metriche

M. Arrigoni Neri & *P. Borghese*

indice

- prestazioni: grandezze - indici statistici
- modalità di misura
- curva di prestazioni (aspetto grafico qualitativo)
- processi (introduzione matematica)

prestazioni
grandezze - indici statistici

qualità di un impianto

- dipende da molti attributi o fattori che devono essere tenuti presenti nella progettazione e gestione:
  - *prestazioni* (performance)
    - efficienza: rendimento (situazione di massima capacità produttiva)
    - efficacia: capacità di produrre pienamente l'effetto voluto
  - *affidabilità* (reliability)
  - *disponibilità* (availability)
  - usabilità (usability)
  - flessibilità (flexibility)
  - ampliabilità o scalabilità (scalability)
  - adeguatezza (adequacy)
  - evolubilità (evolubility)

- QoS (Qualità del Servizio) = soddisfazione dell'utente

prestazioni

- effetto utile prodotto da un dispositivo per il quale quest'ultimo è stato progettato e costruito *(definizione tratta dal dizionario)*
  - es.: prestazioni di un mulino – quantità di farina prodotta nell'unità di tempo

- insieme degli indici tipici del comportamento di un sistema, di una sua parte (sottosistema) o di un suo componente, relativi a un dato intervallo di tempo e con un determinato carico di lavoro *(uso del termine nell'analisi degli impianti)*
  - gli indici possono essere locali o globali, cioè relativi ad un componente o all'intero sistema
  - generalmente si ipotizza di osservare / studiare il sistema in condizioni stazionarie cioè con "piccola" variabilità

grandezze di prestazione

variabili	tipo di metrica
throughput (capacità)	num./ tempo
lunghezza della coda	num. di elementi
tempo di risposta	unità di tempo
tempo di attesa	unità di tempo
memoria	unità di memoria
utilizzazione	% di tempo di occupazione
disponibilità	% del tempo di funzionamento
affidabilità	prob(prestazione a un dato tempo)
safety	prob(assenza di eventi catastrofici)
security	prob(assenza di accessi indesiderati)

- spazio / consumi / calore dissipato ecc. → green IT

parametri di principale interesse

- per gli utenti
  - tempi di risposta (tempo complessivo atteso dall'utente)
  - ritardi per attese di servizio (accodamenti)
  - numero di utenti in attesa (lunghezza delle code)
  - numero di utenti in un componente o nell'intero nel sistema
  - probabilità di dover attendere
  - probabilità di blocco
  - disponibilità / affidabilità delle applicazioni / servizi / componenti
- per i fornitori di servizio
  - utilizzo delle stazioni di servizio
  - utilizzo delle linee di attesa (da cui: dimensione dei buffer)
  - tassi di servizio di clienti/transazioni (throughput)
  - costi
  - qualità del servizio (grado di soddisfazione dell'utente)
    - disponibilità / affidabilità delle applicazioni / servizi / componenti

grandezze di prestazione (cont)

- misure dirette – metriche:
  - tempi (istanti - durate)
  - conteggi di eventi / utenti

cosa si misura

- statistiche:
  - valori medi, momenti, percentili

cosa si calcola

- strumenti di misura e controllo:
  - monitor resource manager

- procedure operative
  - problemi di sincronizzazione
  - integrazione di strumenti di misura
  - caratterizzazione del carico in classi
  - livelli di dettaglio
  - accorgimenti: per esempio, le medie hanno senso se operate in periodi stazionari

come dove quando

digressione sulla media

- *media* - generalmente viene introdotta mediante definizione
  - - aritmetica
  - - geometrica
  - - armonica
  - - …..

$$\frac{\sum x_i}{N} \qquad \sqrt[N]{\prod x_i} \qquad \frac{N}{\sum \frac{1}{x_i}}$$

ogni media ha diritto di cittadinanza in statistica ma impiego diverso

- *media pesata* con pesi p_i
  - - aritmetica
  - - geometrica
  - - armonica
  - - …..

$$\sum x_i \cdot p_i \qquad \sum p_i = 1$$

$$\prod x_i^{p_i} \qquad \frac{1}{\sum \frac{p_i}{x_i}}$$

se $p_i = 1/N$ (per ogni i) si ottengono le medie usuali

digressione sulla media: una definizione più generale

- si abbiano *N* grandezze omogenee $X_1, X_2, \ldots, X_N$
- interessa considerare:
  - la funzione $f(X_1, X_2, \ldots X_N)$
    - simmetrica nelle variabili, nel senso che non cambia al variare dell'ordine di queste
  - il valore X
  - tale che, agli effetti del calcolo di *f*, tutto va come se le *N* variabili X_i avessero tutte il valore *X*

    $f(X_1, X_2, \ldots X_N) = f(X, X, \ldots X)$

- *X* è detta: *media* agli effetti del calcolo di *f*
  - per esempio: la media aritmetica conserva la funzione: $f = \Sigma X_i$

    (se X_j sono tempi di servizio si conserva la funzione utilizzo)

digressione sulla media: esempio di calcolo della media

- un'automobile percorre 225 Km viaggiando a *60 Km/h* per i primi 120 Km e a *105 Km/h* per i restanti 105 Km
- ci chiediamo quale è la velocità (costante) che dovrebbe avere per percorrere *la stessa strada nello stesso tempo*

- $f(V_1,V_2) = S_1 / V_1 + S_2 / V_2 = S_1 / V + S_2 / V$

digressione sulla media: esempio di calcolo della media

- un'automobile percorre 225 Km viaggiando a *60 Km/h* per i primi 120 Km e a *105 Km/h* per i restanti 105 Km
- ci chiediamo quale è la velocità (costante) che dovrebbe avere per percorrere *la stessa strada nello stesso tempo*
  - essa è la media armonica delle velocità pesate sugli spazi percorsi
  - $V = (S_1 + S_2) / (S_1/V_1 + S_2/V_2) = 225 / (120/60 + 105/105) = 75$ Km/h (120 Km a 60 Km/h e 105 Km a 105 Km/h)
  - oppure la media aritmetica delle velocità pesate sui tempi
  - $V = (V_1t_1 + V_2t_2) / (t_1 + t_2) = 60 \times 2/3 + 105 \times 1/3 = 75$ Km/h (2 ore a 60 Km/h e 1 ora a 105 Km/h)
- (l'automobile che percorre entrambi i tratti di strada a velocità V impiega lo stesso *tempo*)

- $f(V_1,V_2) = S_1 / V_1 + S_2 / V_2 = S_1 / V + S_2 / V$

digressione sulla media: esempio di calcolo della media (cont.)

- il consumo di carburante (nell'unità di tempo) è proporzionale al quadrato della velocità
- ci chiediamo quale è la velocità che dovrebbe avere l'automobile per consumare la *stessa quantità di carburante nello stesso tempo*

- $f(V_1,V_2) = S_1 / V_1 + S_2 / V_2 = S_1 / V + S_2 / V$

digressione sulla media: esempio di calcolo della media (cont.)

- il consumo di carburante (nell'unità di tempo) è proporzionale al quadrato della velocità
- ci chiediamo quale è la velocità che dovrebbe avere l'automobile per consumare la *stessa quantità di carburante nello stesso tempo*
- essa è la media quadratica pesata sui tempi
  - consumo = $k \times V^2 = (k \times V_1^2\, t_1 + k \times V_2^2\, t_2) / (t_1 + t_2)$
  - $V^2 = 60^2 \times 2/3 + 105^2 \times 1/3$; $V = 77.94$ Km/h
- agli effetti del *carburante* (che è rimasto lo stesso) è come se avesse viaggiato per 3 ore a 77.94 Km/h
  - (ovviamente lo spazio percorso sarà diverso: 233.82 Km)

- $f(v_1,v_2) = V_1^2 t_1 + V_2^2 t_2 = V^2 t_1 + V^2 t_2$

digressione sulla media: esempio di calcolo della media (cont.)

- ci chiediamo infine quale è la velocità che dovrebbe avere l'automobile per percorrere la *stessa strada con lo stesso consumo*

- $f(V_1,V_2) = S_1 / V_1 + S_2 / V_2 = S_1 / V + S_2 / V$

digressione sulla media: esempio di calcolo della media (cont.)

- ci chiediamo infine quale è la velocità che dovrebbe avere l'automobile per percorrere la *stessa strada con lo stesso consumo*
- è la media aritmetica pesata sugli spazi:
  - $(S_1 + S_2) = VT'$; $S_1 = V_1 t_1$; $S_2 = V_2 t_2$
  - $V^2T' = V_1^2 t_1 + V_2^2 t_2$ (consumi)
  - $VVT' = (S_1 + S_2)\, V$ allora:
  - $(S_1 + S_2)\, V = S_1 V_1 + S_2 V_2$
  - $V = 60 \times 120 / 225 + 105 \times 105 / 225 = 81$ Km/h
- agli effetti del *consumo* (che è rimasto lo stesso) è come se avesse percorso la stessa strada a 81 Km/h
  - (ovviamente il tempo impiegato sarà diverso: 2.78 h)

- $f(V_1, V_2) = S_1 V_1 + S_2 V_2 = S_1 V + S_2 V$

digressione sulla media: esempio di calcolo della media (cont.)

- una *mucca* produce 80 litri di latte secondo la seguente modalità:
  - 30 litri di latte in 3 giorni : (10 / giorno)
  - 30 litri di latte in 2 giorni : (15 / giorno)
  - 20 litri di latte in 1 giorno : (20 / giorno)

- quanti *litri/giorno* dovrebbe produrre la *mucca* per dare luogo alla stessa produzione?

digressione sulla media: esempio di calcolo della media (cont.)

- una *mucca* produce 80 litri di latte secondo la seguente modalità:
  - 30 litri di latte in 3 giorni : (10 / giorno)
  - 30 litri di latte in 2 giorni : (15 / giorno)
  - 20 litri di latte in 1 giorno : (20 / giorno)

- quanti *litri/giorno* dovrebbe produrre la *mucca media* per dare luogo alla stessa produzione?

$$\frac{80}{30\cdot\frac{1}{10}+30\cdot\frac{1}{15}+20\cdot\frac{1}{20}}=\frac{80}{3+2+1}=13.\overline{3}$$

media armonica delle produzioni pesata sui litri di latte prodotti

- infatti *6 giorni* $\times$ *13.(3) litri/giorno = 80 litri*
  le prestazioni (es.: velocità, produzione) vanno mediate armonicamente sul prodotto (Km, litri)

digressione sulla media: (cont.)

- la media gode della *proprietà associativa*, cioè non cambia se a gruppi di dati si sostituisce la loro media
  - *se dividiamo i valori in più classi e facciamo le medie per classe, la media delle medie parziali (con pesi uguali alla somma dei pesi degli elementi in ciascuna classe) resta inalterata*
- le medie associative *(teorema di Nagumo - Kolmogorov)* sono le *trasformate della media aritmetica* secondo una funzione **g**

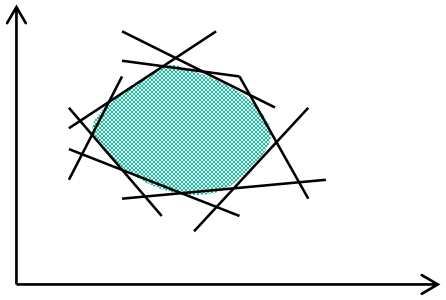
$$M = g^{-1}\left(g(x_1)\cdot p_1 + g(x_2)\cdot p_2 + \cdots\right)$$

$$es: \qquad M\ geo. = \exp\left(\left(\log x_1 + \log x_2 + \cdots\right)/n\right) = \left(x_1 \cdot x_2 \cdots\right)^{1/n}$$

digressione sulla media: (cont.)

- media armonica ≤ geometrica ≤ aritmetica ≤ quadratica ≤ cubica …

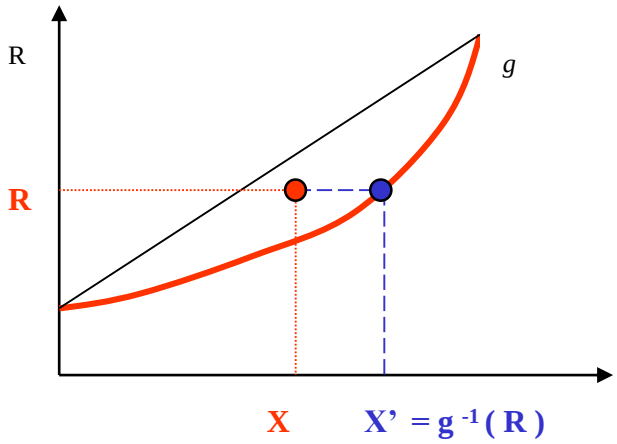
- la media è *interna* al "politopo" convesso delle misure
  - in particolare, per una grandezza scalare, se *min* e *Max* sono rispettivamente il minimo e il massimo valore osservato, per *qualunque media M* vale:
  - $min \leq M \leq Max$



- le medie geometriche conservano i rapporti (il rapporto di due medie$_G$ è la media$_G$ dei rapporti)

digressione sulla media: una applicazione

- tempo di risposta in funzione del flusso: curva di prestazioni nel piano (X,R)
- il punto medio (*X*,*R*) non si trova sulla curva *g(R)*.
- *X'*: flusso medio di transazioni ai fini del tempo di risposta *R*
- durante il periodo di osservazione il punto rappresentativo del fenomeno si muove sulla curva *R = g(X)* al variare di X
- (*X* = *X'* solo se *g* è una funzione lineare)



X' è il tasso di transazioni che darebbe luogo al tempo di risposta ottenuto come media delle misure

esercizio: considerazioni sul tempo medio di risposta

- un insieme di nodi costituisce un sistema “cloud computing”
- job, statisticamente identici, sono eseguiti sui nodi di due diverse tipologie con le seguenti prestazioni:

tipologia	ripartizione	tempo medio	throughput
A	0,4	100	1/100
B	0,6	152	2/152

- ipotesi 1:
  l’utente che richiede l’esecuzione dei job sceglie a caso un nodo:
- valore atteso del tempo di risposta =
  r1 = 100×0.4 + 152×0.6 = ***131.2***
- ipotesi 2:
  il sistema stesso decide il nodo di esecuzione, rispettando le frequenze
- valore atteso del tempo di risposta =
  r2 = (100×0.4×1/100 + 152×0.6×2/152) / (0.4×1/100 + 0.6×2/152)
  = ***134.5***

principali indici: media e varianza

- spesso si trova la notazione *E[X]* per la media (valore atteso: *Expectation*)

  - $E[X] = \Sigma\ p_i \times x_i \qquad (0 \le p_i\ ;\ \Sigma\ p_i = 1)$ — *media = baricentro*
  - $V(X) = M[(X - M(X))^2] = E(X^2) - [M(X)]^2$ — *varianza = momento di inerzia*
  - *C* = coefficiente di variazione — $(C^2 = V / M^2)$

  - $\Sigma\ p_i\ (x_i - M)^2 = \Sigma\ p_i\ (x_i^2 - 2x_iM + M^2) = \Sigma\ p_i x_i^2 - 2\Sigma\ p_i x_i M + M^2$

  - $E[X] = \int x f(x) dx = \int x dF(x)$

principali indici: media e varianza (cont.)

date due grandezze X,Y

- $M(X+Y) = M(X) + M(Y)$ *media della somma*
- $Cov[X,Y] = M[(X-M(X))(Y-M(Y))]$ *covarianza*
  $= M(XY) - M(X)M(Y)$
- $V(X+Y) = V(X)+V(Y)+2Cov[X,Y]$ *varianza della somma*
- *r* = coefficiente di correlazione
  - $(r_{12} = Cov_{12} / sqrt(V_1 \times V_2))$
  - $(-1 \leq r \leq 1)$ *$r^2 = 1$ correlazione perfetta*
- se X, Y indipendenti:
  - $V(X+Y) = V(X)+V(Y)$; $Cov[X,Y] = 0$

principali indici: media e varianza (cont.)

- date N grandezze X_i indipendenti di identica distribuzione con media M e varianza V
  - (indipendenza statistica $\Rightarrow$ $Cov[X_i,Y_j] = 0$; $i \neq j$)
- $V[(X_1+X_2+\ldots X_N)/N] = 1/N^2 \times V[(X_1+X_2+\ldots X_N)] = V(X)/N$

Varianza della media

- nel caso generale:
- $V[\Sigma X_i] = \Sigma V[X_i] + \Sigma_{i<j}\, 2Cov[X_i, X_j]$

- $S_N = X_1 + X_2 + \ldots\ldots X_N$ (N e X grandezze casuali indipendenti)
  - $E[S_N] = E[X] \times E[N]$
  - $V[S_N] = E[N] \times V[X] + V[N] \times E^2[X]$

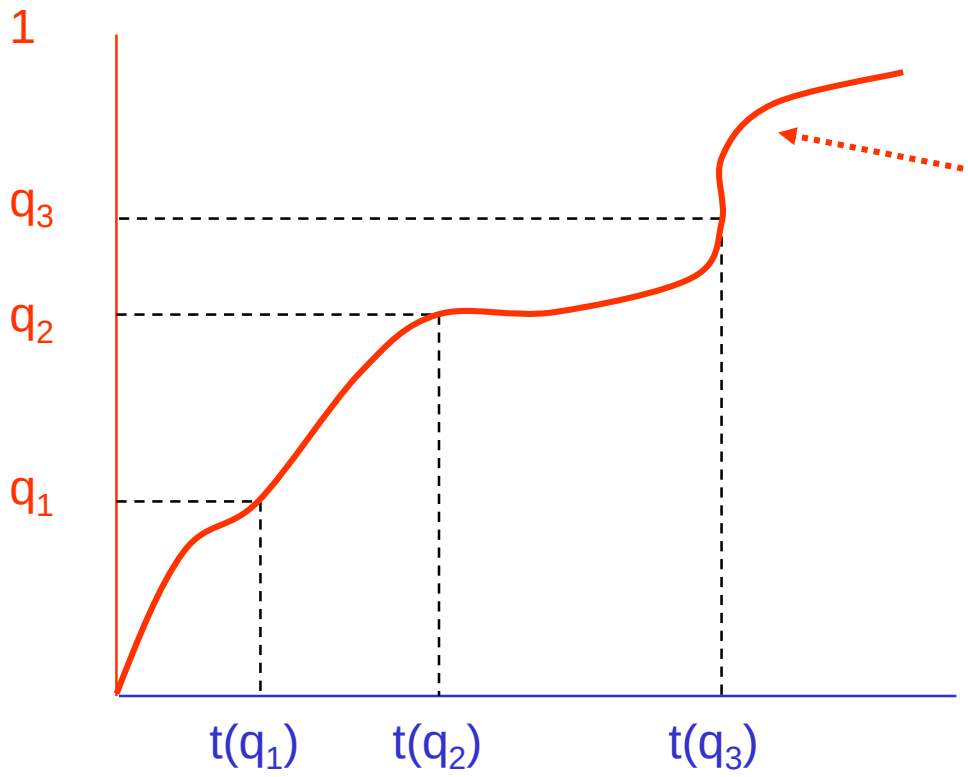
principali indici: media e varianza (cont.)

- il calcolo della *media* e della *varianza* di una serie $\{x_i\}$ di valori può essere eseguito anche in modo *ricorsivo*:
- $X_i = X_{i-1} + (x_i - X_{i-1}) / i$, $\quad i \geqq 1$
- $s_i^2 = ((i - 2) / (i - 1))\, s_{i-1}^2 + (x_i - X_{i-1})^2 / i$, $\quad i \geqq 1$
- con le posizioni iniziali:
- $X_0 = 0$, $\quad s_0^2 = s_1^2 = 0$
- i valori X_i e s_i sono la media e la deviazione standard della porzione di serie fino al termine i-esimo
- $s_n^2 = n/(n-1) \times V$
  - $E[s^2] = V$ (s^2: stima unbiased della Varianza della popolazione)

momenti e percentili

- $E[X^m]$ è detto *momento di ordine m* (per m=1 è la media)
- t_q è detto *100q percentile* di X, ha il seguente significato:

$t < t_q \Rightarrow F_X(t) < q; \qquad t \geq t_q \Rightarrow F_X(t) \geq q$



$F_X(t) = prob(X \leq t)$
$F_X(t)$: distribuzione di X

$F_X(t) = \int_0^t f(x)dx = \int_0^t dF(x)$

$prob(x \leq t) = \Sigma_i\, p_i\, (x_i \leq t)$

momenti e percentili (cont.)

- se si conosce la forma analitica della distribuzione $F_X(t)$, i percentili si possono calcolare direttamente, per esempio se la distribuzione è esponenziale di media <X>:

$$F(t) = 1 - e^{-\frac{t}{<X>}} \qquad P(x \geq t) = 1 - F(t) = e^{-\frac{t}{<X>}}$$

$$90\%le \rightarrow 0.1 = e^{-\frac{t90}{<X>}} \rightarrow t90 = <X> \ln 10 = 2.3 <X>$$

$$95\%le \rightarrow 0.05 = e^{-\frac{t95}{<X>}} \rightarrow t95 = <X> \ln 20 = 3 <X>$$

alcune diseguaglianze utili

- vogliamo considerare alcune diseguaglianze che possono essere applicate per trarre qualche considerazione sulla *coda* di una distribuzione di cui si conoscano solo i primi due momenti (in altre parole la media e la varianza):
- la più semplice è quella di *Markov* che richiede la sola conoscenza della media m.

$$m = \sum_{xi} x_i\, p_i = \sum_{xi<t} x_i\, p_i + \sum_{xi\geq t} x_i\, p_i \geq \sum_{t\leq xi} x_i\, p_i \geq \sum_{t\leq xi} t\, p_i \qquad \{t > m\}$$

$$= t\, P[X \geq t]$$

$$P[X \geq t] \leq \frac{m}{t} \qquad t > 0\,,\; P[X < 0] = 0$$

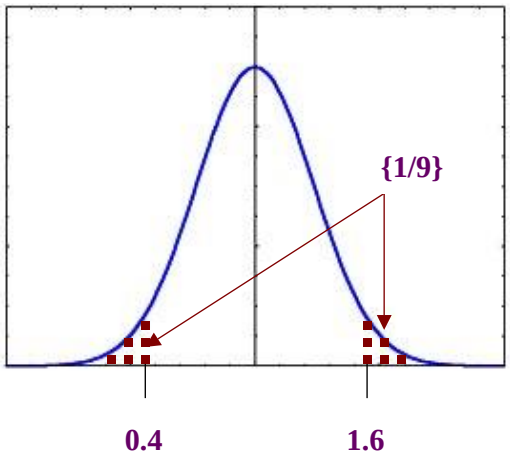
un sistema ha un tempo di risposta stimato di 1 sec. La probabilità che il tempo di risposta X sia maggiore di 2 sec. è allora:
$P[X \geq 2] \leq 1/2$

alcune diseguaglianze utili (cont.)

- diseguaglianza di *Chebychev*
- riguarda la coda (da entrambe le parti) della distribuzione di cui si conosce la media *m* e la deviazione standard σ

$$P\left[\left|X-m\right| \geq t\right] \leq \frac{\sigma^2}{t^2}$$

$$P\left[X \geq t\right] \leq \frac{\sigma^2}{(t-m)^2} \quad \textbf{se } 0 \leq m < t$$



il sistema della pagina precedente ha dev. std. pari a 0.2 -
si vuole conoscere la probabilità che il tempo di risposta sia compreso fra 0.4 e 1.6 sec.

$P[(X \leq 0.4) \cup (X \geq 1.6)] = P[|X-1| \geq 0.6]$
$\leq (0.2/0.6)^2 = 1/9$

$P[0.4 < X < 1.6] = 1 - P[|X-1| \geq 0.6] \geq 8/9$

$|X - m| = 0.6 \quad \sigma = 0.2$

alcune diseguaglianze utili (cont.)

- diseguaglianza di *Cramer*
- riguarda la coda (da un lato) di una distribuzione di cui si conoscono la media *m* e della std. dev. σ

$$P[X > t] \leq \frac{\sigma^2}{\sigma^2 + (t-m)^2} \quad , \quad t > m$$

un modello di un server web fornisce un tempo medio di risposta di 400 msec con una std. dev. di 116 msec
una specifica del progetto è che il 90 percento delle richieste non devono superare i 750 msec

dis. Cramer: $P[X>750] \leq 1/(1+(350/116)^2) = 0.09897$
$|X - m| \geq 350 \quad \sigma = 116$

dis. Chebychev: $P[X \geq 750] = P[X- 400 \geq 350] \leq P[|X- 400| \geq 350]$
$\leq (116/350)^2 = 0.1098$

intervallo di confidenza

- μ è il valore *vero* del parametro che si vuole stimare (es.: valore medio)

$$P\left\{\left|\overline{X}-\mu\right|<\varepsilon\right\}\geq 1-\alpha$$ può essere scritta nelle forme equivalenti:

$$(1)\quad P\left\{\overline{X}-\varepsilon<\mu<\overline{X}+\varepsilon\right\}\geq 1-\alpha$$

$$\left(\overline{X}-\varepsilon\,,\,\overline{X}+\varepsilon\right)$$

intervallo di confidenza (1-α) % per il parametro μ

è un intervallo osservato per l'ipotetico valore $\boldsymbol{\mu}$

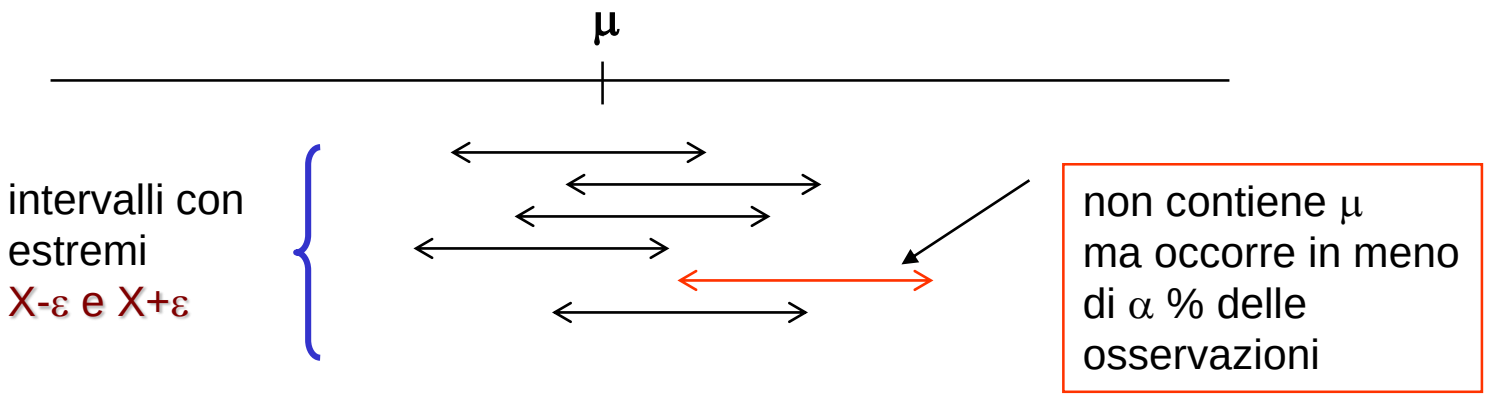
$$(2)\quad P\left\{\mu-\varepsilon<\overline{X}<\mu+\varepsilon\right\}\geq 1-\alpha$$

$$\left(\mu-\varepsilon\,,\,\mu+\varepsilon\right)$$

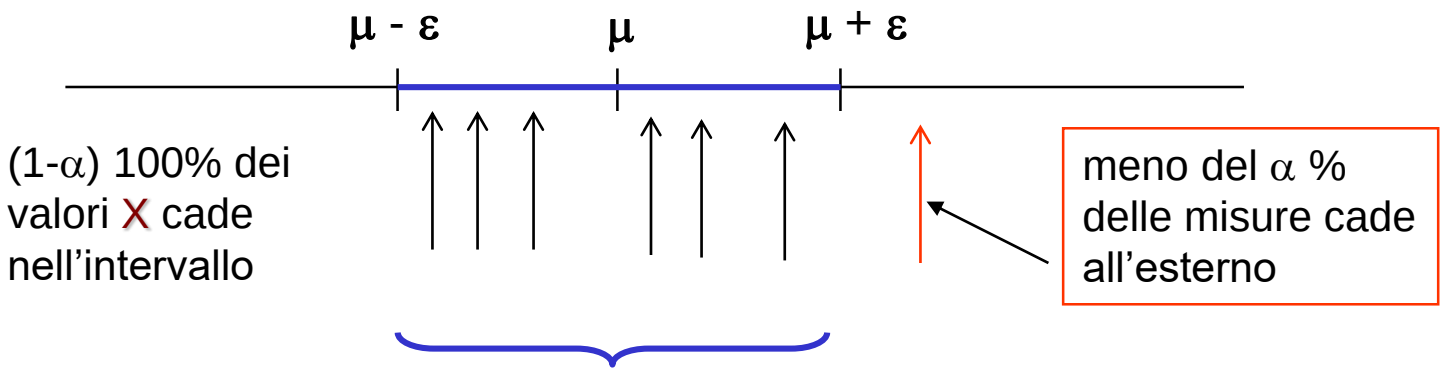
è un ipotetico intervallo per il valore osservato $\boldsymbol{\overline{X}}$

intervallo di confidenza (cont.)

- interpretazione (1)



- interpretazione (2)



modalità di misura
di una grandezza

modalità della misura: diretta

- *esempio: calcolo dell'utilizzo* sia *U(t) = 0, 1*

$$U_a^b = \int_a^b U(t)dt \Big/ (b-a)$$

- *misura diretta:* ogni volta che il componente cambia stato viene tenuta la traccia
  - in t_{2k+1} lo stato varia da 1 a 0 $t_{2K+1} = b$
  - in t_{2k} lo stato varia da 0 a 1 $t_0 = a$

$$t(busy) = \sum_{k=0}^{K} (t_{2k+1} - t_{2k}) \qquad U_a^b = t(busy)/(b-a)$$

modalità della misura: campionamento

- *campionamento:* a intervalli t_i regolari o casuali (non correlati con il comportamento del componente)

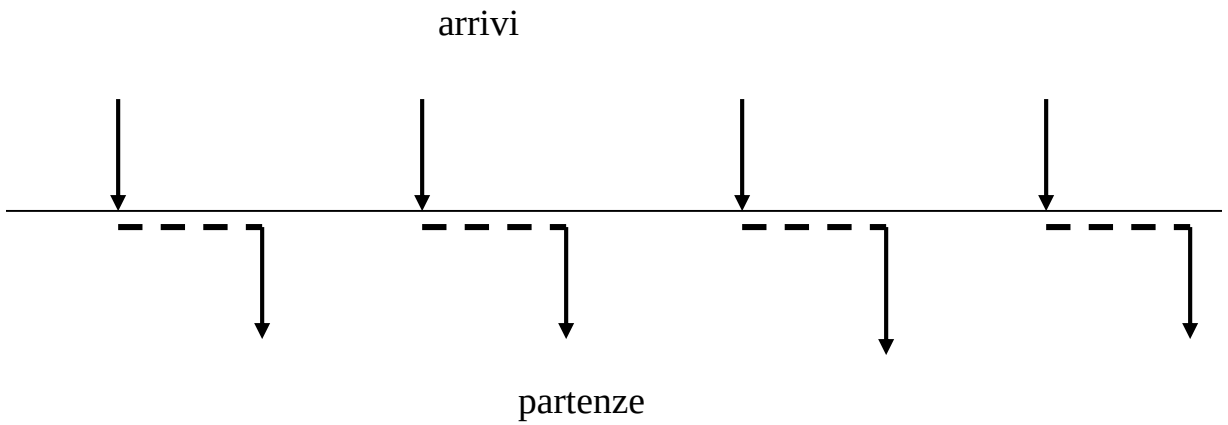
$$\{t_i\} \quad i = 1,2,...N$$

$$U = \frac{\sum_i U(t_i)}{N}$$

- *U* è una *frazione di tempo* e può essere interpretata come:
  - *numero medio di utenti in servizio* o
  - *probabilità* che un utente arrivando *a caso* trovi il componente occupato. $0 \leq U \leq 1$

modalità della misura: campionamento (cont.)

- se gli arrivi non sono casuali:
  - esempio banale:
    - arrivi (e servizi) regolari
    - utilizzo effettivo: 50%
    - utilizzo visto dall'utente: 0% (trova sempre il componente libero)



La probabilità di trovare *busy* è uguale all'utilizzo solo se gli arrivi sono casuali

(Poisson Arrivals See Time Average)

modalità della misura: campionamento (cont.)

- un campionamento di *N* valori x_i indipendenti e identicamente distribuiti da una popolazione di media μ e deviazione standard σ fornisce una stima con la seguente proprietà (dal *teorema del limite centrale*)

$$\Pr\left\{a < \frac{\bar{x}_N - \mu}{\sigma/\sqrt{N}} < b\right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2}\, dz \qquad per\ N \rightarrow \infty$$

la stima $\bar{X}_N$ si comporta allo stesso modo (per $N \rightarrow \infty$) per le infinite distribuzioni con gli stessi valori μ e σ
la convergenza è più rapida se la distribuzione di partenza è già approssimativamente normale

Distribuzione Normale standard N(0,1)
$\mu = 0$; $\sigma = 1$

$\sigma/\sqrt{N}$: std. dev. della media

modalità della misura: campionamento (cont.)

- ponendo ora b = t; a = - b = - t possiamo anche scrivere:

$$\Pr\left\{\left|\bar{x}_N - \mu\right| < t\frac{\sigma}{\sqrt{N}}\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-t}^{t} e^{-z^2/2}\,dz$$

$$\Pr\left\{\bar{x}_N - t\frac{\sigma}{\sqrt{N}} < \mu < \bar{x}_N + t\frac{\sigma}{\sqrt{N}}\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-t}^{t} e^{-z^2/2}\,dz$$

  - t determina univocamente la probabilità
- per dimezzare l'intervallo di confidenza (a parità di confidenza) bisogna quadruplicare il numero delle misure

$$\left|\bar{x}_{N2} - \mu\right| = \tfrac{1}{2}\left|\bar{x}_{N1} - \mu\right| \Rightarrow N2 = 4 \cdot N1$$

$$\frac{t\sigma}{\sqrt{N2}} = \frac{1}{2}\frac{t\sigma}{\sqrt{N1}} \Rightarrow N2 = 4 \cdot N1$$

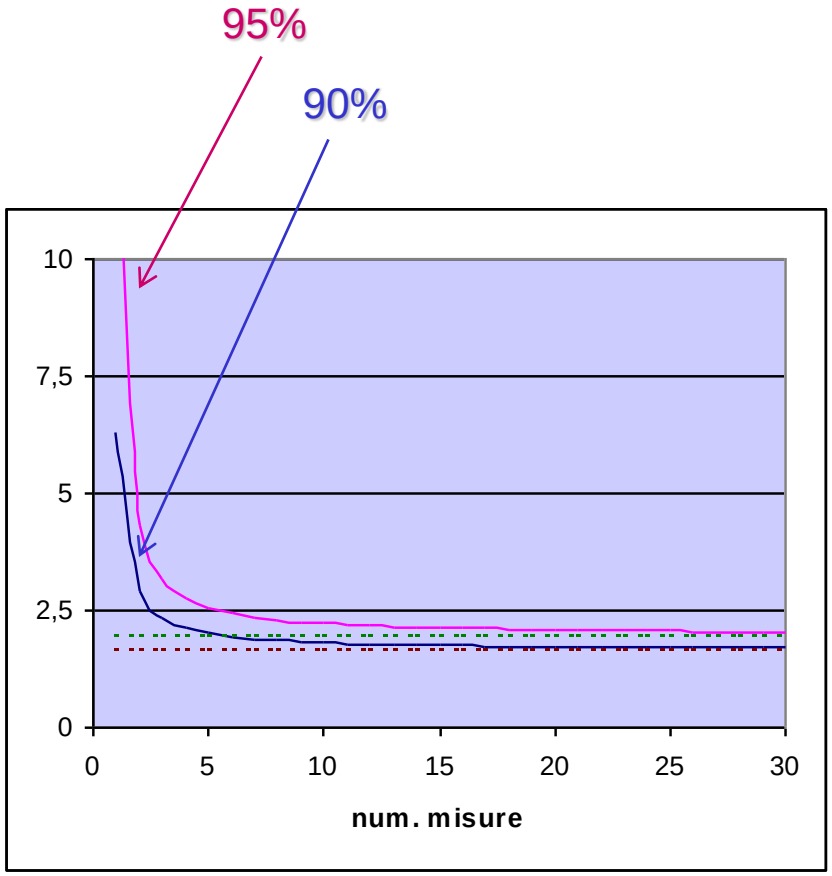
modalità della misura: campionamento (cont.)

- avendo fissato il valore della probabilità (confidenza) si può determinare l'ampiezza dell'intervallo (errore):
- al crescere del numero N delle misure, tende al 95 percento la probabilità che il valore *vero* della media, stimato da M, sia compreso nell'intervallo:
- $(M - 1.96 \times \sigma / (N)^{1/2}, M + 1.96 \times \sigma / (N)^{1/2})$
- la deviazione standard σ, nel caso di una grandezza che, come l'utilizzo, ha due possibili valori puntuali (0: *idle;* 1: *busy*), è al massimo pari a 1/2.
  - $\sigma^2 = p(1^2) + (1-p)0^2 - p^2 = p - p^2$
  - $\sigma^2 = p \times (1-p) \leq ¼$ (p: probabilità dello stato 1 = media)
- (la tabella che appare nel seguito è stata calcolata con questa ipotesi).

modalità della misura: campionamento (cont.)

l'ampiezza dell'intervallo di confidenza dipende da:

- *N* numero osservazioni
- *coefficiente* (1.96 se al 95%; 1.64 se al 90%)
  - anche il coefficiente dipende da *N*
  - vedi grafico (t student)
  - 1.96 e 1.64 sono i valori limite (per *N* "grande")



modalità della misura: campionamento (cont.)

durata (minuti)	semiampiezza intervallo	
1	7.305	La tabella riportata a titolo di esempio, contiene (in funzione della durata della misura) la semiampiezza dell'intervallo di confidenza al 95 percento, per il caso dell'utilizzo percentuale stimato attraverso tre campionamenti al secondo.
5	3.267	
15	1.886	
30	1.334	
60	0.943	
120	0.664	

Esempio: 1 minuto; N = 180 misure
$1.96 \times 0.5/\text{sqrt}(180) = 0.07305$
($\sigma = 0.5$)

modalità della misura: campionamento (cont.)

- ridurre l'intervallo

- aumentare la durata dell'intervallo

modalità della misura: campionamento (cont.)

- ridurre l'intervallo fra misure successive le rende correlate e aumenta il carico di *overhead* e la distorsione dovuta alla misura stessa
  - se le grandezze sono correlate la varianza contiene anche un termine positivo di autocorrelazione perciò l'aumento del numero di osservazioni non aumenta l'accuratezza (non si riduce l'errore)

    autocorrelazione(k) = $Cov(x_i , x_{i+k}) = 1/N \Sigma x_i x_{i+k} – M^2$

- aumentare la durata dell'intervallo di misura è possibile solo se non si coprono periodi *transienti*

  (se la statistica del fenomeno varia nel tempo la media usuale non ha significato)

curva di prestazioni

curva di prestazioni (indicativa)

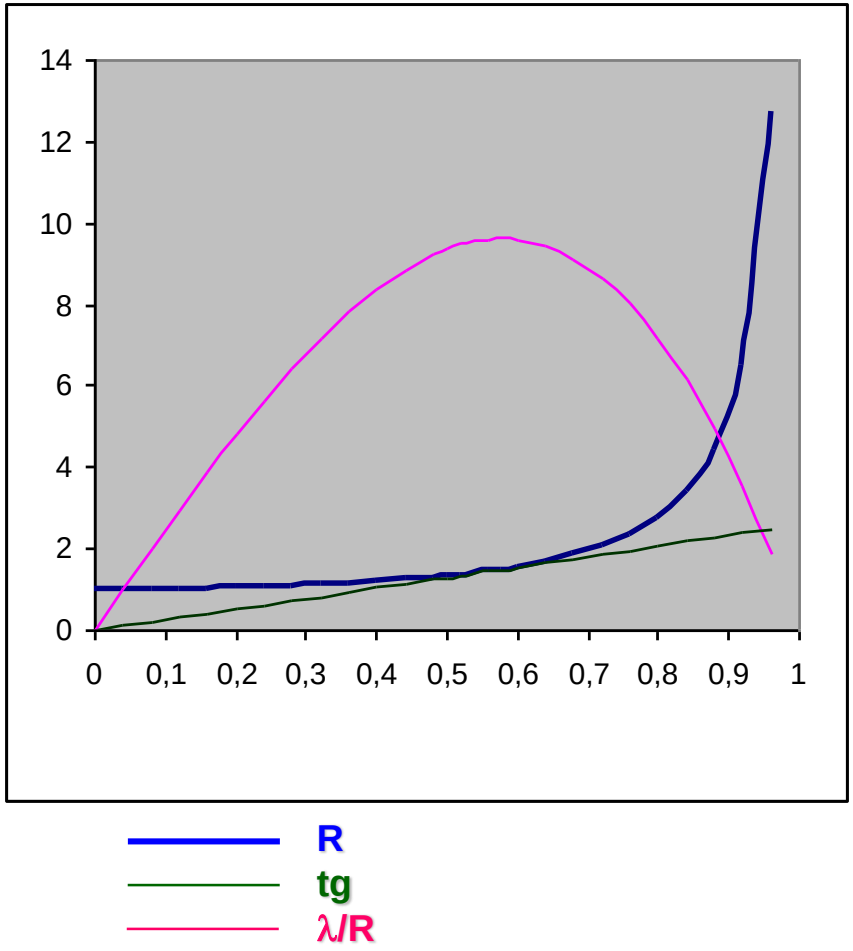
- i legami fra gli indici possono essere rappresentati dalle *curve di prestazione*
- diagrammi che rappresentano l'andamento delle prestazioni al variare di un parametro (per esempio *tempo di risposta* in funzione del *traffico servito*)
- grandezze principali in gioco:
- *N* numero di utenti
- λ tasso di completamento/arrivo
- R tempo di risposta
  - dipende da
    - tempi di servizio
    - tempi di attesa

curva di prestazioni (cont.)

- idealmente si vorrebbe operare con un sistema che presenti il minimo ritardo e la massima produttività
- il punto ottimale di funzionamento è quello di tangenza con la retta passante per l'origine in cui è massimo il rapporto

  $\lambda / R(\lambda)$

  che è una misura di *relativa preferenza*



curva di prestazioni (cont.)

$$d\left(\frac{\lambda}{R}\right)=\frac{d\lambda}{R(\lambda)}-\frac{\lambda\cdot dR(\lambda)}{R^2(\lambda)}=0\Rightarrow\left(\frac{d\lambda}{\lambda}=\frac{dR}{R(\lambda)}\right)_{\lambda=\lambda_0}$$

$$\frac{dR}{d\lambda}=\frac{R(\lambda)}{\lambda}\qquad per\ \lambda=\lambda_0$$

$$\lambda>\lambda_0\Rightarrow\frac{dR}{R}>\frac{d\lambda}{\lambda}$$

$$\lambda<\lambda_0\Rightarrow\frac{dR}{R}<\frac{d\lambda}{\lambda}$$

- λ_0 : punto di massimo di λ / R(λ)
  - (ipotizziamo un'unica classe di carico)
- a sinistra (destra) di λ_0 dR/R cresce meno (più) rapidamente di dλ /λ
- la forma della curva di prestazione determina la posizione di λ_0

processi
introduzione matematica

necessità di un approccio probabilistico

- i meccanismi che determinano il comportamento di un impianto informatico sono complessi e dipendono da fattori diversi:
  - umani, tecnologici, ambientali,...
- le grandezze relative all'affidabilità e alle prestazioni di un sistema non possono essere descritte da leggi deterministiche
  - si è obbligati a considerare queste grandezze in termini statistici
  - le leggi che le governano sono di tipo probabilistico

distribuzione esponenziale

- un *processo* consiste di eventi che si svolgono nel tempo
- se la durata dell'intervallo di tempo t_a fra due eventi successivi ha probabilità che dipende dal tempo Δt secondo la legge:
  - $prob(t_a \leq \Delta t) = \lambda \Delta t + o(\Delta t)$
  - λ parametro costante

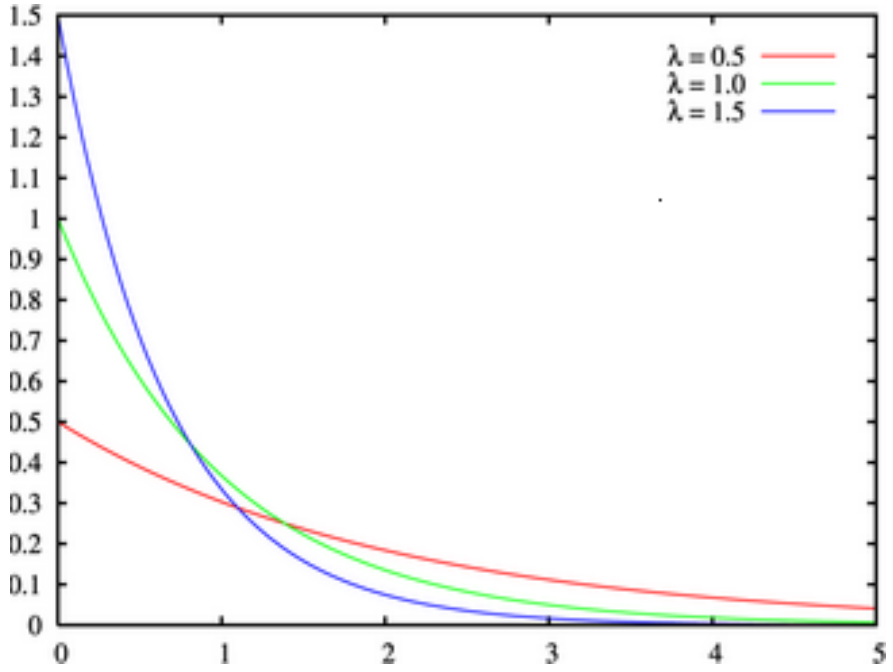
processo di Poisson

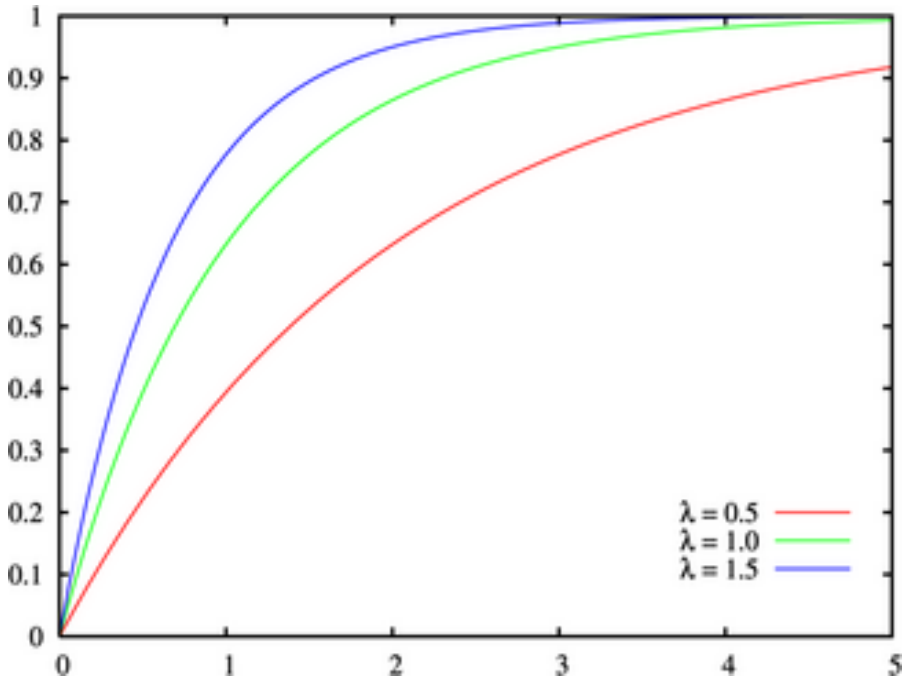
- la probabilità che in un tempo Δt non si verifichi l'evento è:
  - $(1- \lambda \Delta t)$ (a meno di infinitesimi di ordine superiore)
- dividiamo l'intervallo t in n intervallini di durata $\Delta t = t/n$

$$prob(t_a > t) = \left(1 - \frac{\lambda \cdot t}{n}\right)^n \xrightarrow[n \to \infty]{} e^{-\lambda \cdot t}$$

Probabilità che nell'intervallo *di durata t* non si verifichi l'evento

distribuzione esponenziale (cont.)





$$f_X(t) = \begin{cases} 0, & t < 0 \\ \lambda e^{-\lambda t}, & t \geq 0 \end{cases}$$

$$F_X(t) = \begin{cases} 0, & t < 0 \\ 1 - e^{-\lambda t}, & t \geq 0 \end{cases}$$

$$Media = \frac{1}{\lambda} \qquad \sigma^2(t) = \frac{1}{\lambda^2}$$

mancanza di memoria (memoryless)

- la distribuzione esponenziale è l'unica che gode della seguente proprietà:
  - il tempo residuo (t_a - t_0) non dipende da quello già trascorso t_0:

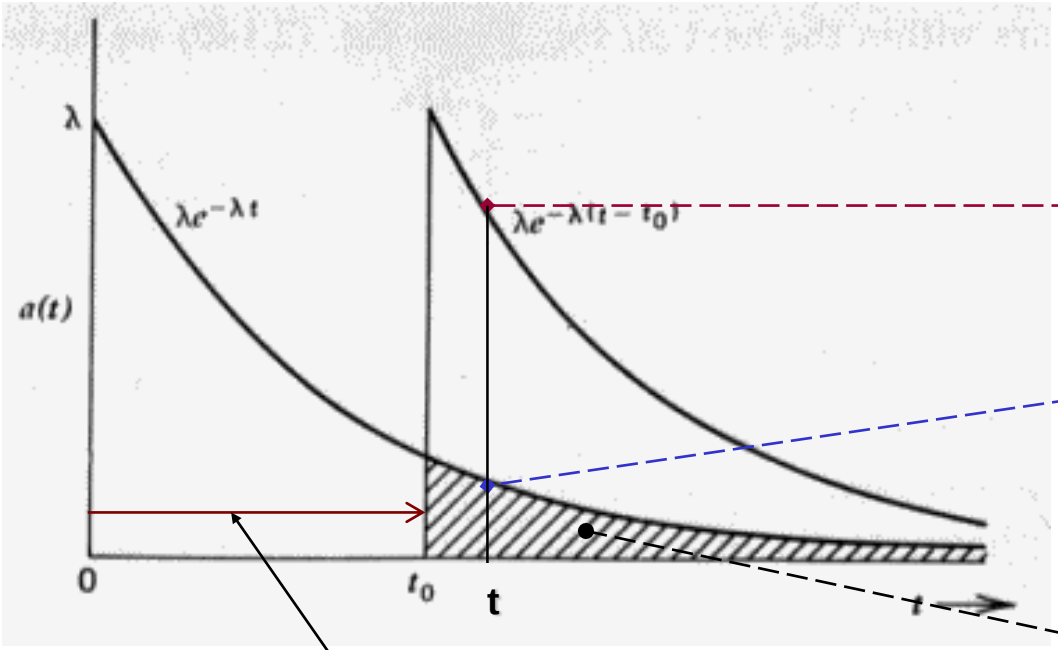
$$prob(t_a \le t_0 + t \mid t_a > t_0) = \frac{prob(t_0 < t_a \le t_0 + t)}{prob(t_a > t_0)} =$$

$$= \frac{prob(t_a \le t_0 + t) - prob(t_a \le t_0)}{prob(t_a > t_0)} = \frac{1 - e^{-\lambda(t_0+t)} - (1 - e^{-\lambda t_0})}{1 - (1 - e^{-\lambda t_0})} = 1 - e^{-\lambda t} =$$

$$= prob(t_a \le t)$$

- all'istante τ=0 si è verificato un evento, fino all'istante τ=t_0 non si sono verificati eventi, si vuole calcolare la probabilità che il prossimo evento si verifichi in (t_0 , t)
- un fenomeno che segue una distribuzione esponenziale è caratterizzato dal fatto che il futuro è indipendente dal passato

mancanza di memoria (memoryless) (cont.)



$$\frac{\lambda e^{-\lambda t}}{e^{-\lambda t_0}} = \lambda e^{-\lambda (t-t_0)}$$

$$\lambda e^{-\lambda t}$$

$$e^{-\lambda t_0}$$

- la densità di probabilità al punto t_0 è identica a quella iniziale (t=0)

processo di Poisson

- il numero n di arrivi in un intervallo t di un processo con *interarrivi* esponenziali ha distribuzione di Poisson (e viceversa)
- dividiamo l'intervallo t in m intervalli di durata $\Delta t = t/m$

$$prob(n;t) = \binom{m}{n} (\lambda\,\Delta t)^n\,(1-\lambda\,\Delta t)^{m-n} = \frac{(\lambda t)^n}{n!}\,\frac{m!}{m^n\,(m-n)!}\left(1-\frac{\lambda t}{m}\right)^{m-n}$$

$$\xrightarrow[m\to\infty]{} \quad \frac{(\lambda t)^n}{n!}\,e^{-\lambda t}$$

In particolare: $F(t) = 1 - prob(0;t) = 1 - e^{-\lambda t}$

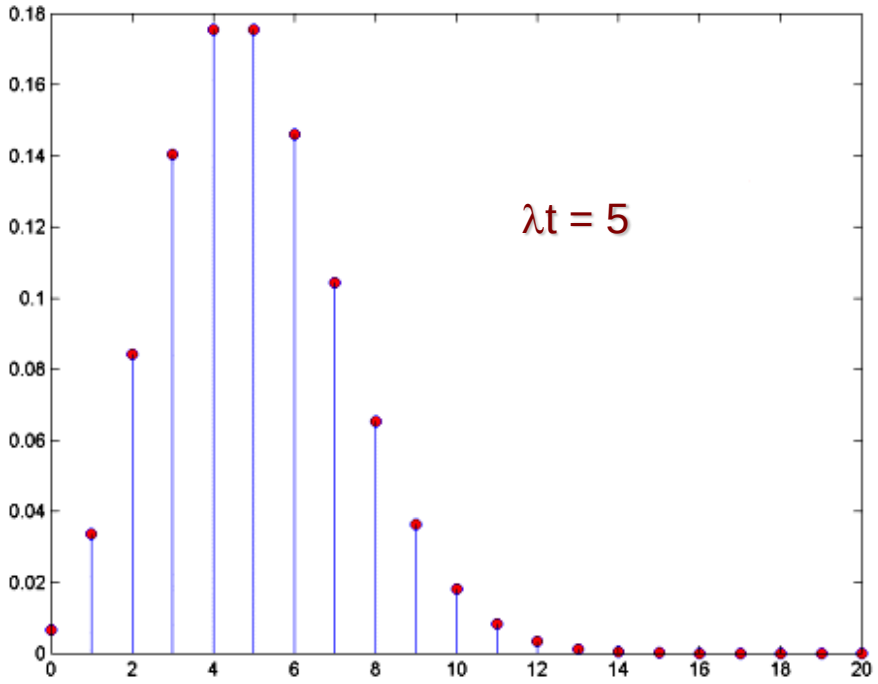
$$\frac{m!}{m^n\,(m-n)!} = \frac{m(m-1)(m-2)\cdots(m-n+1)}{m^n} \underset{m\to\infty}{\to} 1$$

processo di Poisson (cont.)

$$prob(0;t) = e^{-\lambda t}$$

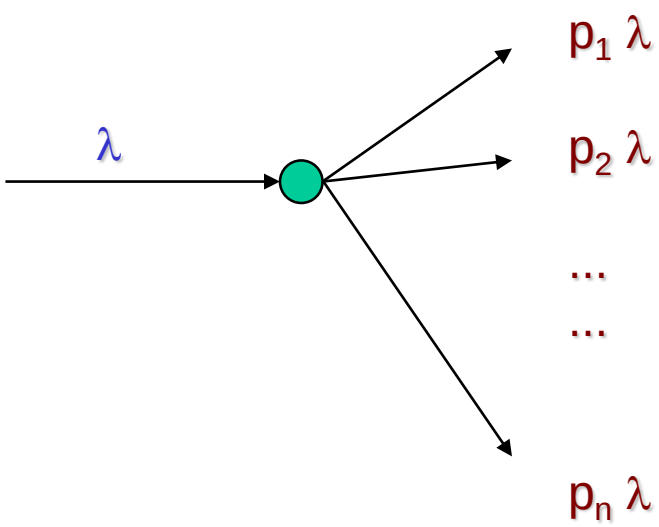
$$F(t) = prob(n > 0;t) = 1 - prob(0;t) = 1 - e^{-\lambda t}$$

prob(n)



λt = numero medio
di eventi nel tempo t
ed è anche il più
Probabile numero di arrivi

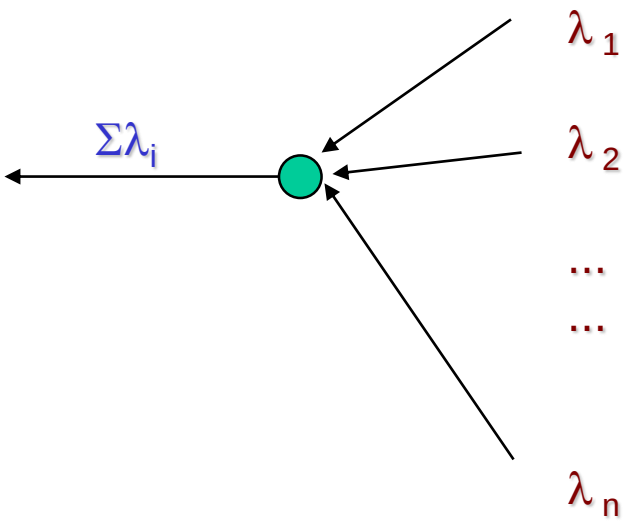
separazione (split) di processi di Poisson



- il ramo *j* a cui viene instradato l'arrivo è scelto con probabilità p_j
- così si generano n processi indipendenti con parametro caratteristico $p_i\lambda$ (i=1,..n)

- se il ramo *j* a cui viene instradato il flusso è invece scelto in modo non casuale, i processi generati non sono più poissoniani
- per esempio, se ogni arrivo è smistato in ordine di arrivo si ottengono n processi con interarrivi n-erlangiani

aggregazione (merge) di processi di Poisson



- il processo che si ottiene aggregando gli n processi originari è ancora poissoniano con parametro $\lambda = \Sigma \lambda_i$

- se vengono aggregati n processi (renewal, ma non necessariamente poissoniani) fra loro indipendenti, al crescere di n se $\lambda >> \lambda_i$ (per ogni i) il processo risultante è approssimativamente *poissoniano*

aggregazione (merge) di processi di Poisson (cont.)

- n processi mutuamente indipendenti sono caratterizzati dalle distribuzioni dei loro interarrivi: F_1, F_2,...., F_n di medie rispettivamente : $1/\lambda_1$, $1/\lambda_2$,... ,$1/\lambda_n$
- in condizioni di stazionarietà il processo aggregato avrà un interarrivo di media pari a $1/\lambda$ (con $\lambda = \lambda_1 + \lambda_2 + ...+ \lambda_n$)
- W_k: tempo di attesa al primo evento del processo k
- W: tempo di attesa al primo evento in assoluto (minimo dei W_k)

$$P\{W_k \le t\} \approx t\lambda_k$$ se t è piccolo rispetto a $1/\lambda_k$

$$P\{W > t\} \approx (1-t\lambda_1)(1-t\lambda_2)\cdots(1-t\lambda_n) \approx e^{-t\lambda}$$ per n grande