

4. Stime & Test

Corso di Simulazione

Anno accademico 2008/09

Media campionaria

X_1, X_2, \dots, X_n v.c. indipendenti con distribuzione F , e:

$$\begin{aligned}E[X_i] &= \mu \\ \text{Var}[X_i] &= \sigma^2, i = 1, \dots, n\end{aligned}$$

Media campionaria: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i,$

\bar{X}_n è uno stimatore di μ .
È uno stimatore corretto:

$$E[\bar{X}_n] = \mu.$$

Risulta poi

$$\text{Var}[\bar{X}_n] = \frac{1}{n} \sigma^2$$

Uno stimatore corretto della varianza σ^2 è

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

Infatti:

$$\begin{aligned}(n-1)E[S_n^2] &= E\left[\sum_{i=1}^n (X_i - \bar{X}_n)^2\right] \\ &= \sum_{i=1}^n E[X_i^2] - nE[\bar{X}_n^2]\end{aligned}$$

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

Dalla

$$E[X_i^2] = \text{Var}[X_i] + E[X_i]^2,$$

Sostituendo si ha:

$$\begin{aligned}(n-1)E[S_n^2] &= \sum_{i=1}^n (\text{Var}[X_i] + E[X_i]^2) - n(\text{Var}[\bar{X}_n] + E[\bar{X}_n]^2) \\ &= n\sigma^2 + n\mu^2 - n\frac{\sigma^2}{n} - n\mu^2 \\ &= (n-1)\sigma^2.\end{aligned}$$

Quanto è accurata la stima di μ fornita dalla media campionaria \bar{X}_n ?

Per il teorema del limite centrale, per n opportunamente grande, è

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim N(0, 1),$$

La stessa cosa vale se sostituiamo σ , che non conosciamo, con la sua stima S_n .

Se Z è una v.c. normale standard, per ogni $\alpha \in (0, 1)$, indichiamo con z_α il valore per cui è

$$P(Z > z_\alpha) = \alpha.$$

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$



$$P(-z_{\alpha/2} < \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} < z_{\alpha/2}) \approx 1 - \alpha,$$



$$P(-z_{\alpha/2} < \sqrt{n} \frac{\mu - \bar{X}_n}{S_n} < z_{\alpha/2}) \approx 1 - \alpha,$$



$$P(\bar{X}_n - z_{\alpha/2} \frac{S_n}{\sqrt{n}} < \mu < \bar{X}_n + z_{\alpha/2} \frac{S_n}{\sqrt{n}}) \approx 1 - \alpha,$$

Con probabilità $1 - \alpha$, il valore μ incognito si trova nell'intervallo

$$\bar{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

Ad esempio: $P(Z < 1.96) = 0.975$

Quindi la probabilità che la media campionaria \bar{X}_n differisca da μ di più di $1.96 \frac{S_n}{\sqrt{n}}$ è circa 0.05.

Calcolo ricorsivo della media

Si voglia stimare la media della v.c. X in modo che la probabilità che di fare un errore maggiore di d sia pari ad α .

Se c è il reale per cui risulta $P(Z < c) = 1 - \alpha/2$, si generano successive realizzazioni di X fino ad averne un numero k tale che risulti $c \frac{S_k}{\sqrt{k}} < d$. (È comunque opportuno che il valore di k non sia inferiore a 30).

Per realizzare in modo efficiente il calcolo è opportuno disporre di formule ricorsive per il calcolo di \bar{X}_k e di S_k^2 .

$$\begin{aligned}\bar{X}_{k+1} &= \frac{1}{k+1} \sum_{j=1}^{k+1} X_j \\ &= \bar{X}_k - \bar{X}_k + \frac{k\bar{X}_k + X_{k+1}}{k+1} \\ &= \bar{X}_k + \frac{X_{k+1} - \bar{X}_k}{k+1}\end{aligned}$$

Calcolo ricorsivo della varianza (1)

$$\begin{aligned} S_{k+1}^2 &= \sum_{j=1}^{k+1} \frac{(X_j - \bar{X}_{k+1})^2}{k} \\ &= \sum_{j=1}^k \frac{(X_j - \bar{X}_k + \bar{X}_k - \bar{X}_{k+1})^2}{k} + \frac{(X_{k+1} - \bar{X}_{k+1})^2}{k} \\ &= \sum_{j=1}^k \frac{(X_j - \bar{X}_k)^2 + (\bar{X}_k - \bar{X}_{k+1})^2 + 2(X_j - \bar{X}_k)(\bar{X}_k - \bar{X}_{k+1})}{k} \\ &\quad + \frac{(X_{k+1} - \bar{X}_{k+1})^2}{k} \\ &= \left(1 - \frac{1}{k}\right) S_k^2 + (\bar{X}_k - \bar{X}_{k+1})^2 + \frac{(X_{k+1} - \bar{X}_{k+1})^2}{k}, \end{aligned}$$

$$(\sum_{j=1}^k (X_j - \bar{X}_k) = 0)$$

Calcolo ricorsivo della varianza (2)

$$S_{k+1}^2 = \left(1 - \frac{1}{k}\right)S_k^2 + (\bar{X}_k - \bar{X}_{k+1})^2 + \frac{(X_{k+1} - \bar{X}_{k+1})^2}{k}$$

Essendo

$$\begin{aligned}X_{k+1} - \bar{X}_{k+1} &= \frac{(k+1)X_{k+1} - \sum_{j=1}^{k+1} X_j}{k+1} \\ &= \frac{kX_{k+1} - \sum_{j=1}^k X_j}{k+1} \\ &= \frac{kX_{k+1} - \sum_{j=1}^k [(k+1)X_j - kX_j]}{k+1} \\ &= k(\bar{X}_{k+1} - \bar{X}_k)\end{aligned}$$

si ha

$$S_{k+1}^2 = \left(1 - \frac{1}{k}\right)S_k^2 + (1+k)(\bar{X}_{k+1} - \bar{X}_k)^2$$

X_1, X_2, \dots, X_n osservazioni indipendenti della variabile casuale X , con funzione di densità $f_\theta(x)$, dove con θ si è indicato il parametro che caratterizza la distribuzione.

Si vuole stimare θ .

Funzione di verosimiglianza:

$$L(\theta) = f_\theta(X_1)f_\theta(X_2)\dots f_\theta(X_n)$$

Il *metodo della massima verosimiglianza* consiste nello scegliere come stimatore il valore di θ che massimizza $L(\theta)$.

Interpretazione di $L(\theta)$

Nel caso di distribuzioni discrete, è la probabilità di avere ottenuto il campione X_1, X_2, \dots, X_n .

Nel caso di distribuzioni continue la probabilità che l'estrazione casuale di un elemento da una popolazione con la distribuzione data sia un valore compreso in un'intorno di raggio $\varepsilon/2$ di X_i è approssimativamente $\varepsilon f_\theta(X_i)$, con un'approssimazione tanto più accurata quanto più piccolo è ε .

Pertanto $L(\theta)$ è approssimativamente proporzionale alla probabilità dell'estrazione di un campione di n elementi, Y_1, Y_2, \dots, Y_n , con $Y_i \in [X_i - \varepsilon, X_i + \varepsilon]$, $i = 1, 2, \dots, n$, e con ε opportunamente piccolo.

Massima verosimiglianza: esempio 1

Si voglia stimare con il metodo della massima verosimiglianza il parametro λ di una esponenziale. Si ha

$$\begin{aligned}L(\lambda) &= (\lambda e^{-\lambda X_1})(\lambda e^{-\lambda X_2}) \dots (\lambda e^{-\lambda X_n}) \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n X_i} \\ &= \lambda^n e^{-\lambda n \bar{X}_n}\end{aligned}$$

$$\frac{dL}{d\lambda} = n\lambda^{n-1} e^{-\lambda n \bar{X}_n} - \lambda^n n \bar{X}_n e^{-\lambda n \bar{X}_n} = 0$$



$$\lambda = \frac{1}{\bar{X}_n}$$

Massima verosimiglianza: esempio 2

X uniforme in $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$

$$f(x; \theta) = I_{[\theta - \frac{1}{2}, \theta + \frac{1}{2}]}(x)$$

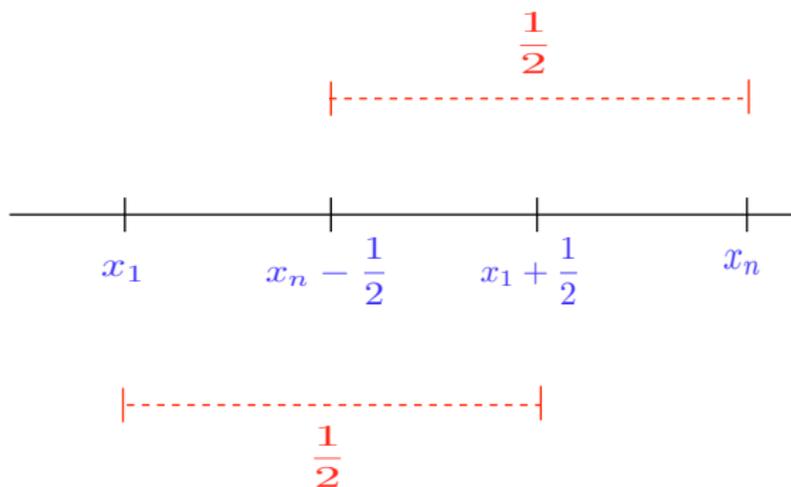
$$x_1 \leq x_2 \leq \dots \leq x_n$$

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta) = \prod_{i=1}^n I_{[\theta - \frac{1}{2}, \theta + \frac{1}{2}]}(x_i)$$

$L(\theta; x_1, x_2, \dots, x_n)$ è una funzione che vale 1 se è $x_i \in [\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ e 0 altrimenti. Quindi:

$$L(\theta; x_1, x_2, \dots, x_n) = I_{[x_n - \frac{1}{2}, x_1 + \frac{1}{2}]}(\theta)$$

Massima verosimiglianza: esempio 2



L assume il massimo valore per $\theta \in [x_n - \frac{1}{2}, x_1 + \frac{1}{2}]$

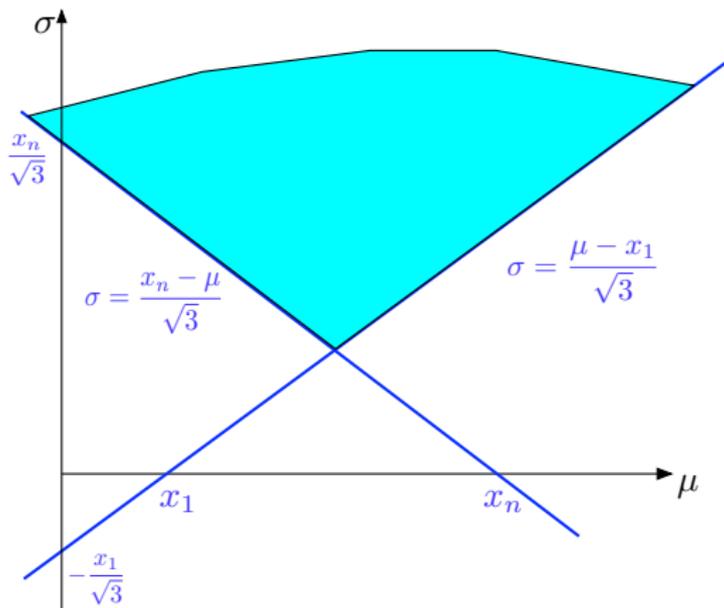
Assumiamo che X sia uniforme, ma non conosciamo né gli estremi né l'ampiezza dell'intervallo.

$$\sigma^2 = \frac{(b-a)^2}{12} \Rightarrow b-a = 2\sqrt{3}\sigma$$

$$f(x; \mu, \sigma) = \frac{1}{2\sqrt{3}\sigma} I_{[\mu-\sqrt{3}\sigma, \mu+\sqrt{3}\sigma]}(x)$$

Massima verosimiglianza: esempio 3

$$\begin{aligned}L(\mu, \sigma; x_1, x_2, \dots, x_n) &= \prod_{i=1}^n \frac{1}{2\sqrt{3}\sigma} I_{[\mu-\sqrt{3}\sigma, \mu+\sqrt{3}\sigma]}(x_i) \\&= \left(\frac{1}{2\sqrt{3}\sigma}\right)^n I_{[\mu-\sqrt{3}\sigma, x_n]}(x_1) I_{[x_1, \mu+\sqrt{3}\sigma]}(x_n) \\&= \left(\frac{1}{2\sqrt{3}\sigma}\right)^n I_{\left[\frac{\mu-x_1}{\sqrt{3}}, +\infty\right]}(\sigma) I_{\left[\frac{x_n-\mu}{\sqrt{3}}, +\infty\right]}(\sigma)\end{aligned}$$



La funzione $L(\mu, \sigma)$ vale $\left(\frac{1}{2\sqrt{3}\sigma}\right)^n$ nell'area che si trova sopra le due rette e 0 altrove. Il massimo si ha allora quando σ è minimo, cioè in corrispondenza dell'incrocio fra le rette:

$$\hat{\mu} = \frac{x_n + x_1}{2} \quad \hat{\sigma} = \frac{x_n - x_1}{2\sqrt{3}}$$

- X_1, X_2, \dots, X_n : v.c. indipendenti con distribuzione F
- $\theta(F)$: parametro di F da stimare
- $g(X_1, X_2, \dots, X_n)$: stimatore usato

Definiamo l'*Errore Quadratico Medio*:

$$EQM(F, g) = E_F[(g(X_1, X_2, \dots, X_n) - \theta(F))^2].$$

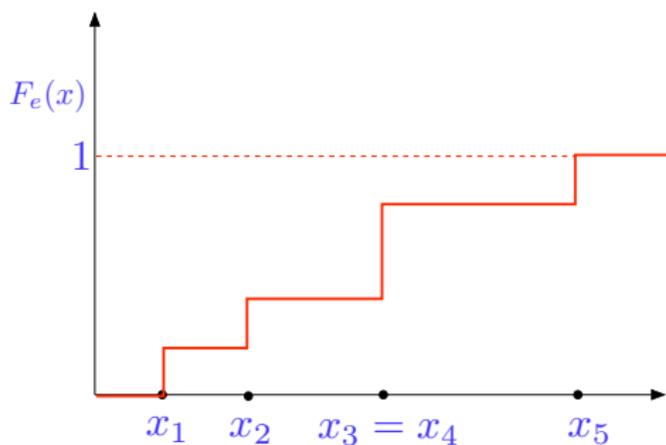
Come stimare $EQM(F, g)$, considerato che F non è nota?

Stima di $EQM(F, g)$

Sia (x_1, x_2, \dots, x_n) una realizzazione delle v.c. X_1, X_2, \dots, X_n

Definiamo la variabile casuale discreta X_e che assume i valori x_1, x_2, \dots, x_n con funzione di distribuzione:

$$F_e(x) = \frac{|\{i : x_i \leq x\}|}{n}$$



$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

$$F_e(x) = \begin{cases} 0, & \text{se } x < x_{(1)} \\ \frac{j}{n}, & \text{se } x_{(j)} \leq x < x_{(j+1)} \\ 1, & \text{se } x_{(n)} \leq x \end{cases}$$

F_e è una *stima empirica* di F ; per la legge dei grandi numeri, è

$$F_e(x) \xrightarrow[n \rightarrow \infty]{} F(x).$$

$\theta(F_e)$ è una approssimazione di $\theta(F)$ e $EQM(F_e, g)$ è una approssimazione di $EQM(F, g)$:

$$EQM(F_e, g) = E_{F_e}[(g(X_1, X_2, \dots, X_n) - \theta(F_e))^2]$$

L'approssimazione è tanto più buona quanto più è grande n

In pratica però il calcolo di $EQM(F_e)$ può risultare notevolmente oneroso essendo

$$EQM(F_e, g) = \sum_{y \in \{x_1, x_2, \dots, x_n\}^n} \frac{(g(y) - \theta(F_e))^2}{n^n},$$

In pratica vengono generati k vettori $y \in \{x_1, x_2, \dots, x_n\}^n$,
 y_1, y_2, \dots, y_k , e si pone

$$EQM(F_e) \simeq \sum_{i=1}^k \frac{(g(y_i) - \theta(F_e))^2}{k}.$$

Infatti le $(g(y_i) - \theta(F_e))^2$ possono essere considerate come valori
assunti da variabili casuali indipendenti con media $EQM(F_e)$, e
quindi la loro media è una stima corretta di $EQM(F_e)$

X v.c. discreta che assume valori $1, 2, \dots, k$
 x_1, x_2, \dots, x_n realizzazioni di X
Ipotesi H_0 (ipotesi nulla):

$$H_0 : P[X = i] = p_i, i = 1, \dots, k$$

dove p_1, p_2, \dots, p_k sono valori dati con somma 1.

$$N_i = |\{j : x_j = i\}|, i = 1, \dots, k$$

Sotto l'ipotesi H_0 , N_i ha distribuzione binomiale con parametri n e p_i , per ogni i , e quindi ha media np_i .

$$T = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}$$

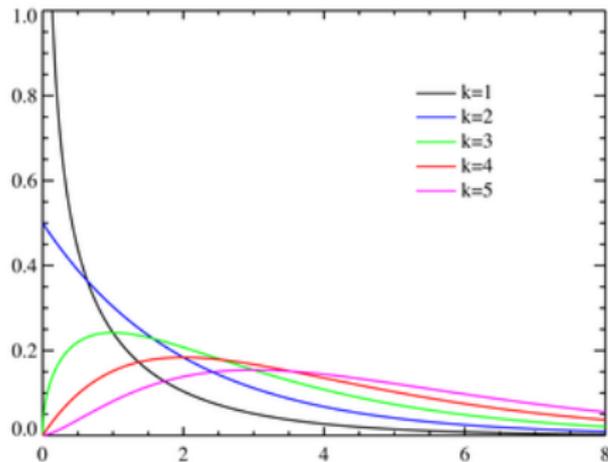
Più grande è T meno è probabile che l'ipotesi H_0 sia corretta.
Per n grande, T ha approssimativamente una distribuzione *Chi-Quadro* con $k - 1$ gradi di libertà

$$P_{H_0}[T \geq t] \cong P[\chi_{k-1}^2 \geq t]$$

$$f_X(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} \left(\frac{1}{2}\right)^{\frac{k}{2}} x^{\frac{k}{2}-1} e^{-\frac{1}{2}x} I_{(0,\infty)}(x)$$

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx, t > 0 \quad [t \text{ intero} \rightarrow \Gamma(t) = (t-1)!]$$

$$E[X] = k$$
$$Var[X] = 2k$$



Distribuzione Gamma(α, β)

La distribuzione Chi-Quadro è un caso particolare della distribuzione generale Gamma(α, β):

$$f_X(x) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} I_{(0,\infty)}(x)$$

La Chi-Quadro con k gradi di libertà è una Gamma($k/2, 2$).

Per α intero la Gamma(α, β) diventa la distribuzione Gamma già vista con parametri $n = \alpha$ e $\lambda = 1/\beta$.

Se è t il valore assunto da T , $P[\chi_{k-1}^2 \geq t]$ fornisce la probabilità di errore nel caso che si decida di scartare l'ipotesi H_0 .

Valori che vengono considerati ragionevoli per rigettare l'ipotesi sono $P[\chi_{k-1}^2 \geq t] = 0.05$ (oppure più conservativamente 0.01).

Una più accurata approssimazione del valore $P_{H_0}[T \geq t]$ può essere ottenuta per mezzo di una simulazione. Si generano a questo scopo le variabili casuali $T^{(1)}, T^{(2)}, \dots, T^{(r)}$, ciascuna con la distribuzione di T sotto l'ipotesi H_0 , e si pone

$$P_{H_0}[T \geq t] \cong \frac{|\{j : T^{(j)} \geq t\}|}{r}$$

Al crescere di r migliora la bontà dell'approssimazione.

X_1, X_2, \dots, X_n variabili indipendenti, identicamente distribuite
 H_0 : che abbiano una comune distribuzione continua F data.
Possiamo ricondurci al caso discreto suddividendo l'insieme dei possibili valori assunti dalle X_i in k intervalli distinti

$$(-\infty, x_1), (x_1, x_2), \dots, (x_{k-2}, x_{k-1}), (x_{k-1}, +\infty).$$

Si considerano le v.c. discrete X_i^d con $X_i^d = i$ se X_i si trova nell'intervallo (x_{i-1}, x_i) , e l'ipotesi H_0 diviene
 $P[X_i^d = i] = F(x_i) - F(x_{i-1}), i = 1, \dots, k.$

Test di Kolmogorov-Smirnov per distribuzioni continue

X_1, X_2, \dots, X_n variabili indipendenti, identicamente distribuite

H_0 : che abbiano una comune distribuzione continua F data.

F_e approssimazione della F :

$$F_e(x) = \frac{|\{i : X_i \leq x\}|}{n}$$

Se l'ipotesi H_0 è corretta allora $F_e(x)$ è una buona approssimazione di $F(x)$. Una misura dello scostamento è

$$D = \max_x |F_e(x) - F(x)|.$$

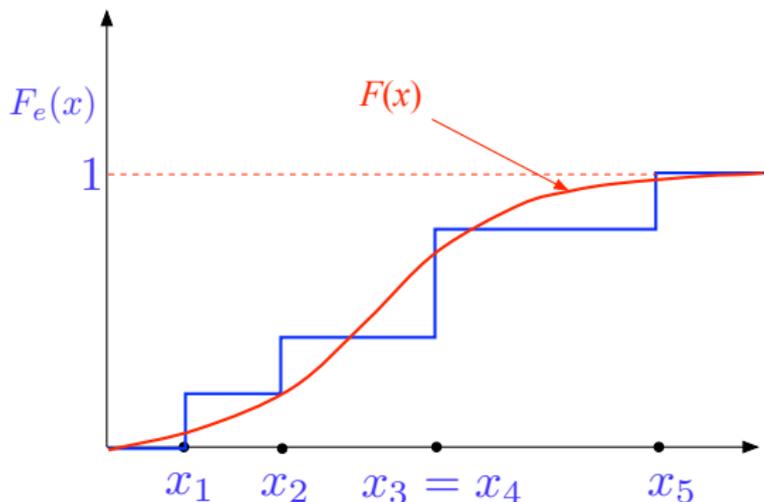
Dati i valori osservati x_1, \dots, x_n di X_1, \dots, X_n , ricaviamo il valore osservato d di D .

$$d = \text{Max}_x \{F_e(x) - F(x), F(x) - F_e(x)\}$$

$P_F(D \geq d)$: probabilità di fare un errore se decidiamo di rigettare H_0 .

$$\text{Max}_x \{F_e(x) - F(x)\} = \text{Max} \left\{ \frac{j}{n} - F(x_{(j)}) : j = 1 \dots n \right\},$$

$$\text{Max}_x \{F(x) - F_e(x)\} = \text{Max} \left\{ F(x_{(j)}) - \frac{j-1}{n} : j = 1 \dots n \right\},$$



$$d = \max \left\{ \frac{j}{n} - F(x_{(j)}), F(x_{(j)}) - \frac{j-1}{n} : j = 1, \dots, n \right\}.$$

$$\begin{aligned}
 P_F[D \geq d] &= P \left[\text{Max}_x \left| \frac{|\{i : X_i \leq x\}|}{n} - F(x) \right| \geq d \right] \\
 &= P \left[\text{Max}_x \left| \frac{|\{i : F(X_i) \leq F(x)\}|}{n} - F(x) \right| \geq d \right] \\
 &= P \left[\text{Max}_x \left| \frac{|\{i : U_i \leq F(x)\}|}{n} - F(x) \right| \geq d \right]
 \end{aligned}$$

U_i , sono v. c. indipendenti uniformi in (0,1).

(Se X è una v.c. con distribuzione continua F , allora $F(X)$ è una v.c. uniforme in (0,1). Infatti, ponendo $Y = F(X)$ si ha $P[Y \leq y] = P[X \leq F^{-1}(y)] = y$.)

$$P[D \geq d] = P \left[\text{Max}_{0 \leq y \leq 1} \left| \frac{|\{i : U_i \leq y\}|}{n} - y \right| \geq d \right].$$

Stima di $P[D \geq d]$:

① Si generano u_1, u_2, \dots, u_n , uniformi in $(0, 1)$,

② Si calcola $\text{Max}_{0 \leq y \leq 1} \left| \frac{|\{i: u_i \leq y\}|}{n} - y \right| =$
 $\text{Max} \left\{ \frac{j}{n} - u_{(j)}, u_{(j)} - \frac{j-1}{n} : j = 1, \dots, n \right\}$.

Si ripete più volte e si prende come valore per $P[D \geq d]$ la proporzione di volte in cui il valore trovato risulta $\geq d$.

Se $P[D \geq d]$ è sufficientemente basso (es. 0.05) l'ipotesi viene rigettata, altrimenti viene accettata.

Il test della somma dei ranghi

Y_1, Y_2, \dots, Y_m : valori osservati di una variabile V del sistema studiato.

(Le Y_i possono essere considerate come v.c. identiche e indipendenti)

X_1, X_2, \dots, X_n : i valori forniti per V dalla simulazione in n run.

(Anche le X_i saranno v.c. identicamente distribuite e indipendenti, con distribuzione F , in generale non nota)

Ipotesi H_0 : le Y_i abbiano la stessa distribuzione delle X_i , cioè

$$X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_m$$

sono v.c. identicamente distribuite e indipendenti.

Ordiniamo le $X_1, \dots, X_n, Y_1, \dots, Y_m$ in ordine crescente di valore e per $i = 1, \dots, n$ sia R_i il rango di X_i , cioè la sua posizione nella lista ordinata.

Ad esempio se le sequenze sono:

$X : 20, 15, 38, 40, 35, 31$

$Y : 25, 30, 29, 34.$

Si ha

$R_1 = 2, R_2 = 1, R_3 = 9, R_4 = 10, R_5 = 8$ e $R_6 = 6.$

$$R = \sum_{i=1}^n R_i$$

è un indicatore di plausibilità dell'ipotesi ($R = 36$ nell'esempio precedente)

Chiaramente un valore troppo piccolo o troppo grande di R falsificherebbe con alta probabilità l'ipotesi H_0 . Supponendo di ritenere accettabile una probabilità α (ad es. 0.05) di sbagliare nel rigettare l'ipotesi, rigetteremo H_0 se risulta

$$2 \min \{ P_{H_0} [R \leq r], P_{H_0} [R \geq r] \} \leq \alpha.$$

Come determinare la distribuzione di R ?

Ponendo $F_{n,m}(r) = P_{H_0}[R \leq r]$, vale la seguente equazione ricorsiva

$$F_{n,m}(r) = \frac{n}{n+m} F_{n-1,m}(r-n-m) + \frac{m}{n+m} F_{n,m-1}(r),$$

con

$$F_{1,0}(r) = \begin{cases} 0, & \text{se } r < 1 \\ 1, & \text{se } r \geq 1 \end{cases}$$
$$F_{0,1}(r) = \begin{cases} 0, & \text{se } r < 0 \\ 1, & \text{se } r \geq 0 \end{cases}$$

Si ha allora un sistema di equazioni ricorsive che consente di calcolare $F_{n,m}(r)$ e quindi la distribuzione di R .

In pratica il calcolo di $F_{n,m}(r)$ utilizzando la formula ricorsiva risulta molto oneroso. Si ricorre allora ad una approssimazione di $F_{n,m}(r)$. È possibile dimostrare che

$$R = \frac{n(n+m+1)}{2} / \sqrt{\frac{nm(n+m+1)}{12}}$$

è, approssimativamente, per n ed m grandi, una normale standard, $N(0, 1)$. Pertanto è

$$P[R \leq r] \cong P[Z \leq r^*]$$

con Z una v.c. $N(0, 1)$ e

$$r^* = \frac{r - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

Arrivi ad uno sportello, pacchetti in arrivo ad un server, chiamate in arrivo ad un call center, ...

- λ : numero medio di arrivi nell'unità di tempo,
- $N(t)$: il numero di arrivi nell'intervallo temporale $[0, t]$.

Si parla di processi di Poisson, che possono essere sia stazionari che non stazionari.

Processo di Poisson stazionario

- 1 arriva un individuo alla volta, cioè non ci sono arrivi di gruppi di individui;
- 2 il numero di arrivi nell'intervallo $(t, t + s]$, $N(t + s) - N(t)$, è indipendente da $N(u)$, per ogni $u \in [0, t]$;
- 3 la distribuzione di $N(t + s) - N(t)$ è indipendente da t per ogni $(t, s) \geq 0$.

$$P[N(t + s) - N(t) = k] = \frac{e^{-\lambda s} (\lambda s)^k}{k!}, \quad k = 0, 1, 2, \dots, \quad t, s \geq 0,$$

In molti casi reali il numero medio di arrivi nell'unità di tempo non è indipendente dal tempo. Sostituiamo allora alla costante λ una funzione del tempo $\lambda(t)$.

- 1 arriva un individuo alla volta, cioè non ci sono arrivi di gruppi di individui;
- 2 il numero di arrivi nell'intervallo $(t, t + s]$, $N(t + s) - N(t)$, è indipendente da $N(u)$, per ogni $u \in [0, t]$.

$$P[N(t+s) - N(t) = k] = \frac{e^{-b(t,s)} b(t,s)^k}{k!}, \quad k = 0, 1, 2, \dots, \quad t, s \geq 0.$$

- $\Lambda(t) = E[N(t)]$
- $b(t, s) = \Lambda(t + s) - \Lambda(t)$

Se $\Lambda(t)$ è differenziabile si ha:

- $\lambda(t) = \frac{d\Lambda(t)}{dt}$
- $b(t, s) = \Lambda(t + s) - \Lambda(t) = \int_t^{t+s} \lambda(y) dy$

Si abbiano i dati relativi agli arrivi nell'intervallo T di n giorni. Dividiamo l'intervallo T in p intervallini di uguale ampiezza Δ , $[t_1, t_2], [t_2, t_3], \dots, [t_p, t_{p+1}]$.

Sia x_{ij} il numero di arrivi nell'intervallo i del giorno j . Possiamo allora calcolare la media del numero di arrivi in ciascuno degli intervallini:

$$\bar{x}_i = \frac{\sum_j x_{ij}}{n},$$

e di conseguenza costruire una approssimazione della funzione $\lambda(t)$:

$$\bar{\lambda}(t) = \frac{\bar{x}_i}{\Delta}, \quad t \in [t_i, t_{i+1}], \quad i = 1, 2, \dots, p.$$

Definiamo poi, per ogni gruppo i , la variabile casuale discreta, B_i , che può assumere valori $1, 2, \dots$. Tale variabile definisce la cardinalità del gruppo.

Il numero di arrivi individuali entro il tempo t è allora dato dalla:

$$X(t) = \sum_{i=1}^{N(t)} B_i, \quad t \geq 0.$$