$$E(X)= \frac{1}{\lambda} \qquad (9.10.3)$$

La *varianza* della v.c. esponenziale negativa è definita dalla notazione:

$$V(X)= \frac{1}{\lambda^2} \qquad (9.10.4)$$

La *deviazione standard* della v.c. esponenziale negativa è definita dalla notazione:

$$\sigma = \sqrt{\frac{1}{\lambda^2}} \qquad (9.10.5)$$

L'*indice di asimmetria* della v.c. esponenziale negativa è definito dalla seguente notazione:

$$I_{AS}=2 \qquad (9.10.6)$$

L'*indice di curtosi* della v.c. esponenziale negativa è definito dalla seguente notazione:

$$I_{CUR} = 6 \qquad (9.10.7)$$

10. Teoremi limite del calcolo delle probabilità.

Oltre ai teoremi classici sulla probabilità presentati al paragrafo 4 si riportano di seguito i teoremi limite rappresentati dalle *Disuguaglianze di Chebyshev e di Markov***;** *dalla Legge empirica dei grandi numeri* e *dal teorema del limite centrale, anche definito teorema centrale del limite.*
Dopo aver definito gli indici di variabilità più comuni e noti come la varianza e la deviazione standard a fronte di distribuzioni conosciute, si può porre il problema di avere informazioni sui valori di variabilità in una situazione in cui non è nota la distribuzione. In questi casi si può utilizzare la *Disuguaglianza di Chebyshev* per avere informazioni sulla varianza. Essa stabilisce che, per ogni distribuzione di dati di una popolazione, la percentuale di essi che non si allontanano dalla media per una certa quantità dello scarto quadratico medio è pari almeno a:

$$(1 - \frac{1}{k^2})x100\% \qquad (10.1)$$

La disuguaglianza può essere espressa, in modo più completo, dalla seguente notazione:

$$\left| x_i - \mu \right| \geq k\sigma \leq \frac{1}{k^2} \qquad (10.2)$$

dove k è la quantità espressa da un numero puro positivo.

Se si pongono a confronto i valori che discendono dalla regola e quelli empirici si avranno le differenze rappresentate nella Tabella 10.1

Tabella 10.1 Valori di Chebyshev e della regola empirica

CHEBYSHEV	**REGOLA EMPIRICA**	
(μ-σ, μ+σ)	almeno 0%	circa il 68%
(μ-2σ, μ+2σ)	almeno 75%	circa il 95%
(μ-3σ, μ+3σ)	almeno 88%	circa il 99,7%

Dalla diseguaglianza di Chebyshev deriva la *Legge dei grandi numeri* che può essere: *debole* e *forte* (quando non si specifica l'aggettivo si sottintende sempre legge debole). Date n variabili mutuamente indipendenti con media μ e varianza σ^2 ed un numero positivo *a* la *legge debole* stabilisce che il limite per n che tende a ∞ della probabilità che in modulo la differenza tra la loro media aritmetica e il valore μ sia maggiore di a è uguale a zero. La notazione conseguente è data da:

$$\lim_{n \to \infty} P\left[\left| (X_1 + X_2 ..... + X_n)/n - \mu \right| > a \right] = 0 \qquad (10.3)$$

La *legge forte* stabilisce che**,** date n variabili mutuamente indipendenti con media μ e varianza σ^2, il limite per n che tende a ∞ della probabilità che in modulo la loro aritmetica sia uguale a μ è uguale a zero. La notazione conseguente è data da:

$$\lim_{n \to \infty} P\left| (X_1 + X_2 ..... + X_n)/n = \mu \right| = 1 \qquad (10.4)$$

Nella situazione in cui non si è a conoscenza della distribuzione della v.c. , si potrebbe avere l'esigenza di definire dei limiti alla probabilità. In questa circostanza , pur con forti limitazioni, si può utilizzare la *disuguaglianza di Markov* dove la probabilità che la v.c. X sia maggiore o uguale alla quantità h, deve essere minore o uguale al rapporto tra la media e la stessa quantità h. La notazione è:

$$P(X \geq h) \leq \frac{\bar{x}}{h} \qquad (10.5)$$

dove X è una v.c. non negativa e $\bar{x}$ è la media o valore atteso.
Dall'osservazione della si può notare che i limiti di probabilità possono essere trovati conoscendo solo il valore atteso $\bar{x}$ e la quantità h.
Nel modello inferenziale assume grande importanza *il teorema centrale del limite*.

Si consideri una successione di variabili casuali X_1, X_2, X_3,……,X_n e si analizzi la *convergenza in distribuzione* delle stesse, mettendo in relazione la funzione di ripartizione $F_n(x)$ con la funzione di ripartizione $F(x)$ di una singola v.c. *X.* Si può affermare che la successione delle variabili casuali X_1, X_2, X_3,……,X_n con funzione di ripartizione
$F_1(x), F_2(x), F_3(x), ...,$

$F_n(x)$ converge in distribuzione se:

$$\lim_{n->\infty} F_n(x) = F(x)$$

Si prendano in considerazione ora una successione di v.c. indipendenti e identicamente distribuite *(i.i.d.),*X_1, X_2, X_3,……,X_n con valore medio μ e con varianza σ^2; stabilito che $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e che la varianza campionaria è pari a σ /n si

ha che la v.c. standardizzata $Z_n = \frac{(\overline{X}_n - \mu)\sqrt{n}}{\sigma}$ *converge in distribuzione, per n* $\rightarrow \infty$*, alla Normale standardizzata* (si suppone che il teorema possa essere valido per valori della numerosità campionaria >30).

N.B. Poiché il termine gradi di libertà è stato incontrato più volte si ritiene opportuno darne una spiegazione. Esso è stato introdotto dal Fisher in analogia al significato di gradi di libertà di un sistema dinamico. Il concetto può essere spiegato considerando, ad esempio, quattro quantità a, b, c, d. Se ad esse facciamo corrispondere quattro valori, si può immaginare un numero infinito di quaterne di valori. Se, però, si impone il vincolo che a+b+c+d = 50 , si ha la libertà di fissare solo tre dei quattro termini in quanto il quarto è automaticamente determinato. Se si fissa, infatti, a =10; b=15; c=18 allora d è univocamente determinato ed è uguale a 7. Poiché si è liberi di fissare solo tre dei quattro valori della precedente equazione, si può affermare che la stessa gode di tre gradi di libertà (d.f. degree free). All'inverso se si considerano k vincoli lineari, o restrizioni con k<n, omogenei all'insieme di osservazioni n, il modello gode di n-k gradi di libertà. Nell'esempio citato in presenza di un vincolo e quattro osservazioni i gradi di libertà sono pari a 3 (4-1=3) e quindi coincidono

11. Teoria dell'inferenza statistica.

Il concetto di *inferenza* è uno dei fondamentali più importanti della scienza statistica. Esso stabilisce che si può avere evidenza empirica sui parametri della popolazione ignoti analizzando una parte di essa definita campione opportunamente rilevato. Si tratta di problemi di induzione che a loro volta si

rifanno ai concetti di *stima e di stimatore.*

11.1. Stimatori e relative proprietà.

Lo stimatore che è una v.c. che assume un valore che può essere definito stima puntuale oppure un intervallo di valori definiti stima intervallare. Esso può essere più o meno affidabile in funzione delle caratteristiche delle sue proprietà. Essi possono essere puntuali se stimano un solo valore oppure intervallari se stimano un intervallo di valori. In questa sede si analizzano solo gli stimatori intervallari per la media e la differenza di medie con varianza nota ed ignota, per la proporzione e la differenza fra proporzioni di popolazioni bernoulliane, per la varianza e il rapporto fra varianze di popolazioni normali.

Le proprietà "*ottimali*" che verranno considerate in questo paragrafo sono la *distorsione o correttezza; efficienza e consistenza.* Lo stimatore T si dice *corretto o non distorto (Unbiased)* se il suo valore medio o atteso è dato da:

$$E(T)= \mu \quad \text{per tutti i possibili valori di } \mu \qquad (11.1.1)$$

La distorsione dello stimatore T è data dalla differenza fra il suo valore medio o atteso e il valore del parametro della popolazione da stimare ovvero:

$$B(T)= E(T)- \mu \qquad (11.1.2)$$

Lo stimatore T si dice *efficiente* se la differenza fra se stesso e il valore del parametro della popolazione da stimare è il più basso possibile ovvero l'efficienza è una misura di dispersione o di variabilità dello stimatore.

Si consideri la quantità in valore assoluto $|T-\mu|$: essa è definita errore di stima; se la stessa quantità è elevata al quadrato $|T-\mu|^2$ essa è definita errore quadratico medio (*MSE-Mean Square Error)* che dà una misura, quindi, dell'efficienza dello stimatore T; quanto minore è il valore di *MSE* tanto più efficiente è lo stimatore T del parametro della popolazione μ.

L'errore quadratico medio dello stimatore T può essere definito, inoltre, dalla seguente notazione:

$$MSE(T)=Var(T) + B(T)^2 \qquad (11.1.3)$$

ovvero esso è dato dalla somma della varianza di T più la sua distorsione.
Nel caso in cui lo stimatore T è corretto allora:

$$MSE(T) = Var(T) \quad \text{per tutti i possibili valori di } \mu \qquad (11.1.4)$$

Se si hanno più stimatori (T_1, T_2,.....,T_n), il confronto tra di essi in termini di efficienza viene svolto attraverso il confronto fra le relative varianze; si dirà, ad esempio, che T_1 è più efficiente di T_2 se la $Var(T_1) < Var(T_2)$ e via di seguito e quindi si ha una *efficienza relativa*. Se invece lo stimatore T_1, ad esempio, è più efficiente di qualsiasi altro stimatore del parametro di interesse si può dire che esiste una *efficienza assoluta.*

Uno stimatore T si dice *consistente* se la sua precisione aumenta all'aumentare della dimensione campionaria. Si dice che uno stimatore T è *asintoticamente consistente* se al tendere all'infinito della numerosità campionaria il suo
valore o realizzazione tende al valore del parametro ignoto della popolazione. Ciò è possibile solo se lo stimatore T è *consistente in media quadratica* ovvero se tende a zero l'errore quadratico medio, ossia se:

$$\lim_{n\to\infty} MSE(T_n) = \lim_{n\to\infty} (T_n - \mu)^2 = 0 \qquad (11.1.5)$$

Esiste inoltre la *consistenza in probabilità* quando è verificata la condizione al limite:

$$\lim_{n\to\infty} P\left[\left|T_n - \mu\right| < a\right] = 1 \qquad (11.1.6)$$

dove $a > 0$ e T_n è lo stimatore funzione delle realizzazioni campionarie

11.2. Stima intervallare e relativi metodi.

Il motivo per cui in questa sede si analizza solo gli stimatori intervallari e le relative stime è che l'evento che uno stimatore produca una stima esattamente uguale al valore del parametro è quasi impossibile. Ed è per questo che il modello inferenziale tende a superare il grosso limite della stima puntuale attraverso l'inserimento di più stime accettabili a cui viene associato un predefinito livello di affidabilità (confidenza) e, quindi, a determinare un cosiddetto intervallo di valori a cui possa appartenere, ad un certo livello di fiducia, il valore del parametro ignoto. Il concetto di stima intervallare (o per intervallo) può essere esplicitato con un esempio. Siano μ il valore atteso e σ la deviazione standard della distribuzione campionaria di uno stimatore T. Se quest'ultima è approssimativamente Normale (condizione che si verifica di larga massima se la numerosità del campione è maggiore di 30), è probabile che la stessa distribuzione di T sia compresa negli intervalli da $\mu-\sigma_T$ a $\mu+\sigma_T$; da $\mu-2\sigma_T$ a $\mu+2\sigma_T$; da $\mu-3\sigma_T$ a $\mu+3\sigma_T$ rispettivamente circa il 68,27%, il 95,45% e il 99,73% delle volte. Allo stesso modo si troverà o si può essere fiduciosi di trovare μ_T negli intervalli da $T-\sigma_T$ a $T+\sigma_T$; da $T-2\sigma_T$ a $T+2\sigma_T$;$T-3\sigma_T$ a $t+3\sigma_T$ circa il 68,27%, il 95,45% e il 99,73% delle volte. Questi valori rappresentano gli intervalli fiduciari rispettivamente al 68,27%, al 95,45% e al 99,73% per la stima di μ, cioè l'intervallo di valori che permettono di stimare il

parametro incognito della popolazione nel caso di uno stimatore T corretto. Gli estremi degli intervalli si definiscono limiti fiduciari al 68,27%, il 95,45% e il 99,73%. La percentuale di fiducia è spesso detta livello di fiducia. Se si vuole esprimere in termini formali la costruzione di un intervallo di stime accettabili, ad esempio per la media, si ipotizza che la v.c. sia Normale e che, quindi, il suo valore medio sia distribuito normalmente con media incognita μ e varianza σ^2. Si voglia stimare, appunto, tale media.

La v.c. Z è data da:

$$Z = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sigma} \qquad (11.2.1)$$

e si distribuisce come una v.c. standardizzata N~(0;1).

Con la notazione seguente:

$$P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha \qquad (11.2.2)$$

si esprime la probabilità che, ad un livello di fiducia o confidenza 1-α ovvero nel 100(1-α)% dei campioni, la media μ della popolazione è ricompresa negli estremi

$$[\ \overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \text{ e } \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\ ],$$

che può essere denotata in modo equivalente come:

$$P(\mu_1 \leq \mu \leq \mu_2) = 1-\alpha \qquad (11.2.1)$$

dove 1-α è appunto il livello di fiducia o confidenza e α è il livello di significatività, ovvero la probabilità di compiere un errore qualora si affermi che il valore del parametro della popolazione di interesse sia compreso nei limiti(μ_1, μ_2). Il livello di fiducia o confidenza 1-α è quell'intervallo di valori campionario che dovrebbe contenere il valore del parametro della popolazione di interesse ad un prefissato livello di significatività α . Nel modello inferenziale esso rappresenta uno strumento attraverso il quale è possibile dare un giudizio di affidabilità sulla stima dei parametri della popolazione. Con un esempio si può facilmente verificare quanto sopra descritto.
Si consideri il caso in cui il valore medio o atteso di una distribuzione campionaria si differenzi dalla media della popolazione μ di ±1,96 σ. Se si stabilisce un livello di confidenza pari a 0,95 (95%), a cui ne corrisponde uno di significatività pari a