha_i (1 - e^{-\lambda_i y}) \text{ con } \lambda_1 \neq \lambda_2 \neq \dots \neq \lambda_n \text{ e } \sum_{i=1}^n \alpha_i = 1, \quad \alpha_i > 0 \quad i = 1..n.$$

Nel caso $n = 2$, con $\alpha_1 = 1 - \varepsilon$, $\alpha_2 = \varepsilon$ e $\lambda_1 \gg \lambda_2 > 0$ si ottiene un effetto “raffica” come mostrato in figura.

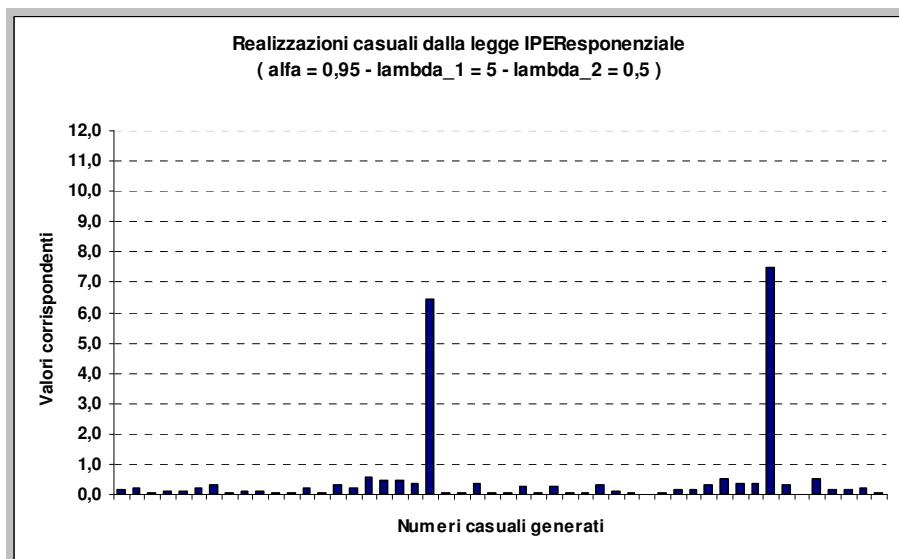


Procedura

```

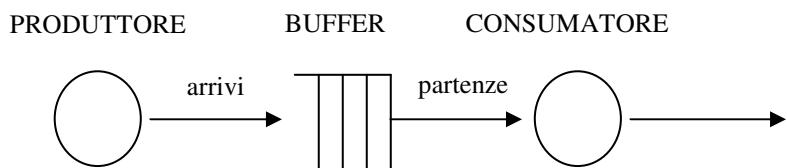
genera u ∈ U{0,1}
if u ≤ α1 then genera(y da 1 - e-λ1y)
else genera(y da 1 - e-λ2y)
  
```

Generazione Monte Carlo di realizzazioni iperesponenziali



Analisi Monte Carlo della coda di accesso ad una risorsa.

In questa sezione si vuole fare riferimento ad un modello gestionale classico composto da due risorse fisiche - un produttore ed un consumatore - separate da un buffer di accumulo degli “oggetti” in transito (clienti nel linguaggio dei modelli, semilavorati nella pratica, ad esempio). Le risorse sono anche dette serventi perché modellano macchine capaci di effettuare un servizio (lavorazione sui semilavorati). I serventi hanno una velocità finita e limitata e la durata di ogni servizio può essere intesa come la realizzazione di una variabile aleatoria afferente ad una distribuzione probabilistica nota. L’aleatorietà del servizio – presso entrambe le risorse - può essere attribuita tanto al servente quanto, più spesso, alla richiesta di servizio effettuata dal cliente. Quale che sia il caso, questa aleatorietà è alla base della formazione della coda d’attesa degli utenti in transito che, vengono accomodati nel buffer intermedio e lì attendono il proprio turno. In particolare la coda può essere limitata o illimitata e qui si adotterà la politica detta di “push”, ovvero il produttore produce/spinge oggetti nel buffer indipendentemente dal consumatore e dallo stato dello stesso buffer. In tal caso, quando i posti in coda risultano tutti occupati, si pone il problema di rigettare e perdere il cliente in arrivo dal produttore, oppure di bloccare il produttore stesso. Si tenga presente che la sola coppia “buffer-risorsa a valle” è anche nota come “stazione di servizio” nella Teoria delle Code classica. Di seguito è raffigurato il sistema produttore - buffer - consumatore:



Osservazione: la coda può crescere all’infinito, se l’intensità degli arrivi è “eccessiva” rispetto alla velocità del servente!

Nel sistema semplificato che vogliamo analizzare il buffer è praticamente inesauribile e gli utenti sono prodotti, “spinti” nel buffer e consumati uno alla volta, sempre secondo l’ordine d’arrivo (disciplina FIFO: First In - First Out). In particolare, vogliamo calcolare

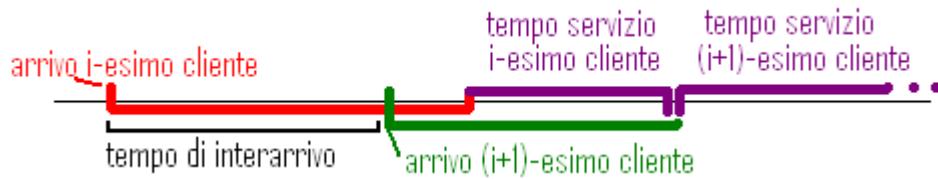
il tempo medio di attesa in coda dei clienti, assumendo che entrambi i serventi siano in una condizione iniziale di ozio e nessuno dei due possa essere soggetto a guasti né ad interruzioni di qualunque sorta, una volta avviato un servizio. Valuteremo anche l'andamento nel tempo del livello di occupazione del buffer e il livello medio su un orizzonte temporale fissato.

Definiamo le seguenti quantità:

W_i = tempo di attesa dell' i -esimo cliente

S_i = tempo di servizio dell' i -esimo cliente

A_i = intervallo di tempo tra l'arrivo dell' i -esimo cliente e quello dell' $(i+1)$ -esimo – in linguaggio “tecnico” *interarrivo* dell' $(i+1)$ -esimo cliente.



Il primo cliente che arriva attende un tempo nullo, dunque per lui: $W_1 = 0$. Allora, il tempo di attesa di ogni cliente che arriva è dato dalla seguente formula ricorsiva, nota come equazione di *Lindley*:

$$W_{i+1} = \max(0, W_i + S_i - A_i)$$

ovvero il tempo di attesa dell' $(i+1)$ -esimo cliente è pari alla somma di tempo d'attesa e tempo di servizio dell' i -esimo cliente meno il tempo di interarrivo dell' $(i+1)$ -esimo cliente.

La formula suddetta restituisce il valore zero quando l'istante di arrivo del cliente $(i+1)$ -esimo risulta successivo all'istante di fine servizio del cliente i -esimo.

Tramite il metodo Monte Carlo generiamo due serie di numeri casuali da cui, applicando le opportune trasformazioni inverse, ricaviamo i valori A_i e S_i .

Immaginiamo che tempi di servizio e di interarrivo seguano entrambi una legge esponenziale:

$$F_X(x) = 1 - e^{-\lambda x}$$

e

$$F_Y(x) = 1 - e^{-\lambda x}$$

dove X rappresenta la variabile aleatoria “tempo di interarrivo” con tasso λ e Y la variabile aleatoria “tempo di servizio” con tasso μ .

I due numeri casuali generati dal metodo Monte Carlo sono rispettivamente i valori di $F_X(x)$ e di $F_Y(x)$. Per ricavare per entrambe la x (ribadiamolo: corrisponde ora alla durata dell’interarrivo, ora alla durata del servizio), ci avvaliamo del metodo della trasformazione inversa, come già fatto nell’esempio precedente.

Posto il tempo di attesa del primo cliente pari a zero, potremo calcolare tutti gli altri tempi d’attesa con la formula di *Lindley*.

Prima di passare ai grafici ottenuti col metodo Montecarlo in Excel, è utile riportare un’osservazione. Accettando, in maniera intuitiva, che il numero medio di arrivi in un intervallo temporale di ampiezza “ $t - 0 \hat{=} t$ ” (ovvero t meno l’istante arbitrariamente fissato come iniziale) sia pari al prodotto $\lambda \cdot t$ (d’altra parte il tasso degli arrivi (λ) è costante!), introduciamo la variabile aleatoria N_S per indicare il numero medio di arrivi durante un servizio del consumatore, a sua volta rappresentato dalla variabile aleatoria S , distribuita ancora secondo una legge esponenziale, ma di parametro μ .

Allora, si può osservare che risulta:

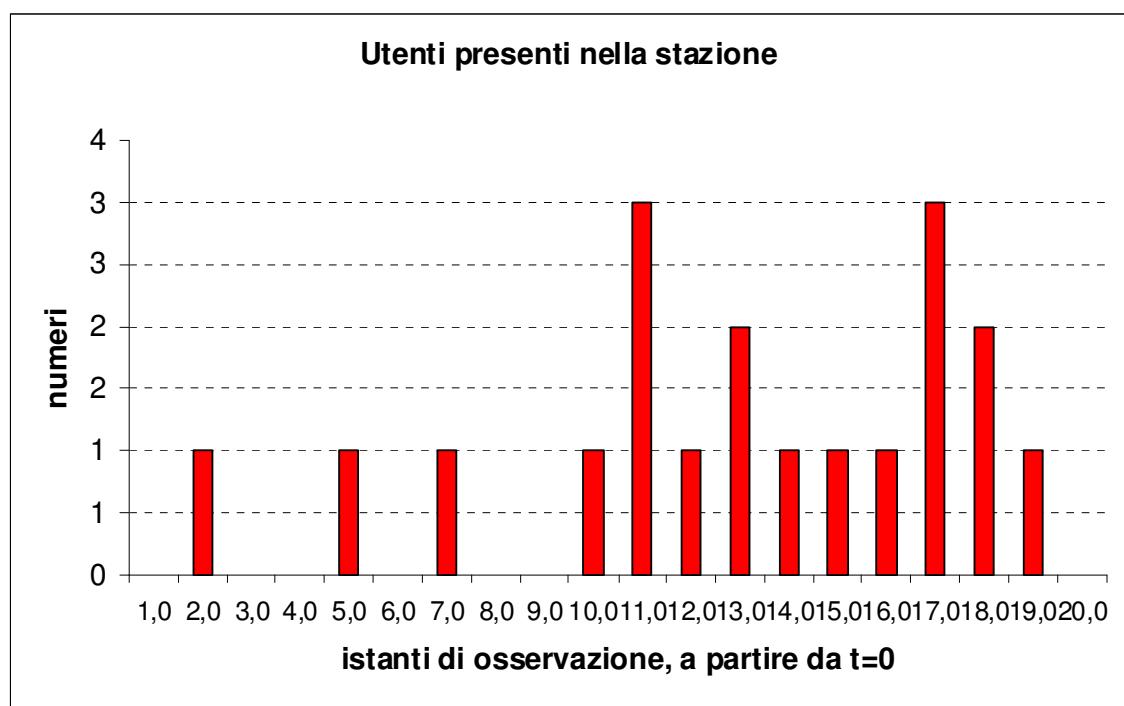
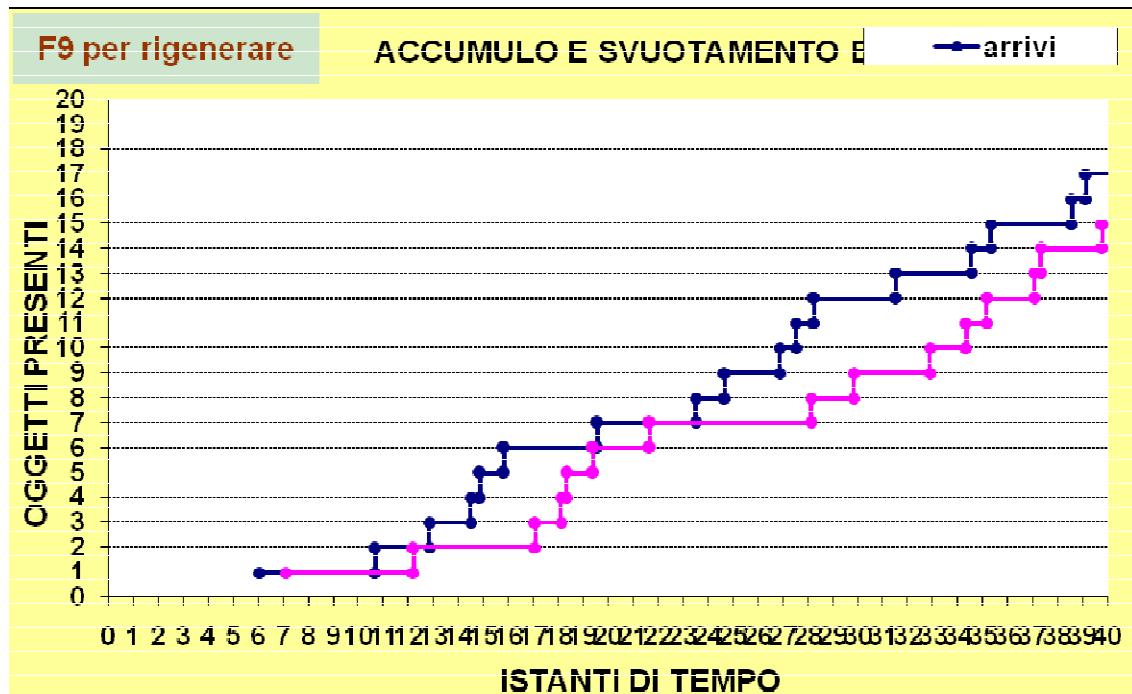
$$E[N_S] = \int_{t=0}^{\infty} E[N_S | S = t] \Pr\{S \in [t, t + dt]\} = \int_{t=0}^{\infty} \lambda t \exp\{-\mu t\} dt = \frac{\lambda}{\mu}$$

E ci si può chiedere: cosa succede quando si verifica la condizione $\lambda / \mu > 1$?

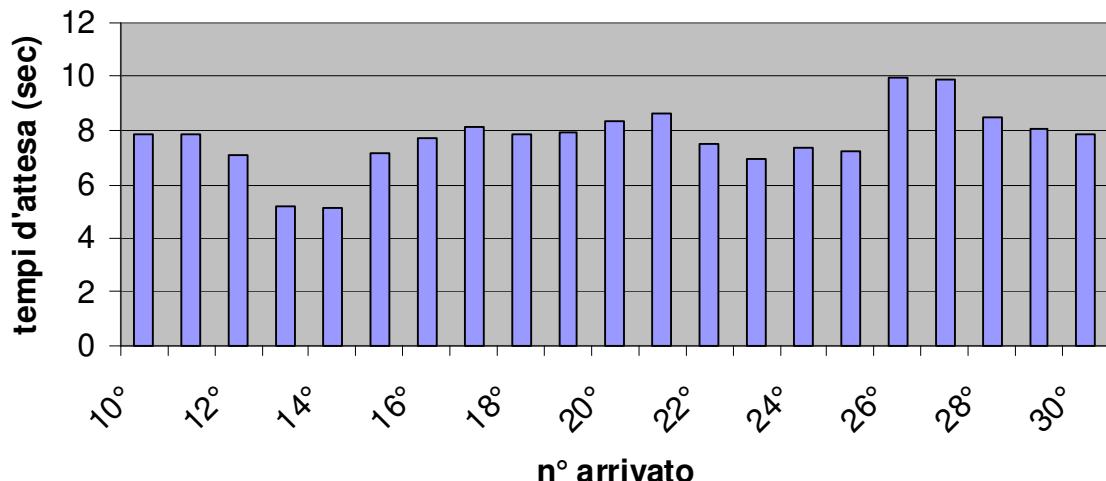
E quando, invece, risulta $\lambda / \mu < 1$?

E varrebbe la pena, nella pratica, di considerare il caso limite/separatore $\lambda / \mu = 1$?

In attesa di dare le risposte corrette, si può passare alla pagina successiva che riporta due grafici costruiti con Excel (vedere fogli Excel Produttore-Consumatore.xls).



Tempi di attesa in coda (dal 10° al 30° utente)



Tempi di attesa in coda

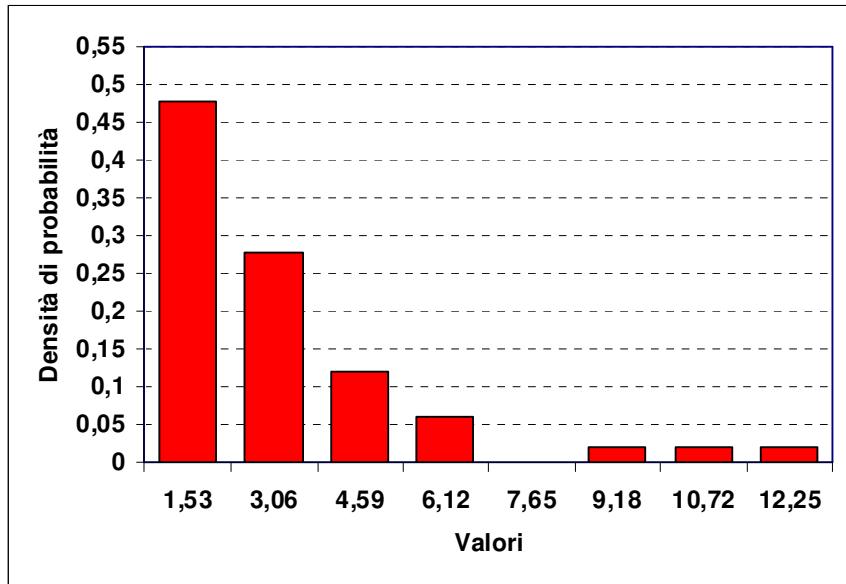
La stima dei parametri (di input)

La stima dei parametri è una fase dell’analisi statistica che gioca un ruolo basilare nella modellazione dei dati che saranno d’ingresso rispetto al modello da studiare. Ad esempio, ritornando al modello produttore-buffer-consumatore, è immediato riconoscere che, per formulare il problema della stima della funzione di distribuzione del tempo d’attesa degli oggetti nel buffer, occorre anzitutto definire o scegliere il modello di distribuzione che meglio si adatta a rappresentare il tempo di produzione (interarrivo) e il tempo di consumo (servizio).

Il primo passo della stima dei parametri è quello di costruire un istogramma – v. appendice – a partire da osservazioni reali, indipendenti, ed individuare in prima approssimazione una famiglia di funzioni (ad esempio la famiglia di Erlang o di Weibull) alla quale potrebbe appartenere quella in esame e per la quale si dovrà, allora, stimare il o i parametri che precisano la funzione (parametri “n” e “l” per la Erlang, parametri “a” e “l” per la Weibull).

In realtà, fase ancora precedente a questa è la verifica che quei dati siano effettivamente estratti da una stessa distribuzione (e.g. test di *Kruskal-Wallis*) e siano indipendenti (e.g. test di *von Neumann*). Non si dimentichi che l’indipendenza delle osservazioni è un’ipotesi fondamentale per l’applicazione di tanti risultati già studiati; in particolare, per la statistica detta “varianza campionaria” e poi per poter applicare il teorema limite centrale alla statistica “media campionaria”.

Prima, però, di qualunque confronto o test, grafico e non, è necessario stimare i parametri caratterizzanti di ognuna delle distribuzioni che si sono selezionate, poiché, nel passo precedente, se ne è semplicemente determinato il “tipo”. Ad esempio osservando l’istogramma seguente, costruito su dati sperimentali si sarebbe indotti a pensare di avere a che fare con una legge esponenziale.



Di essa però non conosciamo il tasso λ , senza il quale non possiamo effettuare, in prima approssimazione, alcun confronto con la reale densità esponenziale.

La stima dei parametri, come il tasso λ cui facevamo riferimento, è una procedura molto delicata e importante che affronteremo ora nel dettaglio.

Ultimi passi per la scelta della distribuzione statistica che meglio modella i nostri dati sono i test di bontà di adattamento (*Goodness of fit texts* – test della chi-quadrato e di *Kolmogorov-Smirnov*), di cui si parlerà più avanti e che si serviranno dei parametri ricavati in questa fase.

Due metodi saranno presi in considerazione:

- la *stima a massima verosimiglianza* (*o maximum likelihood estimation – “MLE”*);
- il *metodo dei momenti* – “MOM”.

Il primo consiste nel definire una funzione, detta “funzione di verosimiglianza”, e stimare i parametri trovandone il massimo (come si vedrà, infatti, la funzione risulterà incognita nei parametri).

Il secondo prevede, invece, di eguagliare i momenti di ordine k della distribuzione ipotizzata, con i rispettivi stimatori di momenti di ordine k dai dati, risolvendo il sistema incognito nei parametri.

Come si noterà, entrambi fanno ancora riferimento ai dati poiché si basano sulle statistiche da essi ricavate (media campionaria, varianza campionaria, ecc).

Il metodo dei momenti

Un generico momento di ordine k di una variabile aleatoria X , cioè il valore atteso della variabile elevata al momento k è così definito:

$$E[X^k] = \begin{cases} \sum_x x^k f(x) & \text{se } X \text{ è discreta} \\ \int_{\mathbb{R}} x^k f(x) dx & \text{se } X \text{ è continua} \end{cases}$$

Consideriamo delle variabili casuali i.i.d. X_1, \dots, X_n estratte da una stessa funzione densità $f(x)$.

Il metodo dei momenti (MOM) afferma che lo stimatore di ordine k è:

$$E[X^k] \rightarrow \sum_{i=1}^n \frac{X_i^k}{n}.$$

che è funzione dei parametri. Il MOM consiste nell'eguagliare a zero i momenti di ordine $1, 2, \dots, n$ con le corrispondenti stime di momenti fino ad ottenere tante equazioni quanti sono i parametri incogniti e poi risolverle contemporaneamente per ottenere le stime richieste.

A volte il MOM dà stimatori su cui è facile lavorare, altre volte può fornire stimatori distorti.

Di seguito alcuni esempi di stimatori MOM:

■ Per $= E[X_i]$ (ordine $k=1$) lo stimatore è $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$

■ Per $E[X_i^2]$ (ordine $k=2$) lo stimatore è $\sum_{i=1}^n \frac{X_i^2}{n}$

■ Per $VAR[X_i]$ lo stimatore è $\frac{n-1}{n} S^2$, poiché $VAR[X_i] = E[X_i^2] - (E[X_i])^2$ da cui:

$$\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n} = \frac{n-1}{n} S^2$$

NB: S^2 è distorto, ma è possibile usare anche solo S^2 .

Esempi (metodo MLE)

Legge esponenziale: solo un parametro da stimare. Dunque:

$$E[X] = \frac{1}{\lambda} = \bar{X} \quad \text{da cui} \rightarrow \hat{\lambda} = \frac{1}{\bar{X}}.$$

Legge di Poisson: supponiamo che X_1, \dots, X_n siano i.i.d da una distribuzione di Poisson di parametro λ . Poiché $\lambda = E[X_i]$, un stimatore di tipo MOM per λ è \bar{X} . Ma dato che $\lambda = Var(X_i)$ allora un altro stimatore MOM per λ è $\frac{n-1}{n}S^2$.

Legge normale: supponiamo di avere una distribuzione normale di media μ e varianza σ^2 . Poiché ci sono due parametri da stimare avremo un sistema in 2 equazioni:

$$\begin{cases} E(X) = \mu = \bar{X} \\ E(X^2) = \mu^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases}$$

Perciò gli stimatori MOM sono:

$$\hat{\mu} = \bar{X} \quad \text{e} \quad \hat{\sigma}^2 = \frac{n-1}{n} S^2.$$

Legge iperesponenziale: nella distribuzione iperesponenziale

$$F_y(y) = \sum_{i=1}^n \alpha_i (1 - e^{-\lambda_i y}) \quad \text{per} \quad y \geq 0$$

Ci sono tre parametri da stimare: α , λ_1 e λ_2 e valgono: $\alpha_1 = \alpha$ e $\alpha_2 = 1 - \alpha$. In questo caso ci servirebbero i momenti del 1°, 2° e 3° ordine. Ma abbiamo a disposizione solo la media e la varianza campionaria (ci rimane un livello di libertà!). Perciò si sceglie di impostare un'ulteriore condizione per rendere determinato il sistema:

$$\frac{1}{\lambda_1 + \lambda_2} = \frac{\frac{\alpha}{\lambda_1} + \frac{1-\alpha}{\lambda_2}}{2}$$

Consideriamo i seguenti risultati:

- | | |
|---|--|
| a. Momento del I ordine | $E[Y] = \sum_{i=1}^n \frac{\alpha_i}{\lambda_i} = \left(\frac{\alpha}{\lambda_1} + \frac{1-\alpha}{\lambda_2} \right)$ |
| b. Momento del II ordine | $E[Y^2] = \sum_{i=1}^n \frac{2\alpha_i}{\lambda_i^2} = 2 \left(\frac{\alpha}{\lambda_1^2} + \frac{1-\alpha}{\lambda_2^2} \right)$ |
| c. Varianza | $VAR[Y] = E[Y^2] - E^2[Y]$ |
| d. Media campionaria | \bar{X} |
| e. Varianza campionaria | $\frac{n-1}{n} S^2$ |
| f. Stimatore: | |
| i. 1° ordine: $M_1 \rightarrow \bar{X}$; | |
| ii. 2° ordine: $M_2 \rightarrow S^2 - E^2[Y]$. | |

Da notare che si è usata la sola S^2 e non il suo stimatore distorto. Il nostro sistema di tre equazioni in tre incognite diventa perciò:

$$\begin{cases} \frac{1}{\lambda_1 + \lambda_2} = \frac{\frac{\alpha}{\lambda_1} + \frac{1-\alpha}{\lambda_2}}{2} \\ \bar{X} = \frac{\alpha}{\lambda_1} + \frac{1-\alpha}{\lambda_2} \\ S^2 = 2 \left(\frac{\alpha}{\lambda_1^2} + \frac{1-\alpha}{\lambda_2^2} \right) - \left(\frac{\alpha}{\lambda_1} + \frac{1-\alpha}{\lambda_2} \right)^2 \end{cases}$$

Osservando che il secondo membro della prima equazione è uguale a quello della seconda:

$$\lambda_1 = \frac{2}{\bar{X}} - \lambda_2$$

Dalla seconda equazione ricaviamo:

$$\hat{\alpha} = \frac{\lambda_1(\lambda_2 \bar{X} - 1)}{\lambda_2 - \lambda_1}$$

Sostituendo λ_1 :

$$\alpha = 1 - \frac{\lambda_2 \bar{X}}{2}$$

Osserviamo che:

$$\frac{\alpha}{\lambda_1} = \frac{1-\alpha}{\lambda_2} = \frac{\bar{X}}{2} \quad \frac{1-\alpha}{\lambda_2^2} = \frac{\bar{X}}{2\lambda_2} \quad \frac{\alpha}{\lambda_1^2} = \frac{\bar{X}^2}{2(2-\lambda_2 \bar{X})}$$

La terza equazione diventa:

$$S^2 = 2 \left(\frac{\bar{X}^2}{2(2-\lambda_2 \bar{X})} + \frac{\bar{X}}{2\lambda_2} \right) - \left(\frac{\bar{X}}{2} + \frac{\bar{X}}{2} \right)^2$$

da cui:

$$\lambda_2(2-\lambda_2 \bar{X}) = \frac{2\bar{X}}{S^2 + \bar{X}^2}$$

Manipolando l'ultimo risultato giungiamo ad un'equazione di 2° grado in λ_2 :

$$\lambda_2^2 \bar{X} (S^2 + \bar{X}^2) - 2\lambda_2 (S^2 + \bar{X}^2) + 2\bar{X} = 0$$

che ammette come soluzioni:

Da cui:

$$\hat{\lambda}_1 = \frac{1}{\bar{X}} + \frac{1}{\bar{X}} \sqrt{\frac{(S^2 - \bar{X}^2)}{(S^2 + \bar{X}^2)}}$$

$$\hat{\lambda}_2 = \frac{1}{\bar{X}} - \frac{1}{\bar{X}} \sqrt{\frac{(S^2 - \bar{X}^2)}{(S^2 + \bar{X}^2)}}$$

(si suppone che $S^2 \geq \bar{X}^2$).

Per completezza riscriviamo il terzo coefficiente:

$$\hat{\alpha} = \frac{\lambda_1(\lambda_2 \bar{X} - 1)}{\lambda_2 - \lambda_1}$$

Esempio dal foglio Excel:

α	λ_1	λ_2
0,9	5	0,5

0,038035	1,182715	0,220085	0,005408	0,09198
0,409635	0,159324	0,128586	0,04661	0,283353
0,181822	0,10365	3,393875	0,407684	0,207476
0,137333	0,042202	0,170603	0,039083	3,563589
0,306768	0,138842	0,044659	0,147523	0,849025
0,293062	0,49593	0,189722	0,06778	0,115149
0,137179	0,130769	0,277224	0,046152	0,00011
0,672394	0,424989	0,152111	0,003551	0,183039

0,275342	0,043464	0,331043	0,101144	0,143709
0,07655	0,143325	4,654475	1,515698	0,396286

$$\hat{\lambda}_1 = 3,827974$$

$$\hat{\lambda}_2 = 0,48794$$

$$\hat{\alpha} = 0,886944$$

Legge beta: per una distribuzione beta di parametri a e b , la corrispondente densità è:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, 0 < x < 1.$$

Avremo che:

$$E[X] = \frac{a}{a+b} \quad \text{e} \quad Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Il metodo della massima verosimiglianza.

Consideriamo delle variabili casuali i.i.d. X_1, \dots, X_n con una funzione di densità $f(x)$.

Definiamo nel modo seguente la **funzione di verosimiglianza**:

$$L(\theta) \equiv f(X_1) \cdot f(X_2) \cdots f(X_n) = \prod_{i=1}^n f(X_i)$$

Poiché conosciamo le X_i , tale funzione ha come unica incognita il parametro (o i parametri) θ che si vuole valutare.

Lo **stimatore a massima verosimiglianza** (MLE) di θ è il valore $\hat{\theta}$ che massimizza la funzione appena definita, ponendo uguali a zero la sua derivata secondo θ . Se i parametri sono più di uno, bisogna risolvere un sistema di equazioni alle derivate parziali rispetto ad ogni parametro. Nella pratica si suole massimizzare, in luogo della $L(\theta)$, il logaritmo di essa, ovvero

$$l(\hat{\theta}) = \ln[L(\theta)]$$

che è nota come **funzione di verosimiglianza ridotta**. Il tutto funziona poiché il logaritmo di una funzione ne segue in qualche modo l'andamento e, data la natura delle distribuzioni più comuni (per esempio l'esponenziale), la funzione di verosimiglianza ridotta rende i calcoli più semplici.

Per essere sicuri che $\hat{\theta}$ sia un massimo, piuttosto che un minimo o un punto di flesso, la derivata di $l(\theta)$, valutata in $\hat{\theta}$ deve essere negativa.

Infine si tenga presente il seguente risultato: se $\hat{\theta}$ è un MLE per un parametro θ e $h(\cdot)$ è una funzione "one-to-one", allora $h(\hat{\theta})$ è il MLE di $h(\theta)$.

Esempi (metodo MLE)

Legge esponenziale: a partire dalla densità esponenziale

$$f_X(X) = \lambda e^{-\lambda X}$$

λ è il parametro che vogliamo stimare. La funzione di verosimiglianza ridotta è pari a:

$$\ell(\lambda) = \ln L(\lambda) = \ln \left(\prod_{i=1}^n \lambda e^{-\lambda X_i} \right) = \ln \left(\lambda^n e^{-\lambda \sum_{i=1}^n X_i} \right) = n \ln \lambda - \lambda \sum_{i=1}^n X_i$$

Le quantità X_i sono le realizzazioni della legge esponenziale, cioè i dati che abbiamo.

Consideriamo il logaritmo della funzione appena trovata:

$$\frac{d\ell}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n X_i = 0$$

da cui si ottiene il valore del parametro cercato:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}}$$

dove \bar{X} è esattamente la media campionaria.

Legge normale: X_1, \dots, X_n sono estratte da una normale con media μ e varianza σ^2 .

Questi ultimi sono i parametri da stimare. La funzione a massima verosimiglianza sarà:

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n f(x_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}} \end{aligned}$$

Calcoliamone il logaritmo:

$$\ln(L(\mu, \sigma^2)) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

E quindi la derivata in μ e in σ^2 :

$$\frac{\partial}{\partial \mu} (L(\mu, \sigma^2)) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) \equiv 0$$

da cui:

$$\hat{\mu} = \bar{X}$$

(con \bar{X} media campionaria) e:

$$\frac{\partial}{\partial \sigma^2} \ln(L(\sigma^2)) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^2 \equiv 0$$

$$\Rightarrow -n\sigma^2 + \sum_{i=1}^n (x_i - \bar{x})^2 = 0$$

da cui:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

che già sappiamo essere pari a $\frac{n-1}{n} S^2$.

Legge di Weibull: $\{t_i, i=1, \dots, n\}$ sono le osservazioni (tempi di vita degli n componenti) e r sono i fallimenti osservati (t_r = tempo al guasto).

Per semplicità chiamiamo $x_i = \min\{t_i, t_r\}$. Ecco distribuzione e densità di Weibull:

$$F_X(t) \triangleq 1 - \exp\{-\lambda t^\alpha\}, \quad \lambda > 0, \alpha > 0, \quad t \geq 0$$

$$f_X(t) \triangleq \lambda \alpha t^{\alpha-1} \exp\{-\lambda t^\alpha\}, \quad \lambda > 0, \alpha > 0, \quad t \geq 0$$

Dobbiamo stimare 2 parametri (λ e α). La funzione a massima verosimiglianza è:

$$\begin{aligned} L(\lambda, \alpha) &= \prod_{i=1}^r f(t_i | \lambda, \alpha) \cdot \prod_{i=r+1}^n R(t_r | \lambda, \alpha) \\ &= \prod_{i=1}^r \lambda \alpha x_i^{\alpha-1} e^{-\lambda x_i^\alpha} \cdot \prod_{i=r+1}^n e^{-\lambda x_i^\alpha} \\ &= \lambda^r \alpha^r \left(\prod_{i=1}^r x_i \right)^{\alpha-1} e^{-\lambda \sum_{i=1}^n x_i^\alpha} \end{aligned}$$

Il suo logaritmo:

$$\ln L(\lambda, \alpha) = r \ln \lambda + r \ln \alpha + (\alpha - 1) \sum_{i=1}^r \ln x_i - \lambda \sum_{i=1}^n x_i^\alpha$$

Deriviamo nei parametri ed eguagliamo a 0. Otteniamo:

$$\frac{r}{\lambda} - \sum_{i=1}^n x_i^\alpha = 0 \quad \text{e} \quad \frac{r}{\alpha} - \sum_{i=1}^r \ln x_i - \lambda \sum_{i=1}^n x_i^\alpha \ln x_i = 0$$

Non esistono soluzioni in forma chiusa per λ e α . Comunque si può scrivere λ in funzione di α :

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^n x_i^\alpha}$$

Sostituendo nella seconda equazione si ottiene:

$$\frac{r}{\alpha} + \sum_{i=1}^r \ln x_i - \frac{r \sum_{i=1}^n x_i^\alpha \ln x_i}{\sum_{i=1}^n x_i^\alpha} = 0$$

che si risolve iterativamente. Trovato il valore di α , si potrà usare per risolvere λ .

La distribuzione gamma

La distribuzione gamma è costruita a partire dalla funzione integrale gamma:

$$\Gamma(\alpha) \hat{=} \int_0^{\infty} x^{\alpha-1} e^{-x} dx , \quad \text{con } \alpha > 0$$

che è stata già incontrata nel calcolo di media e varianza della legge di Weibull.

Verificando le seguenti proprietà:

$$\Gamma(1) = 1, \quad \Gamma(1/2) = \sqrt{\pi}$$

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \quad \text{con } \alpha > 1, \quad \Rightarrow \Gamma(n) = (n - 1)!$$

$$\Gamma(\alpha)/\lambda^\alpha = \int_0^{\infty} x^{\alpha-1} e^{-\lambda x} dx \quad (*)$$

e ricordando la funzione densità del modello di Erlang di ordine “n”:

$$f_X(x) = \frac{(\lambda x)^{n-1}}{(n-1)!} \lambda e^{-\lambda x} \quad 0 \leq x < \infty$$

si decide di generalizzare il fattoriale $(n-1)! = \Gamma(n)$ con la $\Gamma(\alpha)$, ottenendo la seguente nuova funzione che rispetta i requisiti di una densità (grazie alla (*)):

$$f_X(x) = \frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \lambda e^{-\lambda x} \quad \alpha > 0, \lambda > 0, \quad 0 \leq x < \infty$$

Più comunemente (in EXCEL ad esempio) la densità gamma è riportata nella seguente forma:

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \quad \alpha > 0, \beta \hat{=} \frac{1}{\lambda} > 0, \quad 0 \leq x < \infty$$

La funzione integrale di quest'ultima è la distribuzione gamma di parametri α (parametro di forma) e β (parametro di posizione).

La funzione generatrice dei momenti è:

$$L_X(s) = (1 - \beta s)^{-\alpha}, \quad s < \beta^{-1}$$

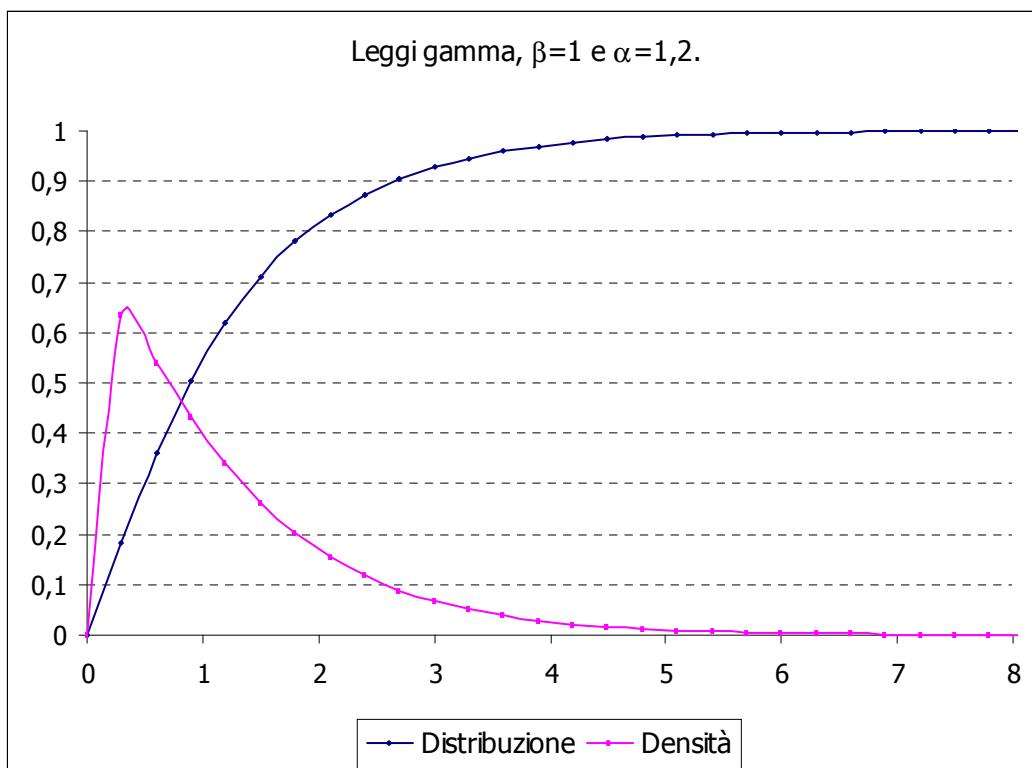
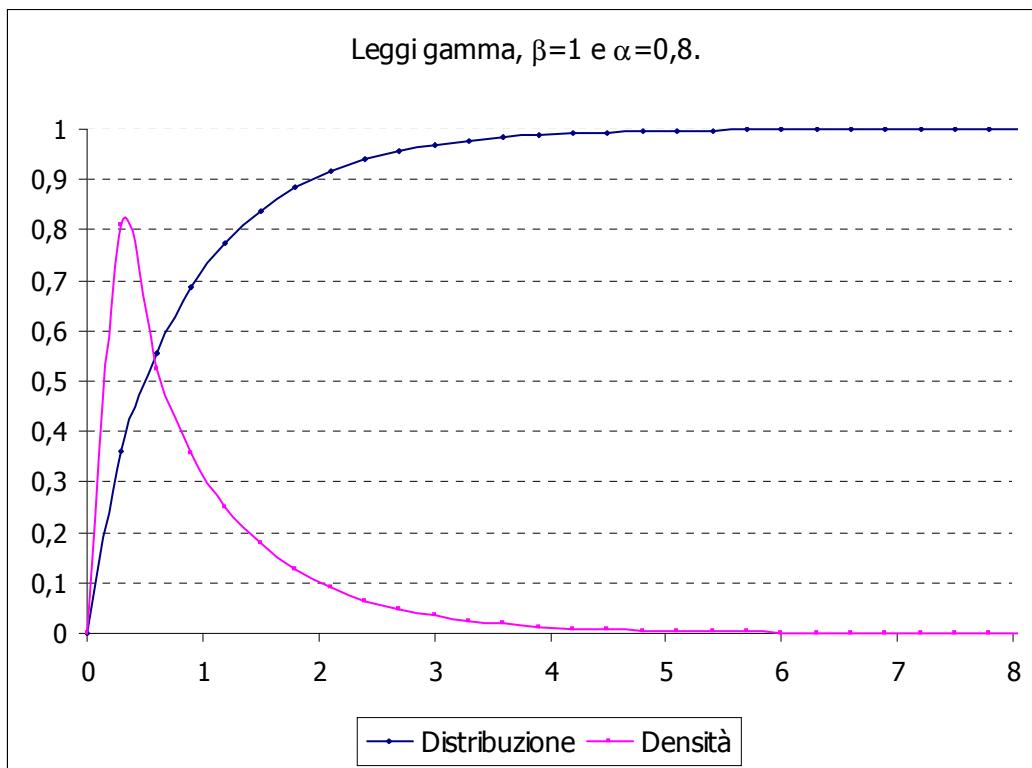
e con essa si possono ricavare media

$$E[X] = \alpha\beta$$

e varianza

$$VAR[X] = \alpha\beta^2,$$

Rappresentazione della Gamma



La distribuzione Student-t

Nei casi in cui non è possibile disporre di un campione di numerosità adeguata, ma si vuole, comunque, costruire un intervallo di confidenza (esatto) per la media si può ricorrere alla distribuzione Student-t. Essa permette di usare la varianza campionaria al posto della varianza vera (non conosciuta), però a patto che le osservazioni sperimentali a disposizione siano quelle di una legge normale.

A partire da una variabile aleatoria normale standard, Z , ed una seconda indipendente dalla prima, X_γ^2 , cioè una chi-quadrato con γ gradi di libertà, si definisce la seguente variabile aleatoria:

$$T_\gamma \hat{=} \frac{Z}{\sqrt{X_\gamma^2 / \gamma}}$$

alla quale è associata (si dimostra) la seguente funzione densità:

$$f_T(t) = \frac{\Gamma((\gamma+1)/2)}{\Gamma(\gamma/2)\sqrt{\pi\gamma}} \left(1 + \frac{t^2}{\gamma}\right)^{-(\gamma+1)/2}, \quad -\infty < t < +\infty$$

detta, appunto, densità della Student-t con γ gradi di libertà.

La funzione generatrice dei momenti non esiste, però media e varianza si calcolano facilmente, risultando:

$$E[T] = 0 \quad \text{e} \quad \text{VAR}[T] = \gamma/(\gamma-2), \quad \{\rightarrow 1 \text{ per } \gamma \rightarrow \infty\}$$

L'utilità della T_γ nella costruzione degli intervalli di confidenza per la media campionaria poggia sui seguenti risultati:

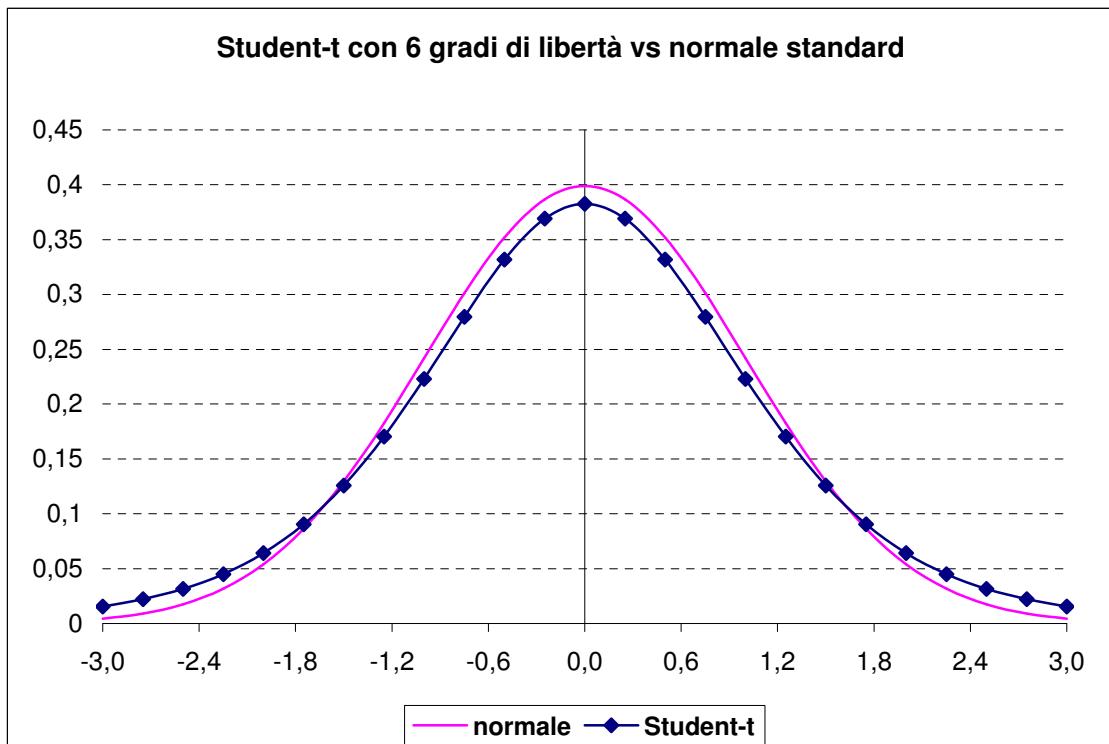
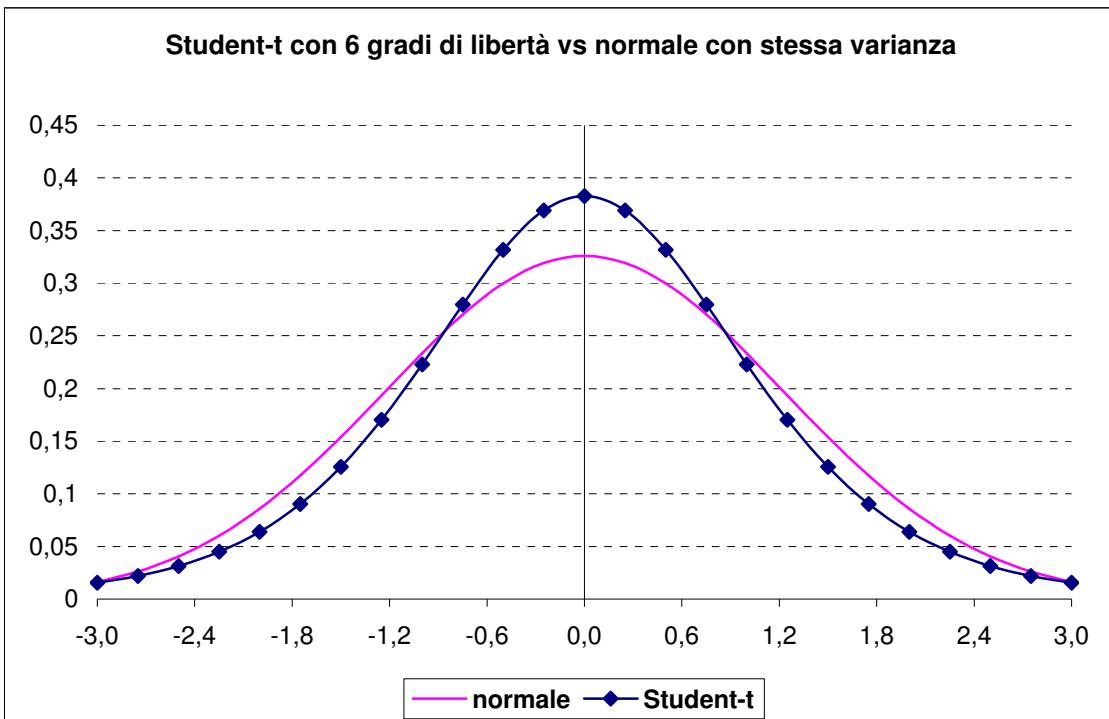
$$(1) \quad \frac{\bar{X} - S/\sqrt{n}}{\sigma/\sqrt{n}} = Z \quad \text{e} \quad (2) \quad \frac{(n-1)S^2}{\sigma^2} = X_{\gamma=n-1}^2$$

Infatti:

$$\frac{\bar{X} - S/\sqrt{n}}{\sigma/\sqrt{n}} = T_{\gamma=n-1} \quad \text{e perciò l'intervallo:} \quad \left[\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2} \right]$$

è un intervallo di confidenza per la media di un campione relativamente piccolo, $n < 30$, di realizzazioni indipendenti estratte da una legge normale di varianza non nota.

Confronti fra la legge Student-T e la normale



Esercizi di Riepilogo

Esercizio 1

I dati indicati in tabella si riferiscono ai tempi di riparazione (espressi in ore) di una macchina sigillatrice presente nei magazzini di un'azienda che assembla prodotti appartenenti ad una filiera agro-alimentare. Il campione è supposto proveniente da una popolazione normale di media incognita (μ) e varianza 4 (deviazione standard $\sigma = 2$).

Tempi di riparazione macchina				
5,03	5,35	4,95	5,17	4,86
4,96	5,04	5,1	5,38	5,35
5,01	4,95	4,75	5,03	4,9
5,42	5,26	4,99	5,27	5,13

Ai fini della valutazione dell'impatto di un eventuale “fermo produzione”, il responsabile dell'azienda vi comunica di voler conoscere nel 95% dei casi la durata del suddetto tempo di riparazione (μ).

Svolgimento

Essendo i tempi di riparazione della macchina sigillatrice provenienti da una popolazione $\sim N(\mu, \sigma^2)$ dove σ^2 è nota, applichiamo direttamente la formula

$$P\left(\bar{X} - z_{\alpha/2} \cdot \sigma / \sqrt{n} < \mu < \bar{X} + z_{\alpha/2} \cdot \sigma / \sqrt{n}\right) = 1 - \alpha$$

dove

$$\bar{X} = 5,095, \sigma = 2, n = 20, \alpha = 0,05 \text{ e } z_{\alpha/2} = 1,96.$$

In numeri

$$P(5,095 - 1,96 \cdot 2 / \sqrt{20} < \mu < 5,095 + 1,96 \cdot 2 / \sqrt{20}) = 1 - 0,05$$

ovvero

$$P(5,095 - 0,877 < \mu < 5,095 + 0,877) = 0,95$$

ed infine

$$P(4,218 < \mu < 5,972) = 0,95$$

cioè in 95 casi su 100 il tempo medio di riparazione della macchina sigillatrice e, quindi, il tempo medio di fermo produzione sarà compreso nell'intervallo $[4,218 - 5,972]$.

Esercizio 2

Riprendendo la situazione descritta nell'esercizio precedente, si supponga di dover rispondere alla medesima richiesta del responsabile di azienda, ma senza conoscere la varianza della popolazione normale da cui sono estratti i tempi di riparazione della macchina sigillatrice riportati in tabella.

Svolgimento

Ferma restando l'ipotesi che il campione provenga da una popolazione normale $N(\mu, \sigma^2)$, nel caso in cui non si conosce σ^2 , per costruire l'intervallo di confidenza occorre basarsi sulla statistica:

$$T = \frac{\bar{X} - \mu}{\sqrt{S/n}}$$

dove T ha distribuzione t di Student con $n-1$ gradi di libertà, anche questi da leggere da tabella.

Quindi, in analogia a quanto visto per il caso con varianza nota:

$$P\left(\bar{X} - t_{\alpha/2} \cdot S/\sqrt{n} < \mu < \bar{X} + t_{\alpha/2} \cdot S/\sqrt{n}\right) = 1 - \alpha$$

con

$$-t_{\alpha/2}, t_{\alpha/2}$$

i centili della distribuzione t di Student associati alle probabilità $\alpha/2, 1 - \alpha/2$.

Applicando direttamente la formula

$$P\left(\bar{X} - t_{\alpha/2} \cdot S/\sqrt{n} < \mu < \bar{X} + t_{\alpha/2} \cdot S/\sqrt{n}\right) = 1 - \alpha$$

dove

$$\bar{X} = 5,095, S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = 0,036 \rightarrow S \cong 0,19, n = 20, \alpha = 0,05 \text{ e } t_{\alpha/2} = 2,09.$$

In numeri

$$P(5,095 - 2,09 \cdot 0,19 / \sqrt{20} < \bar{X} < 5,095 + 2,09 \cdot 0,19 / \sqrt{20}) = 1 - 0,05$$

ovvero

$$P(5,095 - 0,089 < \bar{X} < 5,095 + 0,089) = 0,95$$

ed infine

$$P(5,006 < \bar{X} < 5,184) = 0,95$$

cioè in 95 casi su 100 il tempo medio di riparazione della macchina sigillatrice e, quindi, il tempo medio di fermo produzione sarà compreso nell'intervallo $[5,006 - 5,184]$.

Esercizio 3

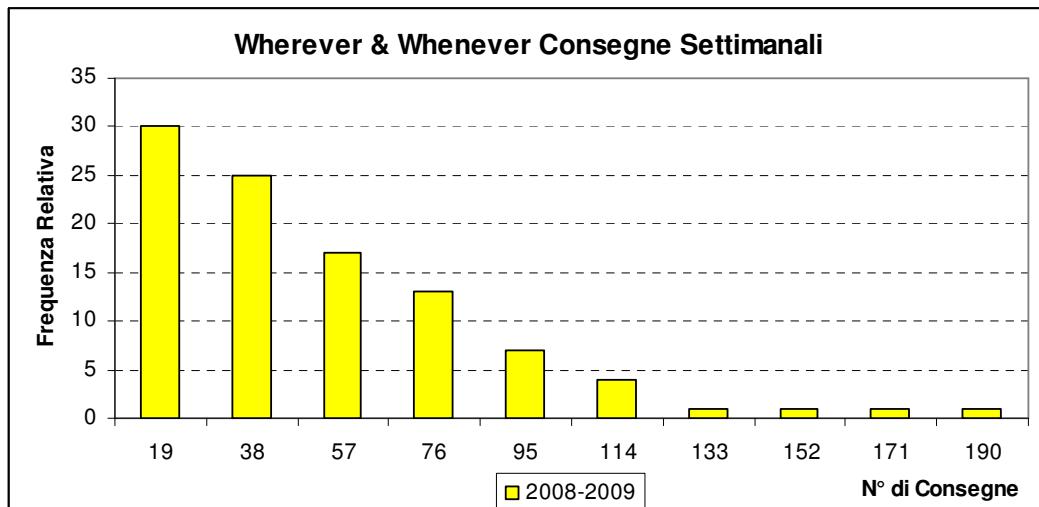
In una settimana di lavoro la flotta di mezzi del corriere *Wherever & Whenever* consegna una quantità variabile di colli. Negli ultimi 2 anni il volume di consegne settimanali è in tabella ottenuti a partire dalle registrazioni del sw di magazzino.

N° di consegne settimanali 2008-2009										
60	75	80	30	20	20	40	20	0	70	
30	10	10	75	30	50	30	30	0	190	
90	40	20	75	10	10	50	10	100	10	
90	20	100	140	20	40	0	50	10	10	
20	25	30	40	60	20	40	10	60	10	
40	160	30	72	30	30	10	10	10	40	
10	30	80	40	10	50	0	90	40	0	
40	30	20	50	10	10	50	70	25	10	
60	10	89	100	110	18	0	30	10	20	
70	10	60	60	30	130	40	10	90	10	

Oltre alla media (), dall'area commerciale vi chiedono altre statistiche sul numero medio di colli consegnati settimanalmente. A tal fine, decidete di presentare un intervallo di confidenza al 95% per la suddetta media a partire dai dati elementari a Vs disposizione, accompagnando, inoltre, i Vs calcoli dalle opportune ipotesi.

Svolgimento

Non è nota la distribuzione della popolazione da cui proviene il campione di dati in tabella, ma da una prima rappresentazione grafica questi non sembrano distribuirsi in accordo ad una legge normale.



Ciò sembrerebbe impedire la generazione di un intervallo di confidenza per la variabile aleatoria \bar{Y} “numero medio di colli consegnati settimanalmente”, ma, grazie all’esistenza del Teorema Limite Centrale il problema è ben presto superato.

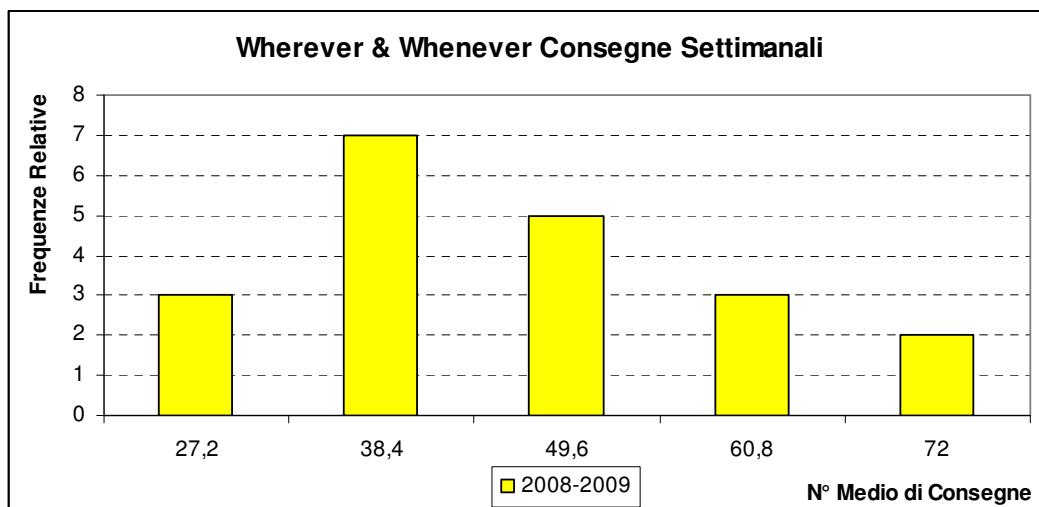
Ciò è possibile a partire dalla riorganizzazione dei 100 dati, per esempio, in 20 gruppi o classi da 5 dati ciascuno

N° medio di consegne settimanali 2008-2009				
58	48	28	32	34
44	55,8	38	20	35
34	72	28	24	58
48	64,4	47,6	42	16

a cui segue la definizione degli altri estremi necessari per generare un istogramma, ovvero il numero di classi (i.e. 59), il minimo (i.e. 16), il massimo (i.e. 72) e l’ampiezza di ciascuna classe dato da (massimo-minimo)/n° classi (i.e. 11,2). Indicando il limite superiore (LS) per ciascuna classe, si possono infine calcolare le frequenze assolute (FA) e le frequenze relative (FR).

Classe	LS	FA	FR
1	27,2	3	3
2	38,4	10	7
3	49,6	15	5
4	60,8	18	3
5	72	20	2

Dalla rappresentazione delle frequenze relative appare ora più vicina la soddisfazione della normalità della distribuzione dei (gruppi di) dati.



Considerando la variabile aleatoria $\bar{Y}_i \quad i = 1, \dots, 20$ ottenuta raggruppando 5 dati relativi al numero di colli consegnati settimanalmente, ci siamo ricondotti ad un caso che sappiamo trattare in cui la varianza non è nota. Quindi, per generare l'intervallo di confidenza al 95% applichiamo la formula

$$P\left(\bar{Y} - t_{\alpha/2} \cdot S/\sqrt{n} < \bar{Y} < \bar{Y} + t_{\alpha/2} \cdot S/\sqrt{n}\right) = 1 - \alpha$$

dove

$$\bar{Y} = \frac{\sum_{i=1}^{20} \bar{Y}_i}{20} = 41,34, \quad S^2 = \frac{\sum_{i=1}^{20} (\bar{Y}_i - \bar{Y})^2}{19} = 230,1 \rightarrow S \cong 15,17, \quad n = 20, \quad \alpha = 0,05 \quad e \quad t_{\alpha/2} = 2,09.$$

In numeri

$$P(41,34 - 2,09 \cdot 15,17 / \sqrt{20} < \bar{Y} < 41,34 + 2,09 \cdot 15,17 / \sqrt{20}) = 1 - 0,05$$

ovvero

$$P(41,34 - 7,09 < \text{ } < 41,34 + 7,09) = 0,95$$

ed infine

$$P(34,25 < \text{ } < 48,43) = 0,95$$

cioè in 95 casi su 100 il numero medio di colli consegnati settimanalmente sarà compreso nell'intervallo $[34,25 - 48,43]$.

Test d'ipotesi: concetti di base e test sulla media

Facendo sempre riferimento al modello produttore-consumatore, non dovrebbe essere difficile immaginare che fra i vari obiettivi dell'analisi statistica del sistema ci possa essere almeno uno dei due seguenti:

1. “stabilire se il tempo medio di soggiorno dei clienti nel buffer d'attesa è superiore, inferiore o approssimativamente uguale ad un valore fissato”
2. “stabilire se la forma della distribuzione della variabile aleatoria che rappresenta il tempo di attesa dei clienti nel buffer è assimilabile ad una forma fissata”

Ancora, si può immaginare di avere a disposizione due sistemi che funzionano con due diverse politiche di gestione del buffer (inserimento di nuovi prodotti da parte del produttore ed estrazione degli stessi da parte del consumatore) e si potrebbe voler stabilire qual'è il sistema dove si registrano tempi medi di permanenza nel buffer minori. In tal caso, si dovrebbe ragionare sulla differenza di due medie.

In tutti i casi appena enunciati, la Statistica suggerisce di ragionare formulando un'opportuna ipotesi e poi di tentando di confutarla con un metodo adeguato. Il metodo si chiama Test d'ipotesi. L'ipotesi formulata è detta ipotesi *nulla* ed è quella che ha un fondamento di convinzione scientifica: per questo è anche detta ipotesi del ricercatore o, più modestamente, ipotesi di lavoro. Ad esempio, nel modello produttore-consumatore con interratrivi esponenziali di parametro $\lambda = 0.8 \text{ u.t.}^{-1}$, tempi di consumo esponenziali di parametro $\mu = 1 \text{ u.t.}^{-1}$ e gestione FIFO del buffer l'ipotesi del ricercatore è che la media dei tempi d'attesa, w_0 , calcolata solo su quelli che aspettano, risulti: $w_0 = 1 / (1 - \rho) = 5 \text{ u.t.}$, dove $\rho \doteq \lambda / \mu = 0.8$. Allora, se si pensa che il sistema reale sotto osservazione si comporti effettivamente secondo quanto previsto dal modello con arrivi e consumi esponenziali, si proverà a confutare l'ipotesi (nulla) che la media dei tempi d'attesa nel sistema reale (w_{re}) sia, appunto, pari a 0.8.

Il punto è quello di stabilire come si possa confutare quell'ipotesi, ovvero di progettare un test.

Prima di entrare nel dettaglio della progettazione del test, si osservi che l'ipotesi nulla rimarrà in campo fino a quando non si riuscirà a rigettarla e, viceversa, se con un certo test (progettato) si riuscirà a rigettare l'ipotesi nulla, allora si potrà concludere che

risulta accettata la cosiddetta ipotesi *alternativa* che, di fatto, è la semplice negazione della prima. Nell'esempio del produttore-consumatore, l'ipotesi alternativa consiste nella dichiarazione seguente: la media dei tempi d'attesa nel sistema reale è diversa da 0.8.

Sul piano formale le ipotesi nulla e alternativa, con le notazioni del modello produttore-consumatore, si formulano, rispettivamente, alla seguente maniera:

$$H_0: w_{re} = w_0$$

$$H_1: w_{re} \neq w_0$$

Passando, finalmente, alla progettazione di uno specifico test per la media dei tempi d'attesa, si osserva, in via preliminare, che la media del sistema reale non può che essere stimata attraverso la media di un campione di osservazioni (indipendenti e di numerosità sufficientemente grande). Infatti, partendo da qui, si può ricorrere al fatto (sperimentale) che la media $\bar{W}(k)$, costruita su un campione di k tempi d'attesa rilevati, tende ad essere distribuita come una normale, di media pari a w_{re} , per $k \rightarrow \infty$. A partire da questo, si conviene di “recuperare” dal sistema reale tanti (n) campioni di dimensione k , in modo da avere tante medie campionarie $\bar{W}_1(k)$, $\bar{W}_2(k)$, ..., $\bar{W}_n(k)$. La cosiddetta grande media, cioè la media delle medie campionarie (è anch'essa una campionaria, ma costruita con n osservazioni “tendenti alla normale”!), indicata con $\bar{\bar{W}}(n, k)$ continua ad essere corretta per stimare la w_{re} , ma, in più, per essa vale anche il seguente risultato:

$$\frac{\bar{\bar{W}}(n, k) - w_0}{S / \sqrt{n}} \approx T_{n-1}, \quad k \rightarrow \infty$$

dove S indica la solita radice quadrata della varianza campionaria, riferita alla grande media e T_{n-1} indica la statistica distribuita secondo la legge di Student con $n-1$ gradi di libertà.

Attorno al risultato appena stabilito può essere progettato un test perché si osserva che, se l'ipotesi nulla è vera, allora la probabilità che $\bar{\bar{W}}(n, k)$ risulti abbastanza vicina a w_0 corrisponde alla probabilità che il valore della statistica T_{n-1} sia compreso in un certo intervallo di valori. Formalmente:

$$\Pr(-t_{n-1;1-\alpha/2} \leq \frac{\bar{W}(n,k) - w_0}{S/\sqrt{n}} \leq t_{n-1;1-\alpha/2}) = 1 - \alpha$$

dove i valori $-t_{n-1;1-\alpha/2}$ e $t_{n-1;1-\alpha/2}$ sono i quantili della legge di Student che risultano dopo aver fissato la probabilità “ $1 - \alpha$ ” ad un valore adeguato, quale potrebbe essere il valore 0.9, oppure 0.95 o, al massimo, 0.99. Nel linguaggio della Teoria dei Test, si dice che “ α ” è il *livello di significatività* del test e l’intervallo $[-t_{n-1;1-\alpha/2}; t_{n-1;1-\alpha/2}]$ è detto *regione di accettazione*, nel senso che, se risulta:

$$\frac{\bar{W}(n,k) - w_0}{S/\sqrt{n}} \in [-t_{n-1;1-\alpha/2}; t_{n-1;1-\alpha/2}]$$

cioè se il valore della statistica appartiene alla regione di accettazione, allora rimane stabilito l’esito (positivo per il ricercatore!) del test nell’evidenza sperimentale che *non consente di rigettare l’ipotesi nulla*. Accanto alla regione di accettazione, rimane pure definita la cosiddetta *regione di rifiuto* che, nel caso in questione, è rappresentata dall’unione dei due intervalli:

$$(-\infty, -t_{n-1;1-\alpha/2}] \text{ e } [t_{n-1;1-\alpha/2}, +\infty).$$

Se il valore della statistica appartiene alla regione di rifiuto, allora rimane stabilito l’esito (negativo per il ricercatore!) del test nell’evidenza sperimentale che *consente di rigettare l’ipotesi nulla*.

Si tenga presente che, nella pratica, l’esito di non rigettare l’ipotesi nulla equivale ad accettarla e ritenerla vera fino a futura prova contraria, cioè fino a che un eventuale altro test stabilisca il contrario. E il contrario consisterebbe nel fatto che il valore di un’altra statistica, diversa da quella di Student ma ugualmente “valida”, vada a cadere nella cosiddetta regione di rifiuto (o regione critica). Per capire in che senso la statistica di Student è stata valida per progettare il test appena definito e quindi per avere le linee guida per cercarne una seconda (se esiste!) occorre riflettere sul fatto che grazie al risultato

$$\frac{\bar{W}(n,k) - w_0}{S/\sqrt{n}} \approx T_{n-1}, \quad k \rightarrow \infty$$

siamo stati in grado di incorporare la differenza di nostro interesse, $\bar{W}(n,k) - w_0$, in una variabile aleatoria avente una distribuzione nota che ci ha permesso di tradurre la piccolezza di quella differenza nell’appartenenza della realizzazione (costruita con i

campioni di osservazioni reali) della variabile aleatoria ad un determinato intervallo (regione di accettazione). Il ragionamento è stato dunque un ragionamento da prova di necessità: se l'ipotesi nulla è vera allora i valori delle possibili realizzazioni (e ogni volta che si esegue il test se ne costruisce una e una sola) della statistica individuata devono appartenere alla regione di accettazione e, viceversa, se la realizzazione costruita in un'esecuzione del test risulta appartenere non alla regione di accettazione ma a quella di rifiuto, allora si conclude che l'ipotesi nulla debba essere rigettata. L'esplicitazione di questo ragionamento ne illustra immediatamente i limiti ed è facile riconoscere che questi limiti si traducono in due possibilità di errore proprie del test. Le possibilità di errore risiedono nella seguente doppia eventualità: la prima è che un'ipotesi (nulla) vera possa apparire falsa solo per un caso e ciò conduce ad un errato rigetto (che mortifica il ricercatore che, invece, aveva visto giusto!); la seconda eventualità è che un'ipotesi (nulla) falsa possa apparire vera solo per un caso e ciò conduce ad una errata accettazione (o, meglio, al mancato rigetto).

In sintesi, si può riconoscere che esistono due tipi di errori:

- considerare falsa una ipotesi nulla che, in realtà, è vera (*errore di prima specie*);
- considerare vera una ipotesi nulla che, in realtà, è falsa (*errore di seconda specie*).

Il primo tipo di errore è considerato peggiore del secondo.

In ogni caso:

$$\Pr(\text{errore di I specie}) = \Pr(\text{rigettare } H_0 \mid H_0 \text{ è vera}) = \alpha$$

$$\Pr(\text{errore di II specie}) = \Pr(\text{accettare } H_0 \mid H_0 \text{ è falsa}) = \beta.$$

È interessante osservare che l'occorrenza di un errore di 1^a specie dipende anche dalla scelta, soggettiva, di fissare α ad un valore relativamente alto. Supponendo di avere calcolato il valore della statistica di test corrispondente alle osservazioni reali, si può pensare di determinare il meno alto livello di significatività, α , che, se adottato, porterebbe a perdere l'accettazione. Nella letteratura in lingua inglese esso è detto “*p-value*”. Traducendolo come “*valore p*” di un test, esso può essere definito come il più piccolo livello di significatività, a partire dal quale verrebbe rigettata l'ipotesi nulla.

In questo senso, riprendendo l'intervallo di rigetto corrispondente ad un generico α :

$$\left| \frac{\overline{\overline{W}}(n, k) - w_0}{S / \sqrt{n}} \right| > t_{n-1; 1-\alpha/2},$$

ponendo per comodità

$$\frac{\overline{\overline{W}}(n, k) - w_0}{S / \sqrt{n}} \doteq t_{n-1; 1-p/2},$$

e ragionando sui corrispondenti valori della distribuzione di Student, $F_T(\cdot)$:

$$F_T(|t_{n-1; 1-p/2}|) > F_T(t_{n-1; 1-\alpha/2}) = 1-\alpha/2$$

si ha

$$F_T(|t_{n-1; 1-p/2}|) > 1-\alpha/2 \text{ e da qui: } \alpha > 2 \cdot (1 - F_T(|t_{n-1; 1-p/2}|)) \doteq p.$$

Ancora a proposito della possibilità di accettare/rigettare un'ipotesi nulla, si può fare un'osservazione sulla problematicità della dimensione (n) del campione: poiché la differenza di nostro interesse, $\overline{\overline{W}}(n, k) - w_0$, viene moltiplicata per \sqrt{n} , nel calcolo della statistica, allora per $n \rightarrow \infty$ si potrebbe finire per rigettare sempre l'ipotesi nulla.

Infine, è importante la seguente definizione, con la quale si vuole misurare la potenza di un test.

DEF: Il **potere** di un test d'ipotesi è la probabilità di rifiutare una ipotesi falsa, ossia:

$$1-\beta$$

Un caso particolare: normalità delle osservazioni e varianza nota.

La possibilità di avere a che fare con un campione di osservazioni reali estratte da una legge normale e, per di più, con varianza nota è considerata in questa sede come un caso particolare. Per completezza, dunque, sarà delineato il relativo test sulla media, cogliendo l'occasione di aggiungere qualche ulteriore dettaglio sulla teoria dei test.

Anzitutto è il caso di precisare che, quando ci si riferisce alla media di uno qualunque dei parametri o degli indici di prestazione di un modello d'interesse, diverso da quello produttore-consumatore, allora si è soliti scrivere così:

$$H_0: \mu = \mu_0 \quad (\text{ipotesi nulla})$$

$$H_1: \mu \neq \mu_0 \quad (\text{ipotesi alternativa})$$

Tale test è detto bilaterale e così lo si distingue dagli altri due possibili, che sono il test monolaterale_1 e il test monolaterale_2, rispettivamente formulati come segue:

$$H_0: \leq 0 \quad H_0: \geq 0$$

$$H_1: > 0 \quad H_1: < 0$$

Con le ipotesi fatte, la statistica del test sulla media diventa la normale standard:

$$\frac{\bar{X} - 0}{\sigma / \sqrt{n}} \approx Z,$$

con \bar{X} media campionaria e σ^2 quale varianza nota.

Da qui:

$$\Pr(-z_{1-\alpha/2} \leq \frac{\bar{X} - 0}{\sigma / \sqrt{n}} \leq z_{1-\alpha/2}) = 1 - \alpha$$

e perciò

$$\left| \frac{\bar{X} - 0}{\sigma / \sqrt{n}} \right| > z_{1-\alpha/2}$$

è la regione di rifiuto (o regione critica) e

$$\left| \frac{\bar{X} - 0}{\sigma / \sqrt{n}} \right| \leq z_{1-\alpha/2}$$

quella di accettazione.

Nel caso di test monolaterale_1, è immediato riconoscere che la regione d'accettazione diventa $(-\infty, z_{1-\alpha}]$ e quindi si rifiuta l'ipotesi se risulta:

$$\frac{\bar{X} - 0}{\sigma / \sqrt{n}} > z_{1-\alpha}.$$

Viceversa, per il test monolaterale_2, l'intervallo di accettazione è $[-z_{1-\alpha}, \infty)$ e l'ipotesi viene rifiutata se risulta:

$$\frac{\bar{X} - 0}{\sigma / \sqrt{n}} < -z_{1-\alpha}.$$

Significatività fissata e potenza garantita

In questo paragrafo verrà determinata la dimensione (n) del campione che assicura un voluto potere del test ($1 - \beta$), una volta fissato il livello di significatività (α). La determinazione ha come riferimento un test bilaterale, nel caso di osservazioni normali e di varianza nota. Assunzione cruciale è che l'utente sia in grado di fornire un valore (μ_1) per la media reale (μ), supposta maggiore di quella (μ_0) ipotizzata con l'ipotesi nulla.

Formalizzando l'assunzione: $\mu_1 = \mu_0 + \delta \geq \delta > 0$ si dimostrerà che la dimensione del campione approssimativamente coerente con la coppia di valori, α fissato, e β , voluto, è la seguente: $n \approx \sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2 / \delta^2$.

La dimostrazione comincia ricavando il valore di β :

$$\begin{aligned}\beta &= \Pr(\text{errore di II specie} \mid \mu = \mu_1 > \mu_0) \\ &= \Pr(\text{fallire nel rigettare } H_0 \mid H_0 \text{ è falsa} (\mu = \mu_1 > \mu_0)) \\ &= \Pr(|Z_0| \leq z_{1-\alpha/2} \mid \mu = \mu_1) = \Pr(-z_{1-\alpha/2} \leq Z_0 \leq z_{1-\alpha/2} \mid \mu = \mu_1) \\ &= \Pr\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \mid \mu = \mu_1\right) \\ &= \Pr\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \mid \mu = \mu_1\right)\end{aligned}$$

Dato che $Z \equiv \frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}}$ è estratta da una normale standard, risulta che:

$$\begin{aligned}\beta &= \Pr\left(-z_{1-\alpha/2} \leq Z + \frac{\sqrt{n}\delta}{\sigma} \leq z_{1-\alpha/2}\right) \\ &= \Pr\left(-z_{1-\alpha/2} - \frac{\sqrt{n}\delta}{\sigma} \leq Z \leq z_{1-\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\right) \\ &= \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\right) - \Phi\left(-z_{1-\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\right)\end{aligned}$$

Adesso si analizza il secondo termine dell'ultima espressione, riconoscendo che $-(z_{1-\alpha/2} + \sqrt{n}\delta/n)$ corrisponde ad un'ascissa negativa abbastanza lontana dall'origine, per cui risulterà:

$$\Phi(-(z_{1-\alpha/2} + \sqrt{n}\delta/n)) \approx 0 \quad \text{e quindi:} \quad \beta \approx \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\right)$$

Per un test monolaterale, sotto le stesse ipotesi, la dimensione del campione diventa:

$$n \approx \sigma^2 (z_{1-\alpha} + z_{1-\beta})^2 / \delta^2.$$

Test sulla differenza di due medie

Come già accennato, il problema di stabilire se c'è differenza (e di che segno) fra le medie (μ_x e μ_y) di due distribuzioni normali, è particolarmente significativo se si immagina che le due leggi normali possano essere quelle secondo le quali è distribuito l'indice di prestazione di un sistema che può essere gestito con due politiche diverse. Col test sulle medie (\bar{X} e \bar{Y}) di due campioni di osservazioni (x_1, x_2, \dots, x_{n_x} e y_1, y_2, \dots, y_{n_y}) reali, indipendenti e relative alle due diverse politiche si cercherà di stabilire se esse siano equivalenti, e, in caso negativo, quale politica sia preferibile.

Qui si comincerà trattando il caso particolare in cui le due rispettive varianze siano note.

Volendo stabilire se le medie sono uguali o diverse, le ipotesi saranno:

$$H_0: \mu_x = \mu_y$$

$$H_1: \mu_x \neq \mu_y$$

La statistica del test è la seguente:

$$Z_0 = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Se l'ipotesi nulla è vera, allora la statistica è distribuita come una normale standard, per cui si rifiuta l'ipotesi se risulta: $|Z_0| > z_{1-\alpha/2}$.

Passando, per completezza, ai risultati per i due test monolaterali:

Per il test_1:

$$H_0: \mu_x \leq \mu_0$$

$$H_1: \mu_x > \mu_0$$

si rifiuta l'ipotesi nulla se risulta:

$$Z_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} > z_{1-\alpha}.$$

Per il test_2:

$$H_0 : \mu \geq \mu_0$$

$$H_1 : \mu < \mu_0$$

si rifiuta l'ipotesi nulla se risulta:

$$Z_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} < -z_{1-\alpha}$$

- **Il caso più generale, con varianze incognite**

Lavorando con varianze campionarie (S_\bullet^2) si può fare ricorso a statistiche di test (\tilde{T}) distribuite in accordo a leggi “ T_γ di Student” con opportuni gradi di libertà (γ) e possono essere distinti tre casi.

CASO 1: “pooled t-test”

Si suppone che le varianze delle due distribuzioni siano uguali, seppur incognite. In una situazione reale, questo caso può essere applicato qualora si possa sostenere l'ipotesi che due diverse politiche di gestione di un sistema non influiscano sulla varianza dell'indice di prestazione d'interesse.

Con $S_p^2 \triangleq \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}$, la statistica del test è: $\tilde{T} = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \approx T_\gamma$

con $\gamma = n_x + n_y - 2$ gradi di libertà.

CASO 2: “approximate t-test”

Si suppone che le varianze delle due distribuzioni siano diverse.

In tal caso la statistica del test è:

$\tilde{T} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}} \approx T_\gamma$, con $\gamma = \frac{\left(S_x^2/n_x + S_y^2/n_y \right)^2}{\left(S_x^2/n_x \right)^2/(n_x+1) + \left(S_y^2/n_y \right)^2/(n_y+1)} - 2$ gradi di libertà.

CASO 3: “paired t-test”

Questo caso ha senso quando le osservazioni con lo stesso indice sono dipendenti, ma si può assumere che le differenze $d_i \hat{=} x_i - y_i$, $i = 1, \dots, n$ siano realizzazioni di una sequenza (D_1, D_2, \dots, D_n) di variabili aleatorie indipendenti e identicamente distribuite, come una normale di media $D \hat{=} \bar{x} - \bar{y}$ che, quindi, dovrebbe risultare pari a zero sotto l’ipotesi nulla, $\bar{x} = \bar{y}$.

La statistica del test è: $\tilde{T} = \frac{\bar{D} - 0}{\sqrt{S_d^2 / n}} \approx T_\gamma$, con $\gamma = n - 1$ gradi di libertà.

Per tutti e tre i casi delineati, valgono le seguenti regole:

- test bilaterale, si rifiuta l’ipotesi nulla se risulta: $|\tilde{T}| > t_{\gamma; 1-\alpha/2}$.
- test monolaterale_1, invece: $\tilde{T} > t_{\gamma; 1-\alpha}$ e per il test monolaterale_2:
$$\tilde{T} < -t_{\gamma; 1-\alpha}.$$

Il test della chi-quadrato per la bontà della forma

A partire da un campione di osservazioni reali (x_1, x_2, \dots, x_n) indipendenti, nel continuo o nel discreto, con questo test si cerca di stabilire se esse possano essere considerate realizzazioni indipendenti di una variabile aleatoria che abbia una funzione densità nota, ovvero scelta fra quelle disponibili dall'analisi probabilistica. La densità ipotizzata sarà detta $\hat{f}(x)$ nel seguito, e non $f(x)$ come al solito, per evidenziare il fatto che la sua espressione analitica contiene parametri che devono essere stimati in via preliminare al test. Ad esempio, se si trattasse della ben nota densità esponenziale, $f(x) = \lambda \cdot \exp(-\lambda x)$, di parametro λ , allora occorrerebbe stimare questo parametro e poi, detta $\hat{\lambda}$ la stima di λ , si avrebbe la densità esponenziale ipotizzata: $f(x) = \hat{\lambda} \cdot \exp(-\hat{\lambda} x)$.

Sul piano formale, l'ipotesi nulla del test può essere espressa come segue:

$$H_0 : x_1, \dots, x_n \leftarrow \hat{f}(x) = f(x; \hat{\theta})$$

come per chiedersi: “può, la densità ipotizzata, essere quella che caratterizza la variabile aleatoria d'interesse e quindi produrre quel campione di realizzazioni?”

Il primo passo dell'esecuzione del test consiste nel raggruppare i valori x_1, x_2, \dots, x_n in un certo numero di intervalli (k) adiacenti. Si tenga presente sin da ora che scegliere questa numerosità è cosa non banale, anzi è il punto debole del metodo perché i risultati sono abbastanza sensibili a quella scelta. Comunque, per il momento, è il caso di andare avanti definendo il seguente:

N_j = numero di osservazioni reali, x_i , $i=1, \dots, n$ raggruppati nel j -esimo dei k intervalli

Il secondo passo consiste nel calcolare la proporzione p_j di valori x_i , $i=1, \dots, n$ che dovrebbero essere riscontrati nel j -esimo dei k intervalli, qualora la distribuzione ipotizzata fosse quella vera.

Nel continuo si ha:

$$p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x) dx.$$

Nel discreto, invece:

$$p_j = \sum_{a_{j-1} \leq x_i < a_j} \hat{p}(x_i)$$

La statistica del test è la seguente:

$$\tilde{T}(n) \triangleq \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} \equiv \sum_{j=1}^k \frac{(ValoriOsservati - valoriAttesi)^2}{valoriAttesi}$$

perché si può dimostrare che risulta:

$$\tilde{T}(n) \rightarrow \chi^2_{n-1} \quad per \quad n \rightarrow \infty$$

Per eseguire il test, occorre ipotizzare un livello di sicurezza α e quindi confrontare il valore assunto dalla statistica del test con il valore $\chi^2_{k-1,1-\alpha}$. Si rigetta l'ipotesi se $\chi^2 > \chi^2_{k-1,1-\alpha}$ (come illustrato nella figura in seguito).

La difficoltà che si incontra usando questo metodo non risiede tanto e solo nel poter disporre di un campione sufficientemente grande di osservazioni reali, quanto e più nella scelta soggettiva dell'ampiezza degli intervalli e quindi della loro numerosità. La prassi ricorrente, riportata come raccomandazione nella letteratura specializzata, è quella di rispettare le seguenti condizioni:

$$\begin{aligned} k &\geq 3 \\ n \cdot p_j &\geq 5 \quad per \quad ogni \quad j \end{aligned}$$

applicando il test anche con un numero di osservazioni limitate a qualche decina.

Quanto alla scelta degli estremi di ciascun intervallo, $(a_0, a_1), \dots, (a_{j-1}, a_j), \dots, (a_{n-1}, a_n)$, di solito viene seguito un approccio detto equiprobabile, ossia tale che le $p_j \quad j=1, \dots, n$ risultino tutte uguali. Se si fissano le probabilità p_j tali che $p_1 = p_2 = \dots = p_k = p = 1/k$, (con k intervalli), si ottiene $\hat{F}(a_j) = j/k$. Questo è giustificato dalla seguente uguaglianza:

$$\begin{aligned} p &= \int_{a_{j-1}}^{a_j} \hat{f}(x) dx = \hat{F}(a_j) - \hat{F}(a_{j-1}) \\ \Rightarrow 1/k &= j/k - (j-1)/k \end{aligned}$$

da cui si ricava

$$a_j = \hat{F}^{-1}(j/k).$$

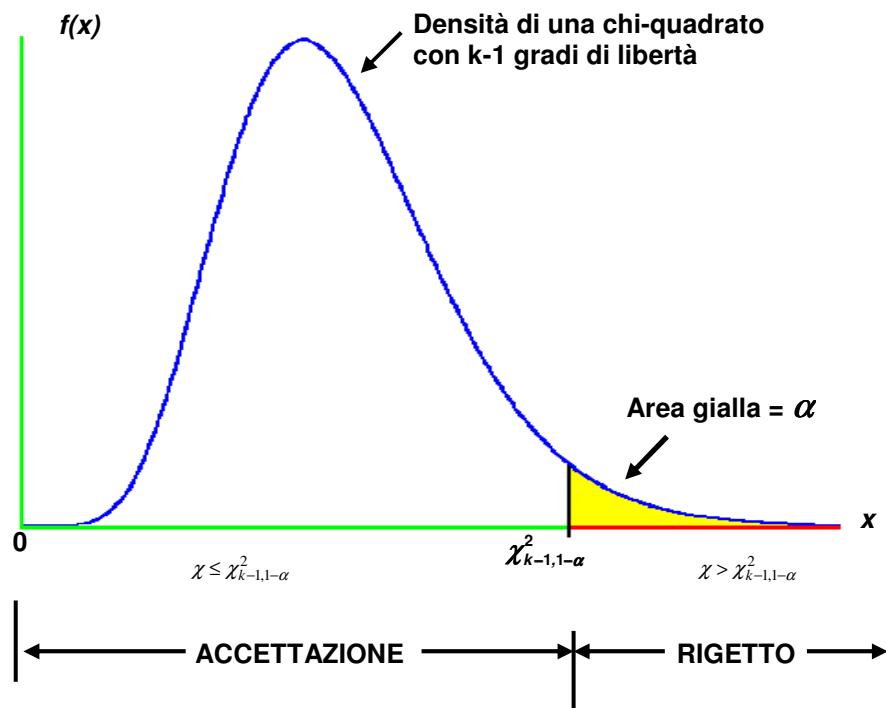
Una difficoltà potrebbe sorgere nel calcolare l'inversa della funzione \hat{F} .

A tal proposito, qui di seguito è descritto il procedimento per una \hat{F} normale.

Fissando la probabilità $p_j = 1/k \Rightarrow \hat{F}(a_j) = j/k$, allora, se la normale è standard il valore di a_j si legge direttamente dalla tabella in corrispondenza del valore di area pari a j/k . Viceversa, se la normale non è standard, si ha: $a_j = \bar{X}(n) + z_j S(n)$, dove z_j è il valore di ascissa per cui si ha la probabilità j/k per la distribuzione normale standard, mentre $S(n)$ e $\bar{X}(n)$ sono gli stimatori, rispettivamente, della deviazione standard e della media. Questo perché $X' \approx N(\mu, \sigma)$ e $X \approx N(0,1)$ sono legate dalla seguente relazione $X' = \bar{X} + \sigma X$. A questo punto si calcola N_j per $j = 1, \dots, n$ e si prosegue.

Alternativamente si possono fissare, anziché le probabilità uguali, gli estremi di integrazioni uguali, ossia si sceglie una certa ampiezza per i k intervalli in cui si è diviso il dominio della distribuzione, uguale per tutti.

Illustrazione del test (della) chi-quadrato



Valori critici della chi-quadrato					
gdl	$\alpha = 0,05$	$\alpha = 0,01$	gdl	$\alpha = 0,05$	$\alpha = 0,01$
1	3,84	6,63	16	26,30	32,00
2	5,99	9,21	17	27,59	33,41
3	7,81	11,34	18	28,87	34,81
4	9,49	13,28	19	30,14	36,19
5	11,07	15,09	20	31,41	37,57
6	12,59	16,81	21	32,67	38,93
7	14,07	18,48	22	33,92	40,29
8	15,51	20,09	23	35,17	41,64
9	16,92	21,67	24	36,42	42,98
10	18,31	23,21	25	37,65	44,31
11	19,68	24,72	26	38,89	45,64
12	21,03	26,22	27	40,11	46,96
13	22,36	27,69	28	41,34	48,28
14	23,68	29,14	29	42,56	49,59
15	25,00	30,58	∞	43,77	50,89

Esempio di esecuzione del test per una legge esponenziale

Facendo riferimento ad una legge esponenziale di parametro fissato, $\lambda = 0.5$, sono stati riprodotti col metodo Monte Carlo, su foglio excel, numerosi campioni di 100 realizzazioni indipendenti. Quindi si è voluto eseguire il test della chi-quadrato su ciascuno di quei campioni per verificare l'occorrenza o meno dell'errore di prima specie, ovvero l'aspetto (in termini di istogramma) dei campioni che portano il test ad accettare l'ipotesi nulla (che in questo caso è ovviamente vera!) e, di contro, l'aspetto di campioni che (pur capitando raramente) portano il test alla conclusione falsa della non appartenenza alle leggi esponenziali in questione, con rigetto dell'ipotesi nulla.

Essendo la legge esponenziale facilmente invertibile, è stato usato il metodo degli intervalli equiprobabili con $p_j = 0.1$, che porta a fissare $k (= p_j)$ in 10.

Allora, da:

$$a_j = F^{-1}(j/k) = -\frac{1}{\lambda} \ln(1 - j/k) = -2 \ln(1 - j/10), \quad j = 0, \dots, 9$$

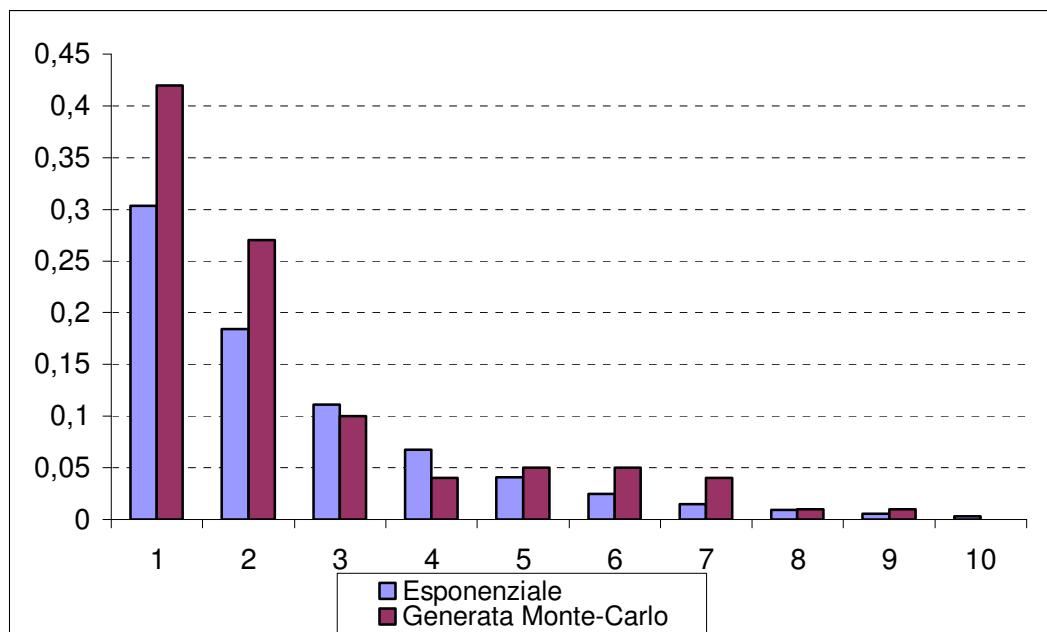
si ricavano i seguenti valori:

$a_0 = 0$	$a_5 = 1.3862$
$a_1 = 0.2107$	$a_6 = 1.8325$
$a_2 = 0.4462$	$a_7 = 2.4079$
$a_3 = 0.7133$	$a_8 = 3.2188$
$a_4 = 1.0216$	$a_9 = 4.6051$

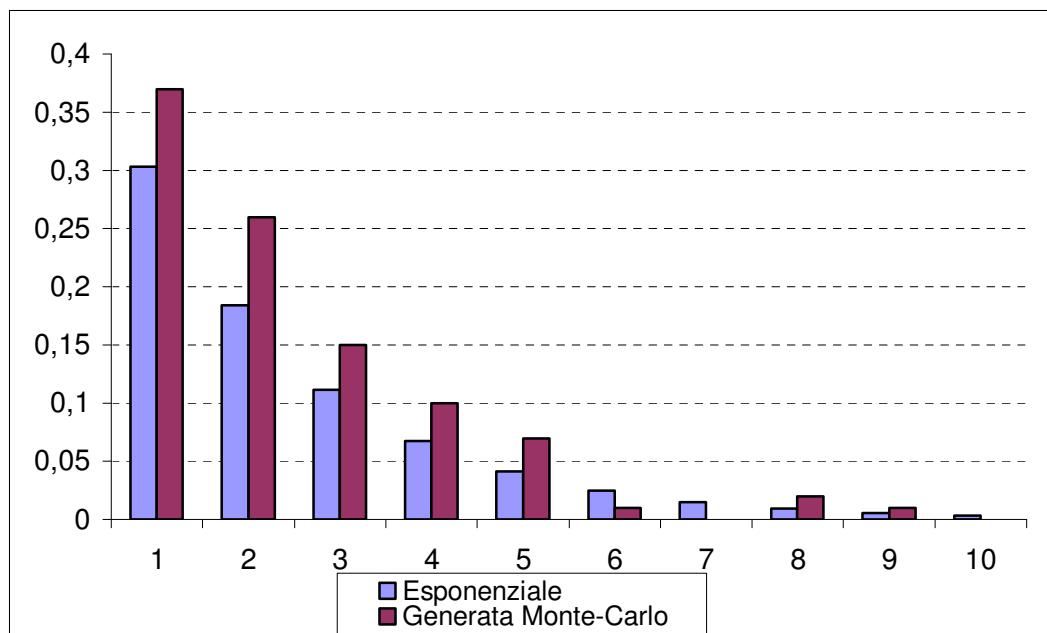
Il livello di significatività del test è stato fissato in $\alpha = 0.05$ e, con 10 intervalli, la statistica del test è una chi-quadrato con 9 gradi di libertà, che offre un valore critico pari a 16.92 ($= \chi^2_{0.95; 9}$).

I due grafici seguenti rappresentano due casi riferiti a due campioni che hanno portato ad esiti opposti del test. Nel primo è stato ottenuto un valore della statistica pari a 6.4 (<16.92) e quindi il campione è stato riconosciuto come generabile dalla legge ipotizzata, mentre nel secondo caso il valore della statistica è risultato pari a 22 (>16.92) e quindi si è verificato l'errore di prima specie.

Caso 1: accettazione dell'ipotesi nulla



Caso 2: rigetto dell'ipotesi nulla



Esercizi di Riepilogo

Esercizio 1

Verificare se la variabile aleatoria X , di cui 50 realizzazioni sono in tabella, si distribuisce con legge esponenziale.

Realizzazioni della X									
79,92	3,027	6,769	18,39	144,7	0,941	0,624	0,59	7,004	3,217
3,081	6,505	59,9	0,141	2,663	0,878	5,38	1,928	31,76	14,38
0,062	0,021	1,192	43,57	17,97	3,371	3,148	0,3	1,005	1,008
1,961	0,013	34,76	24,42	0,091	2,157	7,078	0,002	1,115	2,336
5,845	0,123	5,009	0,433	9,003	7,579	23,96	0,543	0,219	4,562

Per poter applicare il test della chi-quadrato, si **raccomanda** $k \geq 3$, $n \cdot p_i \geq 5$ per ogni i . Inoltre, nel continuo, per la scelta del numero di intervalli, si faccia riferimento ai seguenti *range*.

Numerosità del campione (n)	Numero di Intervalli (k)
20	non usare CG
50	5-10
100	10-20
>100	n-n/5

Per prima cosa, occorre indicare il numero di intervalli in accordo a quanto suggerito dalla tabella sopra riportata. Avendo a disposizione 50 osservazioni, scegliamo un numero di intervalli k pari a 8 (compreso, cioè tra 5 e 10). Precisiamo che si fa riferimento ad intervalli equiprobabili, quindi, la probabilità che un'osservazione cada in un (qualsiasi) intervallo è data da $p = 1/8 = 0,125$.

Per completare la definizione dei singoli intervalli, occorre calcolare i loro estremi, operazione che, a sua volta, richiede l'impiego della media. La media è calcolata a partire dai dati in tabella ovvero:

$$\lambda = 1/\text{media} = 1/11,893 = 0,084$$

Si possono ora calcolare gli estremi degli intervalli a_i a partire dalla funzione di distribuzione di probabilità:

$$F(a_i) = 1 - e^{-\lambda a_i}$$

$$ip = 1 - e^{-\lambda a_i}$$

$$e^{-\lambda a_i} = 1 - ip$$

$$a_i = -\frac{1}{\lambda} \ln(1 - ip)$$

Per qualsiasi valori di λ , gli estremi inferiore e superiori sono $a_0 = 0$ e $a_k = \infty$.

Per gli altri estremi a_i per $i = 1, \dots, 7$:

$$a_1 = -\frac{1}{\lambda} \ln(1 - 0,125) = 1,588 \quad a_2 = -\frac{1}{\lambda} \ln(1 - 0,250) = 3,422$$

$$a_3 = -\frac{1}{\lambda} \ln(1 - 0,375) = 5,590 \quad a_4 = -\frac{1}{\lambda} \ln(1 - 0,500) = 8,244$$

$$a_5 = -\frac{1}{\lambda} \ln(1 - 0,625) = 11,665 \quad a_6 = -\frac{1}{\lambda} \ln(1 - 0,750) = 16,488$$

$$a_7 = -\frac{1}{\lambda} \ln(1 - 0,875) = 24,732$$

Intervallo	O_i	E_i	$(O_i - E_i)^2 / E_i$
[0, 1588)	19	6,25	26,01
[1,588, 3,422)	10	6,25	2,25
[3,422, 5,590)	3	6,25	1,69
[5,590, 8,244)	6	6,25	0,01
[8,244, 11,665)	1	6,25	4,41
[11,665, 16,488)	1	6,25	4,41
[16,488, 24,732)	4	6,25	0,81
[24,732, ∞)	6	6,25	0,01

In quest'ultima tabella, per ogni intervallo i , riportiamo il numero di osservazioni osservate (O_i), il numero di osservazioni attese ($E_i = n \cdot p_i = 50 \cdot 0,125 = 6,25$ $i = 1, \dots, 8$) ed, infine, un calcolo nell'ultima colonna che determina il valore della statistica del caso di interesse.

Sommendo i valori dell'ultima colonna risulta $\chi_0^2 = 39,6$.

Dalla tabella dei valori critici del test, per un livello di significatività $\alpha = 0,05$ risulta un valore di $\chi_{k-s-1,\alpha}^2 = 12,59$ (da notare che, avendo stimato un parametro dell'ipotetica distribuzione, ovvero la media, si considera un numero di gradi di libertà decurtato proprio del numero di parametri stimati).

Essendo $\chi_0^2 = 39,6 > 12,59 = \chi_{k-s-1,\alpha}^2$ si rigetta l'ipotesi nulla secondo la variabile aleatoria X si distribuisce con legge esponenziale. Il risultato del test non cambia per un livello di significatività $\alpha = 0,01$ in quanto $\chi_0^2 = 39,6 > 16,81 = \chi_{k-s-1,\alpha}^2$.

Valore atteso condizionato e curva di regressione

Sia X una prima variabile aleatoria vista quale condizionante e Y una seconda, condizionata dalla prima. Ad esempio, in un modello produttore consumatore, potrebbe essere X = “tempo fra due consumi consecutivi” e Y = “durata della giacenza nel buffer”.

Riprendendo la formula del valore atteso condizionato:

$$E[Y | X = x] \hat{=} \int_{y=0}^{\infty} y \cdot f(y | x) \cdot dy$$

si osserva che essa definisce una funzione sullo spazio delle realizzazioni (continue e non negative) della X :

$$E[Y | X = x], \quad 0 \leq x < \infty.$$

Tale funzione, che descrive l'andamento del valore atteso della Y al variare della x , è detta curva di regressione e si potrebbe ricavare, in linea di principio, a partire dalla conoscenza della densità condizionata

$$f_{Y|X}(y | x)$$

o della congiunta

$$f_{Y,X}(y, x).$$

In pratica, è più spesso compito della statistica stimare i parametri della curva di regressione, a partire da osservazioni sperimentali della coppia (X, Y) .

La retta di regressione

La retta di regressione è quella particolare curva di regressione che si ottiene ponendo:

$$E[Y | X = x] \doteq a \cdot x + b, \quad 0 \leq x < \infty$$

ed è comunemente usata nell'analisi (statistica) della dipendenza della variabile aleatoria Y dalla X , dopo aver stimato i coefficienti reali a e b a partire dalle realizzazioni sperimentali della coppia X, Y .

Si può dimostrare che, indipendentemente dalla forma completa delle funzioni di distribuzione della X e della Y , i coefficienti a e b potrebbero essere ricavati dalle seguenti formule:

$$a = \rho \frac{\sqrt{VAR[Y]}}{\sqrt{VAR[X]}} = \frac{COV(X, Y)}{VAR[X]}; \quad b = E[Y] - \frac{COV(X, Y)}{VAR[X]} E[X];$$

dove ρ è il coefficiente di correlazione di Pearson.

Punto di partenza della dimostrazione è la formula del valore atteso condizionato:

$$\int_{y=0}^{\infty} y \cdot f(y | x) \cdot dy \doteq E[Y | X = x]$$

riscritta usando la densità congiunta e la densità marginale:

$$\frac{1}{f_X^{(m)}(x)} \int_{y=0}^{\infty} y \cdot f_{Y,X}(x, y) \cdot dy = E[Y | X = x]$$

Con essa si può scrivere:

$$\int_0^{\infty} y f_{Y,X}(x, y) dy = (ax + b) f_X^{(m)}(x)$$

e integrando in x ambo i membri

$$\int_{x=0}^{\infty} \int_0^{\infty} y f_{Y,X}(x, y) dy \cdot dx = \int_{x=0}^{\infty} (ax + b) f_X^{(m)}(x) \cdot dx$$

si ottiene:

$$E[Y] = a \cdot E[X] + b. \quad (\text{r1})$$

OSSERVAZIONE: la relazione di linearità tra i valori attesi di Y e X è cosa ben diversa dalla relazione di linearità tra le variabili aleatorie Y e X e non implicante quest'ultima!

Ora, riprendendo la: $E[Y|X=x] \doteq a \cdot x + b$ e integrando ambo i membri, dopo aver moltiplicato entrambi per “ $x \cdot f_X(x)$ ” risulta

$$\begin{aligned} & \int_0^\infty x \cdot f_X(x) \cdot E[Y|x] \cdot dx = \int_0^\infty x \cdot f_X(x) \cdot (ax + b) \cdot dx \\ &= \int_0^\infty x \cdot f_X(x) \cdot \left\{ \int_0^\infty y \cdot f_{Y|X}(y|x) \cdot dy \right\} \cdot dx = b \cdot E[X] + a \cdot E[X^2] \\ &= \int_0^\infty \int_0^\infty xy \cdot f_{X,Y}(x,y) \cdot dxdy = b \cdot E[X] + a \cdot E[X^2] \end{aligned}$$

ovvero:

$$E[X \cdot Y] = b \cdot E[X] + a \cdot E[X^2] \quad (\text{r2})$$

ma, per altra via:

$$\begin{aligned} E[X \cdot Y] &= COV[X, Y] + E[X]E[Y] \\ &= \rho \sqrt{VAR[X]} \sqrt{VAR[Y]} + E[X]E[Y] \end{aligned}$$

Allora, usando la (r2) e ricordando che $E[X^2] = Var[X] + E[X]^2$:

$$\rho \sqrt{VAR[X]} \sqrt{VAR[Y]} - E[X]E[Y] = b \cdot E[X] + a \cdot \{Var[X] + E[X]^2\}$$

A questo punto, invocando la (r1), si riconosce che $b = E[Y] - a \cdot E[X]$ e si ricava:

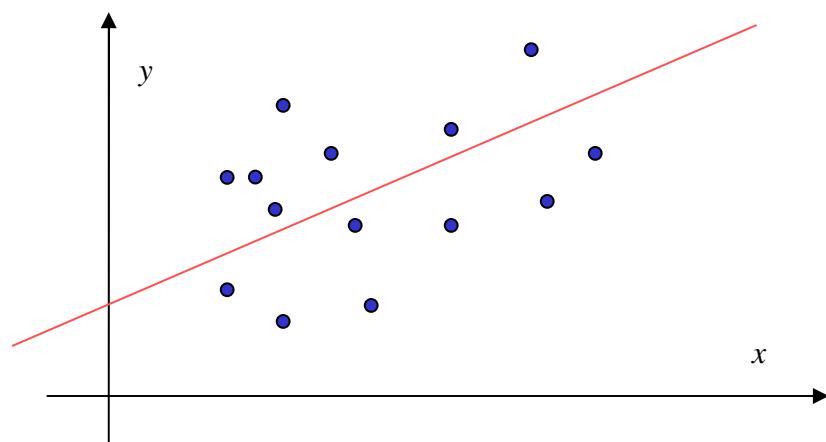
$$a = \rho \frac{\sqrt{VAR[Y]}}{\sqrt{VAR[X]}} \quad \text{e quindi} \quad b = E[Y] - \rho \frac{\sqrt{VAR[Y]}}{\sqrt{VAR[X]}} E[X]$$

In definitiva, è stata dimostrata la seguente:

$$E[Y|X=x] \doteq E[Y] + \rho \frac{\sqrt{VAR[Y]}}{\sqrt{VAR[X]}} (x - E[X]), \quad 0 \leq x < \infty$$

Stima dei parametri della retta di regressione

Qui si farà vedere come possano essere determinati i parametri (a e b) della retta di regressione, a partire da un insieme di coppie $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ di realizzazioni congiunte delle variabili aleatorie X e Y . L'idea è che la retta di regressione debba essere proprio quella che, nel piano euclideo, passa “il più possibile vicino” ai punti corrispondenti alle coppie di realizzazioni $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. La figura seguente illustra l'idea:



Per formalizzare il concetto di “il più possibile vicino”, si conviene di cercare i parametri a e b tali che risulti minima la somma dei quadrati degli scostamenti dei valori y_1, y_2, \dots, y_n osservati rispetto ai valori sulla retta stessa:

$$\sum_{i=1}^n [y_i - (ax_i + b)]^2 .$$

Considerando, dunque, la precedente somma (S) come funzione dei due parametri a e b , si deriva prima rispetto all'uno e poi rispetto all'altro:

$$\begin{cases} \frac{\partial S(a, b)}{\partial b} = -2 \cdot \sum_{i=1}^n (y_i - b - ax_i) = 0 \\ \frac{\partial S(a, b)}{\partial a} = -2 \cdot \sum_{i=1}^n (y_i - b - ax_i) \cdot x_i = 0 \end{cases}$$

da cui:

$$\begin{cases} n \cdot b + a \cdot \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0 \\ b \cdot \sum_{i=1}^n x_i + a \cdot \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i = 0 \end{cases}.$$

Risolvendo rispetto ad a e b si ricavano le stime ai minimi quadrati (denotate come “ \hat{a} ” e “ \hat{b} ”:

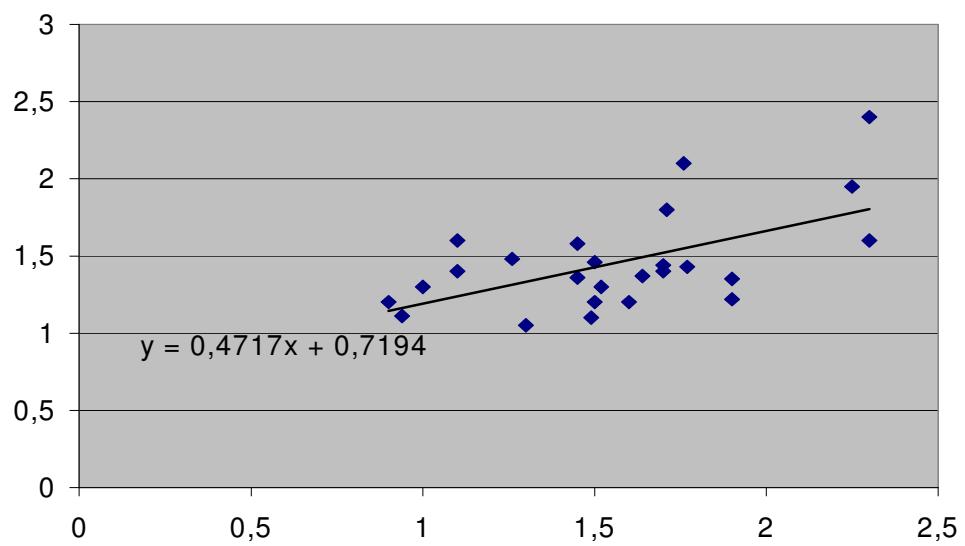
$$\hat{a} = \frac{n \cdot \sum_{i=1}^n x_i \cdot y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad \text{e} \quad \hat{b} = \frac{\sum_{i=1}^n y_i}{n} - \hat{a} \cdot \frac{\sum_{i=1}^n x_i}{n}, \quad (\text{r3})$$

che risultano in accordo con le formule già ricavate:

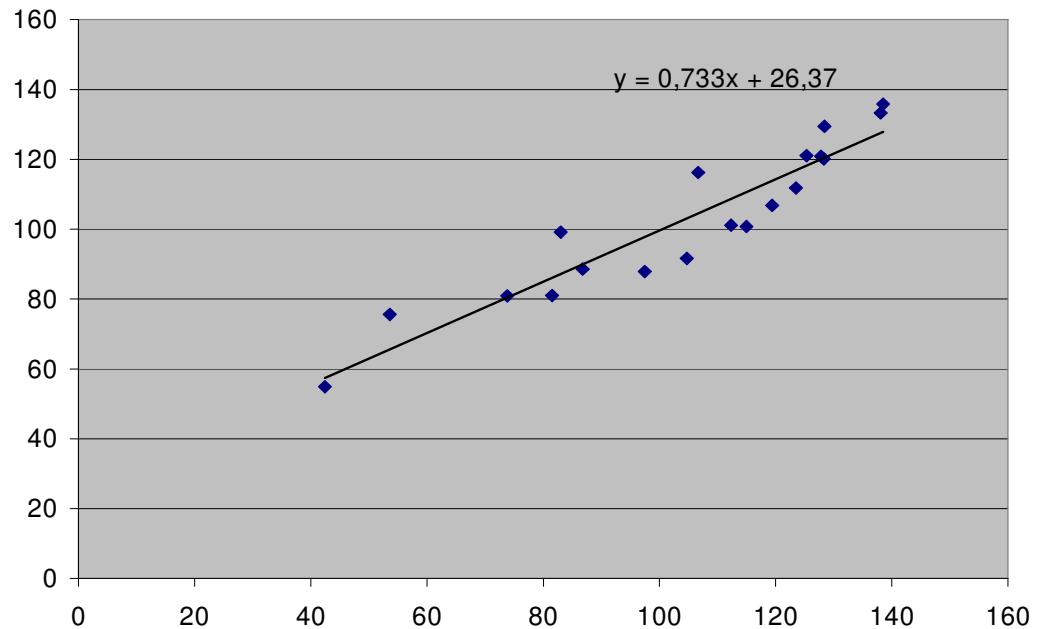
$$a = \frac{COV(X, Y)}{VAR[X]} \quad \text{e} \quad b = E[Y] - a \cdot E[X].$$

Esempi di rette di regressione

Due esempi numerici di costruzione di rette di regressione sono illustrati nelle prossime due figure e sono tratti da un caso reale di analisi della correlazione presso un terminale marittimo per container, dove erano state individuate due sottoaree separate di stoccaggio dei container sul piazzale che potevano ospitare gruppi di container che arrivavano con la stessa nave e con una seconda nave, ancora comune, erano destinati a ripartire. In tal caso, sia i tempi di giacenza nelle due sottoaree sia i rispettivi livelli di occupazione dovevano risultare dipendenti, come confermato dalle rette di regressione ricavate con i dati della pagina seguente.



Retta di regressione per i tempi di giacenza (settimane)



Retta di regressione per i livelli di occupazione (unità)

Dati per la costruzione delle rette di regressione

Rilevazioni congiunte

Tempi di giacenza di singoli container

(settimane)

$$y = 0,4717x + 0,7194$$

Livelli di occupazione delle due aree

(unità)

$$y = 0,733x + 26,37$$

0,9	1,2
1,9	1,35
1,5	1,46
1,45	1,36
1,1	1,6
1,6	1,2
1,45	1,58
0,94	1,11
1,3	1,05
1,1	1,4
1,9	1,22
1,26	1,48
2,3	1,6
1	1,3
1,52	1,3
1,7	1,44
1,5	1,2
1,49	1,1
2,3	2,4
1,7	1,4
1,71	1,8
1,64	1,37
1,76	2,1
2,25	1,95
1,77	1,43

42,4	54,9
73,8	80,9
83	99,2
106,7	116,2
128,4	129,5
138,5	135,8
138,1	133,3
125,3	121,1
128,3	120,1
127,8	120,9
123,5	111,8
119,4	106,8
115	100,8
112,3	101,1
104,7	91,6
86,8	88,6
97,5	87,9
81,5	81
53,6	75,6

Un esempio di analisi (puntuale) della correlazione

Si consideri un semplice modello produttore-consumatore operante con una logica di tipo “pure push”, cioè con un produttore che immette i prodotti appena realizzati nel buffer in accordo al suo ritmo di produzione e senza tenere conto del ritmo di consumo. Il sistema è stato generato col metodo Monte Carlo su foglio Excel (tasso di produzione $\lambda = 1$ e tasso di consumo $\mu = 1$, entrambi riferiti a leggi esponenziali) e sono state effettuate alcune stime del coefficiente di correlazione di Pearson per le seguenti coppie di variabili aleatorie d’interesse:

- “tempo fra due produzioni consecutive” (indicata con P);
- “tempo fra due consumi consecutivi”; (indicata C);
- “tempo di giacenza nel buffer”; (indicata con G);
- “tempo di vita (tempo di giacenza + tempo di consumo)” (indicata con V).

Scopo dell’esperimento era appunto quello di verificare che la stima puntuale del coefficiente di Pearson confermasse:

- 1) l’assenza di correlazione fra tempo di produzione e tempo di consumo, vista la logica di funzionamento di tipo “pure push”;
- 2) una forte correlazione positiva fra tempo di vita dei prodotti e tempo di giacenza, segno che il sistema lavora con livelli e tempi di giacenza alti, visto che il ritmo di consumo è solo del 10% più alto del ritmo di produzione;
- 3) una correlazione più o meno significativa, negativa, fra tempi di produzione collocati in un determinato intervallo temporale e tempi di giacenza collocati in un intervallo successivo, atteso che una dilatazione/riduzione dei tempi di produzione consente di diminuire/aumentare le scorte correnti a beneficio/danno dei tempi di giacenza dei prodotti realizzati successivamente.

Per tutte le variabili aleatorie di cui sopra (P, C, G e V) sono state registrate 30 realizzazioni a partire dal prodotto n° 2001 e fino al 2030. Esse sono riportate nella prossima tabella, colonna per colonna. Per la sola variabile P sono state registrate ulteriori 30 osservazioni, riferite al gruppo di prodotti dal n°1971 al 2000, in accordo alla opportunità dell’analisi di correlazione fra tempi di produzione e tempi di giacenza temporalmente sfalsati (punto 3).

Per ricavare la formula della stima puntuale (stima) del coefficiente di correlazione di Pearson si può partire dalla stima (\hat{a}) del parametro “a” (derivata) della retta di regressione e usare una stima per la varianza, dato che risulta:

$$\rho = a \cdot \frac{\sqrt{VAR[X]}}{\sqrt{VAR[Y]}}, \quad \text{con } X \text{ e } Y \text{ variabili aleatorie generiche.}$$

Allora, riprendendo la formula di stima di “a”, ricavata in precedenza:

$$\hat{a} = \frac{n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

e riscrivendola così:

$$\hat{a} = \frac{\left(n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right) / n^2}{\left[n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] / n^2}$$

si perviene alle seguenti (\bar{x} e \bar{y} indicano le medie aritmetiche degli n rispettivi valori)

$$\begin{aligned} \hat{a} &= \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \quad (+)$$

e ricordando che risultava pure:

$$a = \frac{COV(X, Y)}{VAR[X]},$$

risulta naturale suggerire il numeratore delle (+) quale formula di stima della covarianza e il denominatore delle (+) quale stima della varianza.

A questo punto ritornando alla $\rho = a \cdot \frac{\sqrt{VAR[X]}}{\sqrt{VAR[Y]}}$ e disponendo pure della formula di

stima della varianza, si ottiene la formula di stima del coefficiente di correlazione di Pearson:

$$\hat{\rho} = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2}}$$

Applicando questa formula ai dati in tabella si può verificare che risultano i seguenti valori:

- $\hat{\rho}(P, C) = 7.37 \cdot 10^{-5}$ in risposta al punto 1);
- $\hat{\rho}(V, G) = 0.993$ in risposta al punto 2);
- $\hat{\rho}(P_{[1971-2000]}, G) = -0.433$ e $\hat{\rho}(P_{[2001-2030]}, G) = -0.095$ in risposta al punto 3).

Provando a ripetere l'esperimento con tabelle di dati diversi ci si potrà fare un'idea della variabilità della stima puntuale del coefficiente di correlazione di Pearson. Tale variabilità potrebbe generare dubbi sulle osservazioni puntualizzate sopra, in 1), 2) e 3).

P [2001-2030]	C [2001-2030]	G [2001-2030]	V [2001-2030]		P [1971-2000]
0,0799	1,3904	9,4263	10,8167		1,0167
0,1392	0,0699	10,6775	10,7474		0,6407
0,4812	0,4510	10,2661	10,7172		0,2508
0,5034	1,9104	10,2137	12,1242		0,1321
1,8570	2,5478	10,2671	12,8149		0,0812
3,1738	0,0959	9,6411	9,7370		0,3612
1,2192	0,2642	8,5179	8,7820		1,8262
0,0814	0,3477	8,7006	9,0483		0,1758
0,2612	0,0679	8,7871	8,8550		0,1586
0,1636	0,9072	8,6914	9,5986		2,4498
0,0775	0,1815	9,5211	9,7025		1,3183
1,6235	0,0889	8,0790	8,1679		0,6070
0,5616	0,3442	7,6063	7,9505		0,1328
5,0594	0,6896	2,8912	3,5807		0,9086
0,2806	0,1777	3,3001	3,4778		0,6558
0,6495	0,2149	2,8283	3,0431		1,0054
0,3870	0,2807	2,6561	2,9368		3,9550
0,6724	1,3549	2,2645	3,6193		1,0270
1,1667	0,2506	2,4527	2,7033		0,0155
1,0052	0,2268	1,6981	1,9249		4,5030
1,1825	0,1683	0,7423	0,9106		0,7205
0,6949	0,3446	0,2158	0,5604		2,0135
0,8725	0,8444	0,0000	0,8444		0,0056
0,3876	0,6460	0,4568	1,1029		1,1770
0,8284	0,5893	0,2745	0,8638		1,2998
0,7938	1,8298	0,0700	1,8998		0,1909
0,0002	0,7431	1,8996	2,6427		0,8553
0,2010	0,1206	2,4418	2,5624		0,4934
0,1216	0,4480	2,4408	2,8888		2,4960
0,8675	0,1442	2,0213	2,1655		4,9174

Tabella. Campioni di dati ricavati da un'implementazione del modello produttore-consumatore col metodo Monte Carlo, (tutti espressi in unità di tempo).

Perfetta regressione lineare

Con riferimento al campione di realizzazioni congiunte $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ delle variabili aleatorie X e Y , si può definire perfetta quella retta di regressione con parametri a e b per i quali risulta $y_i = a \cdot x_i + b$, $i = 1, 2, \dots, n$. Ovviamente si tratta di un caso limite, perché è assai raro che gli n punti sperimentali della Y giacciono tutti su un'unica retta.

Più in generale, indicando con δ_i , $i = 1, \dots, n$ le deviazioni dei valori osservati y_i , $i = 1, 2, \dots, n$ rispetto ai corrispondenti valori sulla retta di regressione e ponendo $\delta_i \hat{=} |y_i - (a \cdot x_i + b)|$, $i = 1, 2, \dots, n$, è opportuno interpretare quelle deviazioni come realizzazioni (indipendenti) di una variabile aleatoria Δ . Così si può riconoscere che il criterio dei minimi quadrati per la determinazione della retta di regressione corrisponde a minimizzare la stima del momento del secondo ordine della Δ , $E^2[\Delta]$.

Infatti:

$$\hat{E}^2[\Delta] = \left(\sum_{i=1}^n \delta_i^2 \right) / n$$

Più precisamente, ricordando che:

- se Y e X sono legate da una perfetta relazione lineare allora risulta (r1, a pag.99) che $E[Y] = a \cdot E[X] + b$ e quindi $E[\Delta] = E[Y] - (a \cdot E[X] + b) = 0$,
- $VAR[\Delta] = E^2[\Delta] - (E[\Delta])^2$

si deduce, infine, che il criterio dei minimi quadrati corrisponde a determinare la retta cui corrisponde una variabile aleatoria Δ di media nulla e di varianza minima.

Questa retta è individuata dalle stime, \hat{a} e \hat{b} , (r3 a pag. 102), dei seguenti parametri:

$$a = \rho \frac{\sqrt{VAR[Y]}}{\sqrt{VAR[X]}} \quad (+) \quad b = E[Y] - \rho \frac{\sqrt{VAR[Y]}}{\sqrt{VAR[X]}} E[X]$$

Per capire come si possa stabilire l'esistenza di una perfetta regressione lineare a partire dalla stima ($\hat{\rho}$) del coefficiente di correlazione e, in ultima analisi, dal campione di realizzazioni congiunte $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, è determinante la seguente implicazione:

$$|\rho| = 1 \Rightarrow Y = aX + b$$

Prova

Da $\Delta \hat{=} Y - (aX + b)$ risulta:

$$VAR[\Delta] = VAR[Y] + a^2 VAR[X] - 2a \cdot COV[X, Y]$$

e usando la (+):

$$VAR[\Delta] = VAR[Y] + \rho^2 VAR[Y] - 2\rho^2 \cdot VAR[Y]$$

ovvero:

$$VAR[\Delta] = VAR[Y] \cdot (1 - \rho^2). \quad (\$)$$

Da qui:

$$|\rho| = 1 \Rightarrow \rho^2 = 1 \Rightarrow VAR[\Delta] = 0$$

e, quindi, si avrà:

$$y_i = a \cdot x_i + b, \quad i = 1, 2, \dots, n,$$

ovvero tutte le realizzazioni della Y saranno proprio sulla retta.

OSSERVAZIONE CONCLUSIVA:

la relazione (\$) è interessante e meriterebbe di essere approfondita, perché essa mette in luce che il termine $(1-\rho^2)$ collega la varianza degli scarti alla varianza della Y.

Simulazione Monte Carlo in Excel

La simulazione Monte Carlo è una tecnica utilizzata per riprodurre e risolvere numericamente le funzioni descrittive di un sistema di interesse nella cui formulazione sono coinvolte anche variabili aleatorie. Per modelli semplici, l'implementazione di tale tecnica sotto MS Excel è agevole e riassumibile nei seguenti step:

- Step 1:* Identificare una funzione di ingresso-uscita rispetto al sistema di interesse, $Y = f(X_1, X_2, \dots, X_q)$;
- Step 2:* Generare un insieme di input casuali, $X_{i1}, X_{i2}, \dots, X_{iq}$;
- Step 3:* Valutare la funzione e memorizzare i risultati come Y_i ;
- Step 4:* Ripetere gli step 2 e 3 per $i = 1..n$;
- Step 5:* Analizzare i risultati mediante l'impiego di istogrammi, statistiche descrittive, intervalli di confidenza, etc..

Il modello Produttore Consumatore

Ai fini di illustrare l'implementazione della simulazione Monte Carlo sotto Excel, si riconsidera il modello Produttore – Consumatore illustrato in Figura 1.

Posto che in una realtà industriale la produzione di un certo bene (produttore) destinato a magazzino (buffer) avviene in modo non deterministico e le richieste di consumo (consumatore) di detto bene si manifestano in modo aleatorio, si vuole stimare il tempo di attesa in coda (o di giacenza in magazzino) ed il tempo di soggiorno.

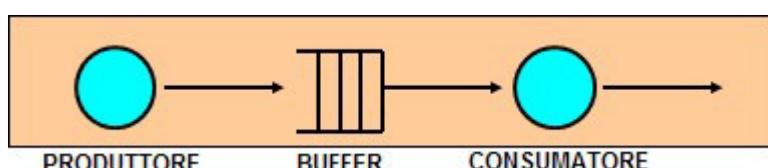


Figura 1 – Il modello Produttore - Consumatore

Si utilizzi la simulazione Monte Carlo per stimare il tempo di soggiorno ed il tempo di attesa in coda (o di giacenza in magazzino).

Identificazione di una funzione di ingresso-uscita

L'identificazione della(e) funzione(i) di ingresso-uscita rispetto al modello Produttore – Consumatore avviene in relazione agli indici di *performance* da stimare. A tal fine, si sceglie di seguire un approccio di tipo *top-down* a partire dalle definizioni delle seguenti quantità:

- W_i = tempo di attesa in coda (o di giacenza in magazzino) dell'i-esimo prodotto;
- S_i = tempo di servizio (o di consumo) dell'i-esimo prodotto;
- SG_i = tempo di soggiorno dell'i-esimo prodotto;
- A_i = intervallo di tempo tra l'arrivo dell'i-esimo prodotto e quello dell'(i+1)-esimo – in linguaggio “tecnico” *interarrivo* dell'(i+1)-esimo prodotto.

Si suppone, inoltre, che la gestione dei prodotti giacenti in magazzino avvenga secondo la disciplina FIFO.

Il primo prodotto che arriva attende un tempo nullo, dunque $W_1 = 0$. Per calcolare il tempo di attesa in coda (o di giacenza in magazzino) dei prodotti successivi, si impiega la seguente formula ricorsiva nota come equazione di *Lindley*:

$$W_{i+1} = \max(0, W_i + S_i - A_i) \quad (1)$$

ovvero il tempo di attesa dell'(i+1)-esimo prodotto è pari alla somma del tempo d'attesa in coda (o di giacenza in magazzino) e del tempo di servizio (o di consumo) dell'i-esimo prodotto meno il tempo di interarrivo dell'(i+1)-esimo prodotto. In sintesi, l'utente (i+1)-esimo può arrivare durante o dopo che l'utente i-esimo sia partito, ma non può arrivare prima; quando ciò succede il valore negativo porta a scegliere lo 0.

Una volta determinato, il tempo di attesa in coda (o di giacenza in magazzino) dell'i-esimo prodotto sommato al corrispondente tempo di servizio (o di consumo) restituisce il tempo di soggiorno dell'i-esimo prodotto, ovvero

$$SG_i = W_i + S_i. \quad (2)$$

Identificate le funzioni di ingresso-uscita (1) e (2), in definitiva per il modello Produttore – Consumatore le variabili di ingresso alla simulazione Monte Carlo sono rappresentate dalle variabili aleatorie (A_i, S_i) .

Generazione degli input

Il concetto principale alla base della simulazione Monte Carlo è la generazione di un insieme di numeri casuali e, quindi, di realizzazioni di variabili aleatorie in ingresso al sistema di interesse. Come è lecito aspettarsi nell'adozione di un qualunque approccio di modellizzazione e/o modello di previsione, vige il principio *garbage in equals garbage out*. In questa sede, non si prendono in considerazione problematiche legate alla corretta scelta né della funzione di distribuzione per rappresentare l'ingresso al sistema, né del generatore di numeri casuali.

Il modello Produttore – Consumatore introdotto prevede l'impiego di una Funzione di Distribuzione Esponenziale per rappresentare la variabile aleatoria di ingresso “interarrivo” ($F_{X_1} = 1 - e^{-\lambda x_1}$ con parametro λ posto pari ad 1 u.t.^{-1} nella cella D1) e “tempo di servizio” (o di consumo) ($F_{X_2} = 1 - e^{-x_2}$ con parametro pari ad $1,1 \text{ u.t.}^{-1}$ nella cella H1).

Per generare una realizzazione di una variabile aleatoria distribuita secondo una legge esponenziale, si impiegano le seguenti formule in Excel (con richiamo delle opportune celle di riferimento):

$$nc = \text{CASUALE}()$$

da inserire in

$$\begin{aligned} &= -\text{LN}(nc) * (1/\lambda) \\ &= -\text{LN}(nc) * (1/) \end{aligned}$$

dove $-\text{LN}(nc) * (1/\lambda)$ e $-\text{LN}(nc) * (1/)$ sono rispettivamente il valore delle funzioni inverse ($F_{X_1}^{-1}(nc)$) e ($F_{X_2}^{-1}(nc)$), ciascuna valutata in corrispondenza di un numero casuale, appunto nc , appartenente all’intervallo $[0,1]$. Inoltre, come mostrato più avanti, il ricorso all’uso della funzione **CASUALE()** di Excel consente il ricalcolo del foglio di lavoro ad ogni nuovo numero casuale generato.

Impostate le formule, si decide di fissare un numero di realizzazioni molto alto (e.g. $n = 3200$ realizzazioni) perché, per la legge forte dei grandi numeri, la media campionaria tende al valore atteso ($1/\lambda$ o $1/$) con probabilità 1 per n tendente ad infinito.

Sotto Excel un modo molto conveniente per organizzare le realizzazioni di ciascuna variabile aleatoria di interesse consiste nel dedicare una colonna per ogni passaggio descritto precedentemente, come illustrato di seguito in Figura 2.

	C	D	E	F	G	H	I	J
7	Numeri Casuali	Tempi di Interarrivo	Istanti di Arrivo			Numeri Casuali	Tempi di Servizio	Istanti di Partenza
8								
9	0,397378926	0,922864981	0,922864981			0,869756995	0,126855839	1,049720819
10	0,908582227	0,095869887	1,018734867			0,094773897	2,142055683	3,191776502
11	0,167161924	1,788792332	2,807527199			0,985773417	0,013026137	3,204802639
12	0,155071361	1,863869875	4,671397074			0,492057418	0,644690786	5,316087861
13	0,684649602	0,378848101	5,050245175			0,707233528	0,314903964	5,630991824
14	0,250939143	1,382544828	6,432790003			0,951801191	0,044908272	6,477698275
15	0,703292904	0,351981825	6,784771828			0,850820789	0,146867056	6,931638884
16	0,597748902	0,51458451	7,299356338			0,12229481	1,910291522	9,20964786
17	0,503246759	0,6866674654	7,986030992			0,268655659	1,194840723	10,40448858
18	0,96378087	0,036891324	8,022922316			0,678558491	0,352531452	10,75702003
19	0,281208884	1,268657525	9,291579841			0,334267619	0,996193956	11,75321399
20	0,568584653	0,56460507	9,856184911			0,053614981	2,859933407	14,4131474

Figura 2 - Input del modello Produttore - Consumatore

In particolare, la realizzazione della variabile aleatoria “istante di arrivo” contenuta nella cella E10 è ottenuta con la formula:

$$= \$E9 + \$D10$$

in base alla quale l’istante di arrivo del secondo prodotto (E10) si differenzia dall’istante di arrivo del prodotto precedente (E9) per un intervallo di tempo pari al proprio tempo di interarrivo (D10). Ovviamente, per il primo prodotto l’istante di arrivo è pari al corrispondente valore di interarrivo (E9=D9).

Analogamente, la realizzazione della variabile aleatoria “tempo di partenza” (o di fine consumo) contenuta nella cella J10 è ottenuta con la formula:

$$= MAX(\$J9 + \$I10; \$E10 + \$I10)$$

in virtù della quale il “tempo di partenza” (o di fine consumo) del secondo prodotto (J10) si differenzia da quello del prodotto precedente (J9) per un intervallo di tempo pari al massimo tra il tempo di partenza (o di fine consumo) del suo predecessore (J9) sommato al proprio tempo di servizio (o consumo) (D10) ed il proprio istante di arrivo (E10) sommato al proprio tempo di servizio (o consumo) (D10). Ovviamente, per il primo prodotto il “tempo di partenza” (o di fine consumo) è dato dal proprio istante di arrivo incrementato del corrispondente tempo di servizio (o consumo) (J9=E9+I9).

Per generare 3200 realizzazioni di entrambe le variabili aleatorie ora descritte, non resta che copiare le suddette formule nelle rimanenti 3198 righe delle rispettive colonne.

Valutazione del modello

Dal momento che, sotto le opportune ipotesi, il modello Produttore – Consumatore non è molto complesso, per la sua valutazione è sufficiente dedicare una colonna ad ogni indice di performance da stimare e completare tali colonne con le relative formule come mostrato in Figura 3.

	N	P
7	Tempi di Soggiorno	Equazione di Lindley Wi
8	2,081370295	0
9	3,182472296	1,88264185
10	2,974087577	2,880963746
11	1,030790104	1,027841217
12	1,720562433	0,90069226
13	0,915615513	0,621033326
14	0,032429828	0
15	2,842610658	0
16	2,494670148	2,407983845
17	1,705167193	1,699322142
18	2,175127598	0,982201997
19	1,983055445	0,695862801
20		

Figura 3 - Output del modello Produttore - Consumatore

In particolare, le formule (1) e (2) sono state inserite, con gli opportuni riferimenti, nelle colonne P e N.

Run della simulazione

La simulazione si esegue copiando le formule impostate per generare le realizzazioni delle variabili aleatorie tempo di soggiorno e tempo di attesa in coda (o di giacenza in magazzino) in tutte le 3200 righe (nelle formule si raccomanda di fare uso di riferimenti relativi, senza cioè il segno \$).

A questo punto, sebbene non sia stata ancora affrontata la fase di analisi dei dati, di fatto una simulazione Monte Carlo è stata completata. Dal momento che si è ricorso all'uso della funzione volatile CASUALE(), per effettuare un nuovo run della simulazione è sufficiente rieseguire il calcolo del foglio di lavoro premendo il tasto F9.

A valle di ciascun run di simulazione, mediante il *frame* riassuntivo proposto nella Figura 4, si può osservare come per ciascuna delle variabili aleatorie elencate, per la legge forte dei grandi numeri, la media calcolata su un numero grande di osservazioni (n) tenda al valore atteso riportato.

Modello Produttore - Consumatore		
Variabili aleatorie	Valore atteso	Media su 3200 osservazioni
Interarrivi (in input)	1	1,02723572
Tempo di servizio (in input)	0,9090909	0,91075472
Tempo di soggiorno (in output)	10	9,69459189
Tempo di attesa in coda (in output)	9,0909091	8,783837173

Figura 4 - Variabili aleatorie di I/O nel modello Produttore - Consumatore

Analisi dei risultati

La simulazione Monte Carlo si conclude con l'analisi dei risultati. Il primo passo consiste nel generare un istogramma (visualizzazione grafica di una serie di dati in funzione di una classe di intervalli possibili) in Excel.

Dalla semplice osservazione dell'istogramma si possono acquisire diverse informazioni. A titolo di esempio, l'istogramma in Figura 5 rivela che:

- la maggior parte dei prodotti ha un tempo di attesa in coda (o di giacenza in magazzino) $\neq 0$;
- l'incertezza è abbastanza ampia, variabile tra 0 e 23 unità di tempo;
- i tempi di attesa in coda (o di giacenza in magazzino) non sembrano distribuirsi secondo una distribuzione Normale; non si individuano valori anomali (*outlier*), raggruppamenti multipli, etc..

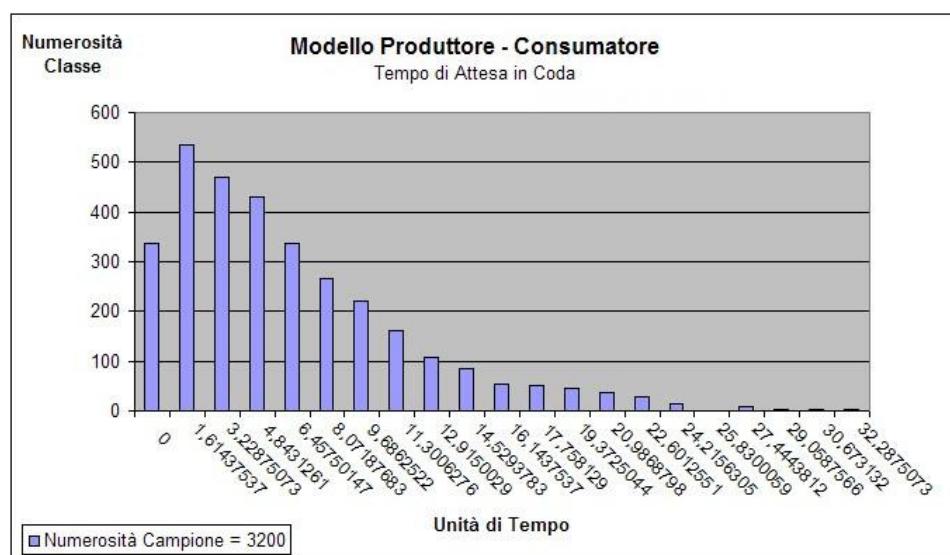


Figura 5 - Istogramma dei tempi di attesa in coda (o di giacenza in magazzino)

Creazione di un istogramma in Excel

Per creare un istogramma sotto Excel si ricorre all'uso della funzione FREQUENZA in base alla seguente procedura.

Creare un vettore di classi

La Figura 6 illustra un possibile vettore dinamico alla base della definizione di 20 classi (intervalli) di pari ampiezza.

	S	T
8	Min	0
9	Max	41,14822172
10	N° Intervalli	20
11		
12		Classi
13		0
14		2,057411086
15		4,114822172
16		6,172233258
17		8,229644344
18		10,28705543
19		12,34446652
20		14,4018776
21		16,45928869
22		18,51669977
23		20,57411086
24		22,63152195
25		24,68893303
26		26,74634412
27		28,8037552
28		30,86116629
29		32,91857738
30		34,97598846
31		37,03339955
32		39,09081063
33		41,14822172

Figura 6 – Vettore dinamico per la creazione di 20 classi

Per creare un vettore di detta dimensione si impostano le seguenti formule:

- $T8 = \text{MIN}(P9 : P3208)$
- $T9 = \text{MAX}(P9 : P3208)$
- $T10 = 20$
- $T13 = \$T\8
- $T14 = T13 + (\$T\$9 - \$T\$8)/\$T\10

Si copia poi la formula in T14 fino alla cella T33.

Impiego della funzione FREQUENZA di Excel

La Figura 7 mostra una porzione dei risultati ottenuti per la stima del tempo di attesa in coda (o di giacenza in magazzino) con la simulazione Monte Carlo. Il numero di realizzazioni sulle 3200 generate che ricadono nel generico intervallo $T_i - (T_{i+1} - 1)$ si ricava con l'impiego della funzione FREQUENZA di Excel.

	P	Q	R	S	T	U
6						
7						
8	Equazione di Lindley Vi					
9	0			Costruzione Iстограмма		
10	0,000550906			Min	0	
11	0			Max	24,84414454	
12	0,380269565			N° Intervalli	20	
13	0					
14	0					
15	0					
16	0,06294376					
17	0,497396759					
18	0,740672538					
19	1,12065026					
20	5,32714225					
21	7,074376482					
22	9,078672221					
23	9,269553903					
24	11,9924707					
25	12,18852022					
26	15,08618826					
27	15,60822231					
28	14,83546626					
29	18,3789766					
30	18,4484979					
31	18,88477441					
32	19,08867079					
33	15,898888884					

Figura 7 - Calcolo frequenza per classe dell'istogramma “tempi di attesa in coda”

In questo particolare caso, per immettere la funzione FREQUENZA(vettore_dati; vettore_frequenze) occorre:

- selezionare le celle U13:U33;
- inserire la formula = FREQUENZA(P : P; T13 : T33);
- Premere Ctrl+Shift+Invio¹.

Affinché sia possibile confrontare l'istogramma con la funzione di densità di probabilità, allora in questa sede si provvede anche a normalizzare l'istogramma in

¹ Questa modalità di inserimento, ossia la triplice combinazione di tasti, è necessaria per immettere la formula in forma di matrice. Se così non fosse, il risultato singolo sarebbe pari a 1. Si rimanda all'*Help* di Excel per maggiori dettagli circa l'uso della funzione FREQUENZA .

modo che l'area ad esso sottesa sia pari a 1. A valle dell'operazione di normalizzazione, l'asse delle y diviene la frequenza (ben diversa dalle frequenze nella colonna U).

Per normalizzare l'istogramma secondo la formula $Normalizzazione = (Frequenza/N^{\circ}Osservazioni)/(passo)$ ove per passo si intende l'ampiezza della singola classe è necessario:

- porre $V13 = (U13/\$T\$11)/(\$T\$14 - \$T\$13)$;
- copiare la cella V13 fino alla cella V33;
- premere F9 per ricalcolare il foglio Excel.

La Figura 8 si illustra un esempio della suddetta operazione di normalizzazione.

	S	T	U	V
6	Costruzione Iistogramma			
7				
8	Min		0	
9	Max		32,28750733	
10	N° Intervalli		20	
11	N° Osservazioni		3200	
12	Classi	Frequenze	Normalizzazione	
13	0	336	0,065040636	
14	1,614375366	535	0,103561726	
15	3,228750733	471	0,091173034	
16	4,843126099	430	0,083236528	
17	6,457501466	338	0,065427782	
18	8,071876832	267	0,051684077	
19	9,686252199	222	0,042973277	
20	11,30062756	161	0,031165305	
21	12,91500293	108	0,020905919	
22	14,5293783	84	0,016260159	
23	16,14375366	54	0,010452959	
24	17,75812903	51	0,009872239	
25	19,3725044	45	0,008710799	
26	20,98687976	37	0,007162213	
27	22,60125513	27	0,00522648	
28	24,2156305	14	0,002710026	
29	25,83000586	1	0,000193573	
30	27,44438123	8	0,001548587	
31	29,0587566	4	0,000774293	
32	30,67313196	3	0,00058072	
33	32,28750733	4	0,000774293	

Figura 8 - Calcoli per normalizzare l'istogramma “tempi di attesa in coda”

Creazione dell'istogramma

La creazione dell'istogramma si avvia cliccando sul pulsante *Creazione guidata Grafico* di Excel. Con la colonna delle “Classi” si forniscono le etichette lungo l'asse delle x, mentre i valori contenuti nella colonna V “Normalizzazione” si collocano lungo l'asse delle y. Le scelte di formattazione (i.e. dimensioni, colori, etc.) sono a gusto dell'utente.

In Figura 9 si riporta un esempio dell'istogramma normalizzato costruito secondo le modalità descritte nei paragrafi precedenti.

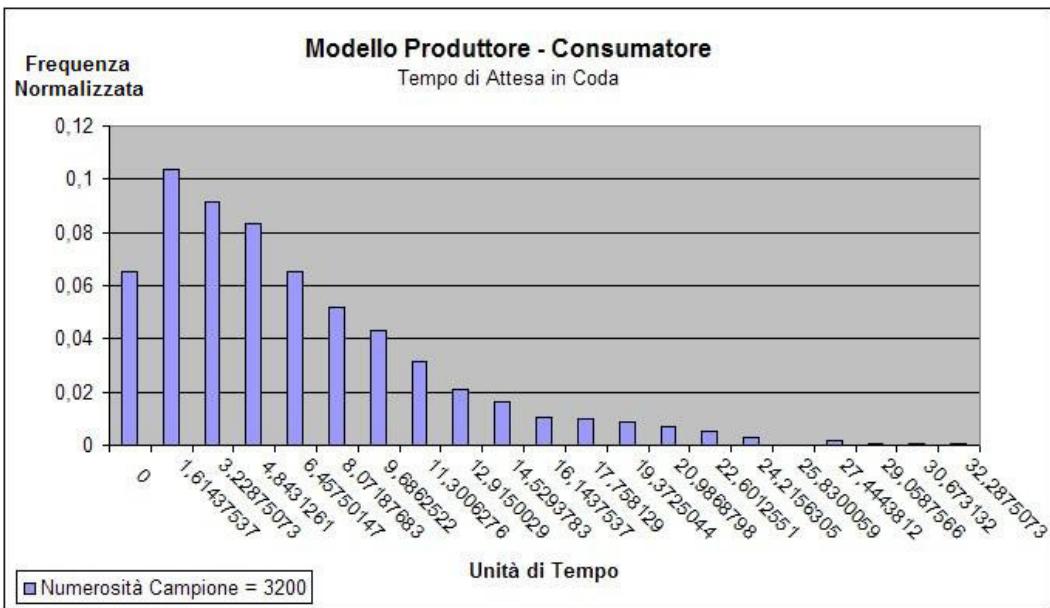


Figura 9 - Istogramma normalizzato dei tempi di attesa in coda

Creazione di una funzione di distribuzione di probabilità

L'utilità dell'istogramma è indiscussa, ma in molti casi non è sufficiente; basti pensare alla richiesta di stimare anche la probabilità che una certo indice di *performance* sia minore (o maggiore) di un certo valore di soglia o compreso in un determinato intervallo.

In questa sezione quindi, alle informazioni già leggibili dall'istogramma, si aggiungono quelle ottenibili dal grafico della relativa funzione di distribuzione di probabilità (FDP) come mostrato a seguire in Figura 10.

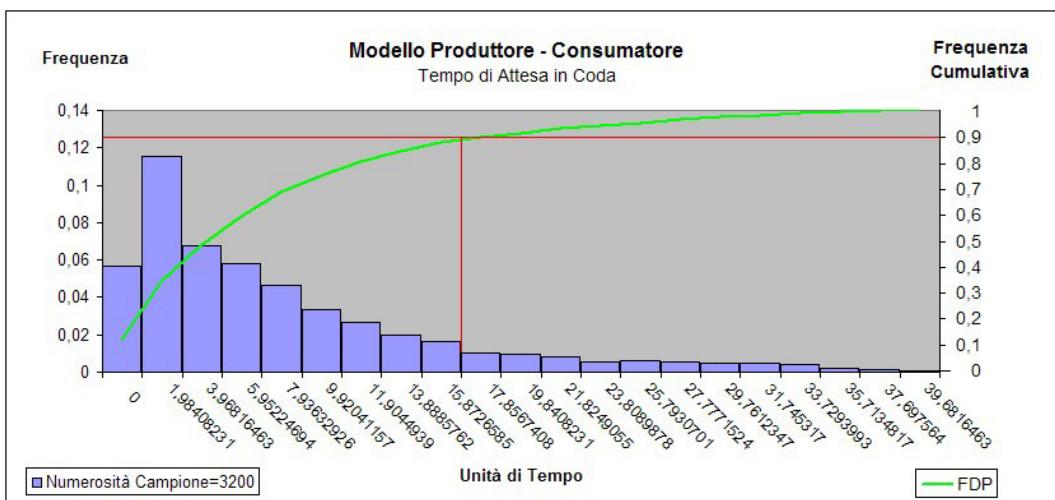


Figura 10 - Istogramma normalizzato e FDP dei tempi di attesa in coda

In sostanza, la ragione per cui si mostra l'istogramma normalizzato insieme alla FDP è per mostrare come la stima di una probabilità cumulativa sia semplicemente la percentuale di osservazioni nell'area sottesa alla curva a sinistra del punto di interesse.

Per esempio, dal grafico dei tempi di attesa in coda (o di giacenza in magazzino), tracciando una linea rossa verticale in corrispondenza di 15,87 unità di tempo, è possibile rilevare che la che con probabilità 0,9 (i.e. nel 90% dei casi) il tempo di attesa in coda è inferiore alla suddetta soglia di tempo.

La Figura 11 illustra come sia possibile stimare la FDP mediante il calcolo delle probabilità attraverso una somma cumulativa praticata sulla colonna Frequenze: è sufficiente dividere la somma cumulativa ottenuta sulle celle in U per il numero totale di osservazioni (T11).

	S	T	U	V	W
6	Costruzione Istogramma				
7					
8	Min	0			
9	Max	31,08549527			
10	N° Intervalli	20			
11	N° Osservazioni	3200			
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					

Figura 11 - Costruzione FDP dei tempi di attesa in coda

Numero di repliche e qualità dell'Intervallo di Confidenza

Nell'approccio introdotto, che ha per obiettivo la costruzione di un intervallo di confidenza per $\mu = E(X)$, dove X è l'indice di performance di un sistema calcolato lungo un orizzonte temporale finito, non si conosce a priori la semiampiezza di detto

intervallo definita come $\frac{z_{1-\alpha/2} * \sigma}{\sqrt{n}}$. Da una breve analisi, si osserva che tale termine,

noto anche come "errore" poiché rappresentativo della differenza tra la il vero valore di μ e la sua stima, dipende:

- dalla varianza di X (mediante σ), quantità non controllabile in quanto generalmente sconosciuta;
- dal numero n di repliche prodotte, grandezza controllabile in quanto fissata anteriormente alla produzione delle medesime repliche.

Ciò premesso, si introduce una metodologia mediante la quale è possibile determinare proprio il numero di repliche n di simulazione che consente di contenere l'errore entro un valore limite predefinito. La procedura in esame permette di ottenere una stima dell'indice di performance $\mu = E(X)$ tale da soddisfare uno dei seguenti criteri:

- **criterio dell'errore assoluto:** si definisce a priori un valore massimo assoluto dell'errore $\varepsilon > 0$ con l'obiettivo di ottenere una stima $\bar{X}(n)$ che soddisfi $|\bar{X}(n) - \mu| < \varepsilon$ con probabilità $1 - \alpha$ (i.e. lo stimatore deve essere entro ε dal vero valore μ con probabilità $1 - \alpha$; ad esempio, supponendo di voler stimare il tempo medio di attesa in coda davanti ad uno sportello bancario, se si vuole che lo stimatore sia entro 10 minuti dal vero valore con probabilità $1 - \alpha$, si fissa $\varepsilon = 10$).
- **criterio dell'errore relativo:** si definisce a priori un valore massimo relativo dell'errore $\varepsilon > 0$ con l'obiettivo di ottenere una stima $\bar{X}(n)$ che soddisfi $|\bar{X}(n) - \mu| / |\mu| < \varepsilon$ con probabilità $1 - \alpha$ (i.e. lo stimatore dovrebbe essere entro il $100\varepsilon\%$ del vero valore μ con probabilità $1 - \alpha$; se si pensa ancora al tempo medio di attesa in coda davanti ad uno sportello bancario e si vuole che lo stimatore sia entro il 5% del vero valore con probabilità $1 - \alpha$, allora si fissa $\varepsilon = 0,05$).

In base al criterio dell'errore assoluto:

$$|\bar{X}(n) - \mu| < \varepsilon \text{ con } \varepsilon \approx \frac{z_{1-\alpha/2} * \sigma}{\sqrt{n}}.$$

Segue che:

$$P\{-\varepsilon \leq \bar{X}(n) - \mu \leq \varepsilon\} \approx 1 - \alpha$$

dove lo stimatore indicato soddisfa il criterio dell'errore assoluto. Più precisamente sia

$N_a(\varepsilon) = \left\lceil \frac{z_{1-\alpha/2}^2 * \sigma^2}{\varepsilon^2} \right\rceil$ il valore intero più piccolo maggiore o uguale alla quantità $\frac{z_{1-\alpha/2}^2 * \sigma^2}{\varepsilon^2}$. Allora, per sostituzione:

$$P\{-\varepsilon \leq \bar{X}(N_a(\varepsilon)) - \mu \leq \varepsilon\} \approx 1 - \alpha \text{ per } \varepsilon \rightarrow 0.$$

Fissato un ε target, se fosse noto σ , ricavare n dalla formula dell'errore mediante pochi passaggi algebrici sarebbe immediato. In realtà σ non è noto, ma si può stimare tramite la varianza campionaria $S^2(n)$ che a sua volta dipende da n . In definitiva, il numero di repliche da produrre è dato dall'espressione:

$$N_a(\varepsilon) = \left\lceil \frac{z_{1-\alpha/2}^2 * S^2(n)}{\varepsilon^2} \right\rceil$$

Per determinare $N_a(\varepsilon)$ si ricorre all'uso di una procedura a due fasi: nella prima si calcola lo stimatore di σ^2 da n_0 *run di prova*; nella seconda si impiega la stima della varianza per calcolare la dimensione del campione $N_a(\varepsilon)$ dei *run di produzione* da realizzare.

Procedura a due fasi per intervalli di confidenza di precisione assoluta

1. selezionare n_0 , il numero di *run di prova* (si consiglia $n_0 \geq 10$), e l'errore di precisione assoluta ε (si consiglia un valore “piccolo” che, ovviamente, dipende dal contesto);
2. generare n_0 *run di prova* (indipendenti) dai quali ottenere le osservazioni campionarie X_1, X_2, \dots, X_{n_0} ;

3. calcolare $S_1^2(n_0) = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (X_i - \bar{X}(n_0))^2$, lo stimatore di $\sigma^2 = \text{Var}(X)$

ottenuto dai *run di prova*;

4. calcolare $N_a(\varepsilon) = \left\lceil \frac{z_{1-\alpha/2}^2 * S_1^2(n_0)}{\varepsilon^2} \right\rceil$;

5. generare $N_a(\varepsilon)$ *run di produzione* $X_{n_0+1}, X_{n_0+2}, \dots, X_{n_0+N_a(\varepsilon)}$ indipendenti tra loro ed indipendenti da X_1, X_2, \dots, X_{n_0} ;

6. calcolare $\tilde{X}(\varepsilon) = \frac{1}{N_a(\varepsilon)} \sum_{j=n_0+1}^{n_0+N_a(\varepsilon)} X_j$ e $\tilde{S}^2(\varepsilon) = \frac{1}{N_a(\varepsilon)-1} \sum_{j=n_0+1}^{n_0+N_a(\varepsilon)} (X_j - \tilde{X}(\varepsilon))^2$,

gli stimatori media campionaria e varianza campionaria, rispettivamente, ottenuti dai soli *run di produzione*;

7. allora $\left[\tilde{X}(\varepsilon) - z_{1-\alpha/2} \frac{\tilde{S}(\varepsilon)}{\sqrt{N_a(\varepsilon)}}, \tilde{X}(\varepsilon) + z_{1-\alpha/2} \frac{\tilde{S}(\varepsilon)}{\sqrt{N_a(\varepsilon)}} \right]$ è approssimativamente

un intervallo di confidenza al $100(1-\alpha)\%$ per ε , la cui semiampiezza dovrebbe essere approssimativamente ε .

Osservazioni

Se ε è piccolo allora $N_a(\varepsilon) \gg n_0$, quindi, l'esclusione delle prime n_0 osservazioni dal calcolo degli stimatori $\bar{X}(\varepsilon)$ e $\bar{S}^2(\varepsilon)$ non inficia la bontà della procedura.

Un vantaggio di tale approccio risiede nel fatto che gli stimatori calcolati allo step 6 sono ottenuti da realizzazioni $(X_{n_0+1}, X_{n_0+2}, \dots, X_{n_0+N_a(\varepsilon)})$ indipendenti ed identicamente distribuiti.

In modo analogo, si definisce una procedura a due fasi per costruire intervalli di confidenza di precisione relativa.

Procedura a due fasi per intervalli di confidenza di precisione relativa

- selezionare n_0 , il numero di *run di prova* (si consiglia $n_0 \geq 10$), e l'errore di precisione relativa ε (si consiglia $\varepsilon \leq 0,10$);

2. generare n_0 run di prova (indipendenti) dai quali ottenere le osservazioni campionarie X_1, X_2, \dots, X_{n_0} ;
3. calcolare $\hat{\bar{X}} = \frac{i}{n_0} \sum_{i=1}^{n_0} X_i$ e $\hat{S}_1^2(n_0) = \frac{1}{n_0-1} \sum_{i=1}^{n_0} (X_i - \bar{X}(n_0))^2$, rispettivamente la media campionaria e la varianza campionaria ottenute dai run di prova;
4. calcolare $N_r(\varepsilon) = \left\lceil \frac{z_{1-\alpha/2}^2 * S_1^2(n_0)}{\hat{\bar{X}}^2 \varepsilon^2} \right\rceil$;
5. generare $N_r(\varepsilon)$ run di produzione $X_{n_0+1}, X_{n_0+2}, \dots, X_{n_0+N_r(\varepsilon)}$ indipendenti tra loro ed indipendenti da X_1, X_2, \dots, X_{n_0} ;
6. calcolare $\tilde{X}(\varepsilon) = \frac{1}{N_r(\varepsilon)} \sum_{j=n_0+1}^{n_0+N_r(\varepsilon)} X_j$ e $\tilde{S}^2(\varepsilon) = \frac{1}{N_r(\varepsilon)-1} \sum_{j=n_0+1}^{n_0+N_r(\varepsilon)} (X_j - \tilde{X}(\varepsilon))^2$, ossia la media campionaria e la varianza campionaria, rispettivamente, ottenuti dai soli run di produzione;
7. allora $\left[\tilde{X}(\varepsilon) - z_{1-\alpha/2} \frac{\tilde{S}(\varepsilon)}{\sqrt{N_r(\varepsilon)}}, \tilde{X}(\varepsilon) + z_{1-\alpha/2} \frac{\tilde{S}(\varepsilon)}{\sqrt{N_r(\varepsilon)}} \right]$ è approssimativamente un intervallo di confidenza al $100(1-\alpha)\%$ per ε , la cui semiampiezza dovrebbe essere approssimativamente $|z_{1-\alpha/2}| \frac{\tilde{S}(\varepsilon)}{\sqrt{N_r(\varepsilon)}}$.

In alternativa alla procedura a due fasi, di seguito si propone un procedimento iterativo a passi variabili che restituisce risultati la cui bontà è dello stesso ordine dei risultati ottenuti con l'impiego della procedura iterativa a passo unitario indicata da Law e Kelton.

Procedura alternativa del “punto fisso” per intervalli di confidenza di precisione relativa

Step 0. $k = 0$, leggi n_k , ε_{target} da input;

Step 1. calcola ε_k :

Step 1.1. calcola $\bar{X}(n_k)$, $S^2(n_k)$

$$\text{Step 1.2.} \quad \varepsilon_k = \frac{z_{1-\alpha/2} * S(n_k)}{\sqrt{n_k} * \bar{X}(n_k)}$$

Step 2. confronta ε_k con ε_{target} :

se $\varepsilon_k > \varepsilon_{target}$

$$\text{Step 2.1.} \quad \text{calcola } n_{k+1} = \left\lceil \frac{z_{1-\alpha/2}^2 * S^2(n_k)}{\varepsilon_{target}^2 * \bar{X}^2(n_k)} \right\rceil$$

Step 2.2. aggiorna $k = k + 1$ e go to Step 1.

altrimenti STOP.

Si consideri una stazione di servizio M/M/5 con interarrivi distribuiti secondo una legge esponenziale di parametro $\lambda = 1$ e tempo di servizio distribuito secondo una legge esponenziale di parametro $\mu = 0,25$. Si determini un intervallo di confidenza di precisione relativa $\varepsilon = 0,09$ per il tempo di attesa in coda.

Secondo il metodo a due fasi, per ottenere un intervallo di confidenza rispondente a detti criteri, occorrono 121 repliche di simulazione (vedere Tabella 1).

$z_{1-\alpha/2}$	1,645		
$\varepsilon_{relativo}$	0,09		
n	$X(k)$	$S^2(k)$	ε_k
10	1,55	0,88	0,3153
121	1,91	1,52	0,1001

Tabella 1 – Risultati procedura a due fasi per calcolo intervallo di confidenza di precisione relativa

L’intervallo risultante è $[1,81 - 2,01]$. In verità, come si evince dalla Tabella 1, la procedura a due fasi si ferma restituendo un errore relativo effettivo superiore a quello target stabilito in fase di input.

Secondo la procedura iterativa a passo variabile, con 139 iterazioni e, quindi, repliche di simulazione (vedere Tabella 2) si ottiene una stima del tempo di attesa in coda che è

contenuto nell'intervallo $[1,78 - 1,96]$ con probabilità $1 - \alpha$ ed è entro il 9% del vero valore di detto indice di prestazione.

$z_{1-\alpha/2}$	1,645					
$\epsilon_{relativo}$	0,09					
n	k	$X(k)$	$S^2(k)$	ϵ_k	Risultato	Iterazione
10	0	1,55	0,88	0,3153	-	
121	1	1,88	1,47	0,0966	-	
137	2	1,87	1,47	0,0913	-	
139	3	1,87	1,45	0,0899	STOP! N° repliche = 139	

Tabella 2 – Risultati procedura iterativa a passo variabile per calcolo intervallo di confidenza di precisione relativa