CORSO DI LAUREA IN STATISTICA
Statistica per le decisioni (Note didattiche)
Bruno Chiandotto

CAP. 3 - CAMPIONI CASUALI e DISTRIBUZIONI CAMPIONARIE

3.1 Introduzione

Nel capitolo introduttivo di queste note si è avuto modo di distinguere la **statistica descrittiva** dalla **statistica induttiva (inferenza statistica)** sottolineando che si opera nel primo ambito quando si dispone di tutte le manifestazioni del fenomeno d'interesse, in tali circostanze la statistica (descrittiva) si risolve in un insieme di metodi che consentono una *compattazione* adeguata delle informazioni disponibili per rendere possibile o, quantomeno, facilitare la comprensione degli aspetti del fenomeno che più interessano (a fini conoscitivi e/o decisionali).

Se per qualche motivo (perché impossibile o perché non conveniente) non si dispone di tutte le manifestazioni del fenomeno ma soltanto di un sottoinsieme di queste, si dispone cioè di un **campione** di manifestazioni del fenomeno d'interesse, la statistica (induttiva) si caratterizza come l'insieme delle teorie e dei metodi che consentono di pervenire, utilizzando i dati campionari, a delle conclusioni che siano "*il più vicino possibile*" a quelle cui si sarebbe pervenuti disponendo di tutte le manifestazioni del fenomeno.

3.2 Campioni casuali

Se con P si indica l'insieme di tutte le possibili manifestazioni del fenomeno di interesse e con C_p un suo sottoinsieme ($C_p \subset P$), operando su C_p si vogliono trarre conclusioni *valide* per P, si vuole, cioè, inferire da C_p a P.

Se è vero che un campione C_p è un qualunque sottoinsieme di P, si deve sottolineare che la statistica (induttiva) tratta in modo quasi esclusivo dei **campioni**

casuali (campioni probabilistici), cioè, dei sottoinsiemi C_p di P cui si perviene attraverso l'applicazione di un qualche meccanismo di selezione avente natura probabilistica. Non costituisce, quindi, parte integrante della statistica (induttiva) l'analisi dei campioni non probabilistici; rientrano in quest'ultima categoria i cosiddetti **campioni ragionati** e quelli per i quali non è noto il meccanismo generatore.

E' campionamento ragionato quello che individua le unità campionarie, cioè le unità statistiche portatrici delle informazioni (manifestazioni del fenomeno d'interesse), attraverso l'applicazione di procedure basate sull'*impiego ragionato* dell'informazione disponibile al momento in cui si procede all'individuazione delle unità che andranno a costituire il campione. In proposito si deve, comunque, sottolineare che le informazioni disponibili costituiscono spesso la base di schemi di campionamento probabilistico più o meno complessi (campionamento stratificato, campionamento a grappolo, campionamento a più stadi, campionamento stratificato a più stadi, ecc.), ma in tali circostanze le informazioni disponibili vengono utilizzate solo per incrementare l'*efficienza* del campione, cioè l'efficienza del processo di induzione dal campione alla popolazione, e non per individuare le singole unità che andranno a costituire il campione.

In questa sede si tratterà esclusivamente del **campionamento casuale semplice**; cioè, dei campioni cui si perviene procedendo all'estrazione (con o senza ripetizione) di n (**dimensione del campione**) elementi che hanno la stessa probabilità di essere inclusi nel campione.

Nell'ambito del campionamento semplice si ipotizzerà sempre (almeno a livello teorico) l'esistenza di un modello probabilistico capace di rappresentare adeguatamente il fenomeno che interessa analizzare. In altre parole, si assumerà che la popolazione P sia rappresentata da una variabile casuale semplice o multipla con una propria funzione di distribuzione non completamente nota. Ovviamente, se la funzione di distribuzione fosse completamente nota si tornerebbe al caso di disponibilità completa di tutte le possibili manifestazioni del fenomeno d'interesse.

Se si fa riferimento al caso univariato (ed è quello considererato in queste note) la situazione di riferimento è quella di una variabile casuale X con funzione di distribuzione $F(x\,;\,\theta_1,\theta_2,....,\theta_k)=F(x\,;\,\underline{\theta})$, dove $(\theta_1,\theta_2,....,\theta_k)=\underline{\theta}$ è l'insieme

(vettore) dei parametri caratteristici del modello definiti nello **spazio parametrico** Θ_k $(\underline{\theta} \in \Theta_k)$; cioè, dei parametri che caratterizzano lo specifico modello, rappresentativo della specifica situazione reale, nell'ambito della famiglia di distribuzioni espressa dalla funzione $F(\cdot,\cdot)$.

Se, come avviene usualmente, si considera la funzione di massa (caso discreto) o di densità (caso continuo) di probabilità della variabile casuale X, si dirà che si sta trattando della variabile casuale semplice X con funzione di massa o di densità di probabilità $f(x;\theta_1,\theta_2,....,\theta_k) = f(x;\underline{\theta})$.

Si è detto che esiste un problema di induzione statistica quando la funzione di distribuzione $F(\cdot,\cdot)$ non è completamente nota; ovviamente, tale affermazione vale anche nei confronti della funzione $f(\cdot,\cdot)$. In proposito si possono distinguere almeno due situazioni di mancanza di conoscenza: la prima situazione è quella caratterizzata da una conoscenza parziale della funzione $f(x;\theta_1,\theta_2,....,\theta_k) = f(x;\underline{\theta})$ nel senso che si conosce la forma analitica della funzione ma non si conosce il valore di tutti o di alcuni parametri caratteristici della funzione stessa, in questa circostanza si parla di **inferenza statistica parametrica**. La seconda situazione è quella d'ignoranza completa: non si conosce ne il valore dei parametri ne la forma analitica della funzione di massa o di densità di probabilità; in questa circostanza si parla di **inferenza statistica non parametrica.** Una terza situazione, intermedia rispetto alle due precedenti, è quella in cui si specificano certe componenti del modello (ad esempio si suppone che la v.c. appartenga alla famiglia esponenziale ma non si specifica la sottofamiglia: forma funzionale della funzione di massa o di densità). Se si opera in tale contesto si parla di **inferenza statistica semi-parametrica**, nel senso che il modello statistico per l'analisi del fenomeno è specificato solo parzialmente.

Da sottolineare che la dizione inferenza statistica non parametrica non è certamente la più appropriata in quanto interpretabile come se, in questo ambito, le procedure di statistica induttiva non riguardassero i parametri. Ovviamente, questa interpretazione è fuorviante, infatti, con la dizione "non paramatrica" si vuole, molto semplicemente, caratterizzare le situazioni inferenziali nelle quali non si conosce forma analitica e valore dei parametri caratteristici, elementi questi entrambi coinvolti nelle

procedure inferenziali: La dizione corretta per caratterizzare tali situazioni è quella di inferenza statistica libera da distribuzione (***distribution free***).

E' già stato sottolineato che in queste note si parlerà. in modo quasi esclusivo, di campionamento probabilistico semplice, in realtà il limite è ancora più rigido; infatti, la trattazione sarà limitata al campionamento semplice con ripetizione **(campionamento bernoulliano**), in questo contesto le variabili casuali associate a ciascuna unità campionaria risultano **indipendenti e identicamente distribuite** (**i.i.d**.). Al riguardo si deve, comunque, segnalare che nelle situazioni reali il campionamento che si realizza è quello esaustivo (senza ripetizione), ma è anche vero che nella generalità dei casi le differenze tra i due schemi di campionamento diventa operativamente irrilevante avendo a che fare con popolazioni di dimensione molto elevate, dimensione che diventa infinita nel caso di variabili casuali continue. Tale motivazione giustifica la trattazione del campionamento bernoulliano molto più semplice dal punto di vista analitico.

Definizione 1 Se $X_1, X_2,...,X_n$ costituiscono un insieme di variabili casuali indipendenti e identicamente distribuite (i.i.d.), la loro funzione di massa o di densità di probabilità congiunta soddisfa l'uguaglianza

$$f(x_1,x_2,....,x_n\ ;\ \theta_1,\theta_2,....,\theta_k) = f(\underline{x}\ ;\ \underline{\theta}) =$$

$$= f(x_1\ ;\ \underline{\theta}) \cdot f(x_2\ ;\ \underline{\theta}) \cdot ....\cdot f(x_i\ ;\ \underline{\theta}) \cdot ....\cdot f(x_n\ ;\ \underline{\theta}) = \prod_{i=1}^{n} f(x_i\ ;\ \underline{\theta})$$

allora si dice che l'insieme di variabili casuali i.i.d. $X_1, X_2,...,X_n$ =costituisce un **campione casuale semplice** di n osservazioni indipendenti relativo alla variabile casuale X che ha funzione di massa o di densità di probabilità equivalente a quella (comune) di ciascuna componente X_i del campione. Il **punto campionario** $\underline{X} = (X_1, X_2,..., X_n)$ è definito nello **spazio o universo dei campioni** ad n dimensioni C $(\underline{X} \in C)$.

Nella formula sopra riportata con $f(x_i\ ;\ \underline{\theta})$, per $i = 1, 2,..,n$, si è indicata la funzione di massa, o di densità di probabilità, dell'i-esimo elemento costituente il

campione. Avendo supposto l'indipendenza tra le osservazioni campionarie, si avrà, come sottolineato, l'uguaglianza (equivalenza) tra la distribuzione della variabile casuale X relativa alla popolazione e la variabile X_i (tale deve essere intesa a priori, cioè prima dell'effettiva estrazione del campione) relativa all'i-esimo elemento campionario (per $i = 1, 2 \ldots, n$).

Dalla definizione risulta che se, ad esempio, si volesse estrarre un campione di n elementi da una popolazione distribuita normalmente, con media μ e varianza σ^2, la funzione di densità di probabilità del campione casuale è

$$f(x_1, x_2,\ldots,x_n) = f(x_1, x_2,\ldots,x_n; \mu,\sigma^2) = \prod_{i=1}^{n} f(x_i; \mu,\sigma^2) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} = \frac{1}{\left(2\mu\sigma^2\right)^{n/2}} \cdot e^{-\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(x_i-\mu)^2}$$

Se l'estrazione del campione di n elementi riguardasse una popolazione poissoniana caratterizzata dal parametro λ, la funzione di massa di probabilità del campione casuale è

$$f(x_1,x_2, \ldots, x_n) = f(x_1,x_2, \ldots, x_n; \lambda) =$$

$$= \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

Alle due funzioni $f(x_1,x_2, \ldots, x_n; \lambda)$ e $f(x_1, x_2,\ldots,x_n; \mu,\sigma^2)$ sopra riportate e, in generale, ad ogni funzione di massa o di densità di probabilità campionaria

$$f\left(x_1, x_2,\ldots,x_i,\ldots,x_n ;\underline{\theta}\right) = \prod_{i=1}^{n} f\left(x_i ;\underline{\theta}\right)$$

dove $\underline{\theta}$ rappresenta uno o più parametri caratteristici della distribuzione di riferimento, può essere associata una seconda interpretazione che introduce nella trattazione un concetto di estrema rilevanza: la **funzione di verosimiglianza**. Si tratta di una funzione del tutto equivalente, in termini formali, alla funzione di massa o di densità di probabilità campionaria sopra introdotta, ma che da questa si diversifica sostanzialmente. Infatti, la **funzione**

$$f\left(x_1, x_2,\ldots,x_i,\ldots,x_n ;\underline{\theta}\right) = \prod_{i=1}^{n} f\left(x_i ;\underline{\theta}\right)$$

viene detta **di verosimiglianza** se la si interpreta come funzione del parametro (o dei

parametri) $\underline{\theta}$ per un campione prefissato e non come funzione degli elementi campionari. Per evidenziare questa particolare interpretazione si può rappresentare algebricamente la funzione di verosimiglianza con l'espressione

$$L(\underline{\theta}) = L(\underline{\theta} / \underline{X} = \underline{x}) = \prod_{i=1}^{n} f(\underline{\theta} / x_1, x_2, ..., x_n)$$

dove $\underline{X} = (X_1, X_2, ..., X_n)$ rappresenta la variabile casuale ad n dimensioni (vettore casuale) associata alle n rilevazioni campionarie, mentre $\underline{x} = (x_1, x_2, ..., x_n)$ rappresenta il punto campionario, cioè una specifica determinazione del vettore casuale $\underline{X}$, definito nello spazio o universo dei campioni a n dimensioni C..

Pertanto, nella prima interpretazione, la funzione

$$f(x_1, x_2, ..., x_i, ..., x_n ; \underline{\theta}) = \prod_{i=1}^{n} f(x_i ; \underline{\theta})$$

fa riferimento all'universo dei campioni, si tratta, come già sottolineato, di un riferimento a priori, cioè prima dell'effettiva estrazione del campione. In questo contesto, le variabili che interessano sono, appunto, $X_1, X_2, ..., X_n$, associate a ciascun punto campionario.

Nella seconda interpretazione, la variabile di riferimento è il parametro, o il vettore dei parametri incognito $\underline{\theta}$, in quanto si assume l'avvenuta estrazione campionaria delle unità statistiche di osservazione e le variabili associate a ciascuna unità (punto campionario) hanno assunto una specifica determinazione, sono cioè delle costanti note, mentre assume la natura di variabile $\underline{\theta}$ (parametro o vettore dei parametri) essendo tale entità un'incognita del problema.

Esempio 1

Si consideri una popolazione bernoulliana (variabile casuale di bernoulli X che può assumere i due valori 0, assenza del carattere, ed 1, presenza del carattere) con parametro caratteristico $\theta = p$ e si supponga che da tale popolazione si voglia procedere all'estrazione di $n = 6$, $n = 12$ ed $n = 36$ unità campionarie rimettendo ogni volta l'unità estratta nella popolazione (campionamento bernoulliano). In tali situazioni la funzione di massa di probabilità è quella sotto riportata

$$f(x_1,x_2,\ldots,x_n;p) = \prod_{i=1}^{n} f(x_i;p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}$$

dove basterà sostituire ad n i valori 6, 12 e 36.

Se si procede alla rilevazione campionaria nei tre casi sopra considerati e le sequenze osservate sono, rispettivamente:

- (1,0,1,1,1,1) per n = 6 (x=5);

- (1,1,0,1,1,1,1,1,1,1,0,1) per n = 12 (x=10);

- (0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,0,0,1,1,1,1,1,1,0,1,0,1,1) per n = 36 (x=30).

Le funzioni di verosimiglianza sono :

$$L(p) = p^5 (1-p)^1 \quad per \quad 0 \le p \le 1$$
$$L(p) = p^{10} (1-p)^2 \quad per \quad 0 \le p \le 1$$
$$L(p) = p^{30} (1-p)^6 \quad per \quad 0 \le p \le 1$$

la cui rappresentazione grafica è riportata alla pagina successiva.

Osservando la figura si rileva in modo molto evidente la tendenza alla normalità della funzione di verosimiglianza al crescere della dimensione campionaria.

Per avere una più immediata comprensione sia dei metodi statistici che verranno trattati in seguito, sia delle loro proprietà, occorre sempre tenere presente la distinzione fra variabile casuale e le determinazioni (valori osservati) della variabile casuale stessa: prima di effettuare l'estrazione campionaria gli elementi costituenti il campione X_1, $X_2,\ldots,X_n$, sono variabili casuali; infatti, l'elemento generico X_i ($i = 1, 2 \ldots, n$) ha, come già sottolineato, una struttura del tutto analoga a quella della variabile casuale X, ha cioè la stessa funzione di distribuzione. Dopo aver osservato i risultati campionari, le quantità $x_1,x_2,\ldots,x_n$, costituiscono particolari determinazioni della variabile casuale X.

Poiché gli elementi costituenti un campione sono delle variabili casuali, è variabile casuale anche ogni funzione $T(X_1,X_2,\ldots,X_n)$ non costante degli stessi. Tale funzione, che non dipende dai parametri incogniti $\theta_1, \theta_2, \ldots, \theta_k$, viene usualmente detta **statistica** (dall'inglese **statistic**). Sarà, quindi, possibile derivare la funzione di massa o di densità di probabilità di tale variabile in funzione della distribuzione di massa o di densità di probabilità delle variabili casuali associate ai singoli elementi campionari.

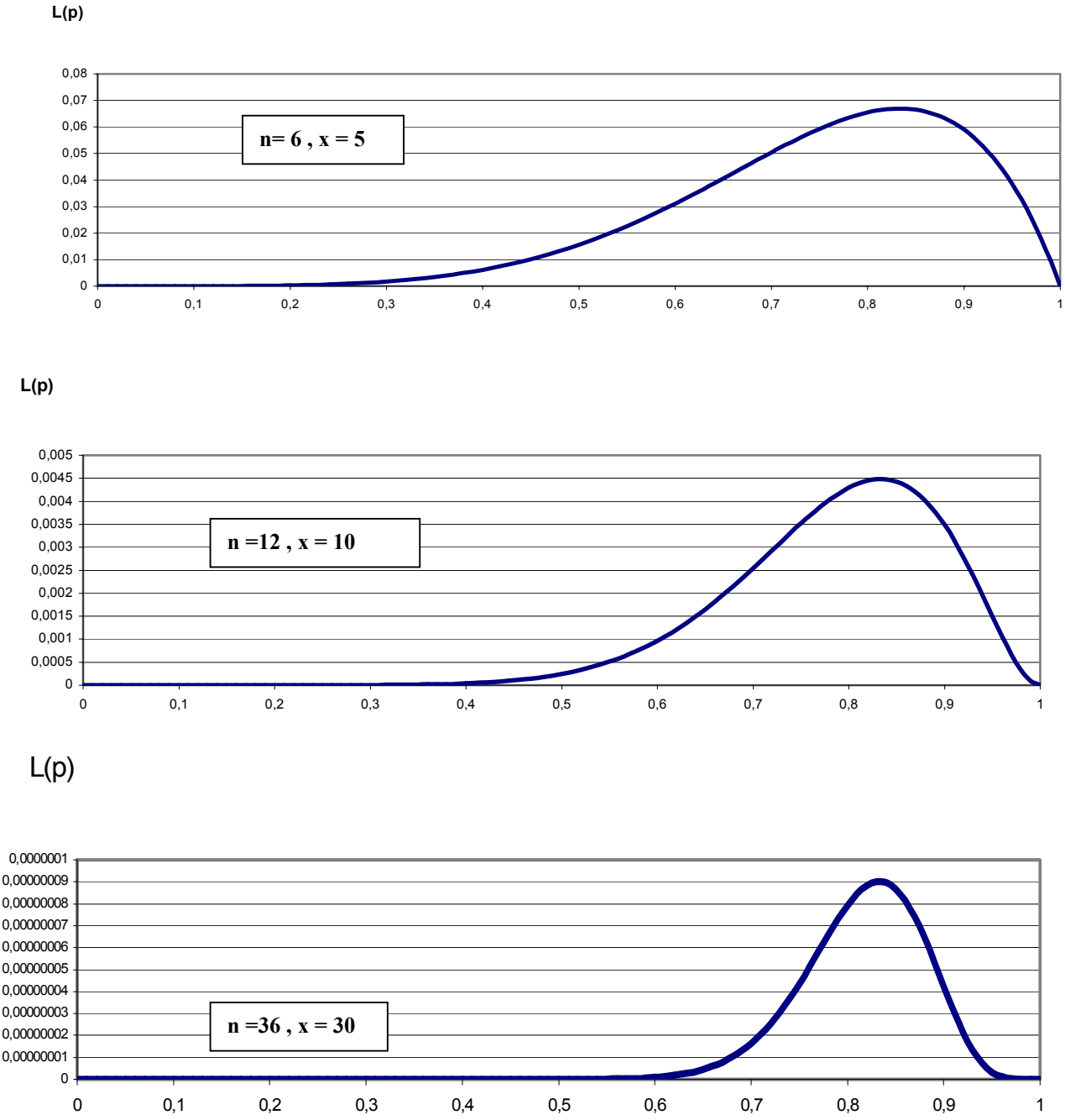

Fig. 1 – Funzione di verosimiglianza relativa a n prove senza ripetizione effettuate su una popolazione in cui ciascuna unità è caratterizzata dalla presenza o assenza di uno specifico carattere

3.3 Momenti campionari e distribuzioni campionarie

Definizione 2 Si dice **distribuzione campionaria**, ogni distribuzione di probabilità che evidenzia la relazione esistente tra i possibili valori che possono essere assunti (**nell'universo dei campioni**) da una qualsiasi funzione $T(X_1,X_2,\ldots,X_n)$ (ad es. un indice sintetico) applicata agli n elementi campionari (casuali) e la distribuzione di massa o di densità di probabilità associata agli n elementi costituenti il campione stesso.

Si consideri la funzione, definita sugli elementi $X_1, X_2,\ldots,X_n$, di un campione casuale semplice con ripetizione relativo ad una certa variabile X che ha momento s-esimo ($s = 1,2,3,\ldots$) pari a μ_s e varianza pari a σ^2:

$$\overline{X}_s = T_s(X_1,X_2,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^s \quad ; \quad s=1,2,\ldots$$

che viene usualmente detto **momento campionario, o momento empirico, di ordine s rispetto all'origine**. Evidentemente tale momento, varierà al variare del campione e descriverà una variabile casuale, la cui funzione di massa o di densità di probabilità dipenderà dalla funzione di massa o di densità di probabilità delle variabili casuali $X_1,X_2,\ldots,X_n$, e qundi, dalla funzione di massa o di densità di probabilità della variabile casuale X.

È facile verificare che il valore medio di $\overline{X}_s$ è pari al momento s-esimo della variabile X, infatti

$$E\left(\overline{X}_s\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^s\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^s\right) = E\left(X^s\right) = \mu_s$$

e quindi, per $s=1$ si avrà

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X\right) = E(X) = \mu_1 = \mu$$

cioè il valor medio della media campionaria è uguale alla media della popolazione.

La varianza della media campionaria è data da

$$Var\left(\overline{X}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) = \frac{\sigma^2}{n} = \sigma_{\bar{x}}^2$$

cioè, la varianza della media campionaria è pari alla varianza della popolazione divisa per la dimensione del campione.

Nel caso di campionamento semplice esaustivo (senza ripetizione) si ha:

$$\mathrm{V}\,ar\left(\overline{X}\right) = \sigma_{\bar{x}}^2 = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} Var\left(X_i\right) + \sum_{i=1}^{n}\sum_{i\neq j} Cov\left(X_i, X_j\right)\right)$$

$$= \frac{1}{n^2}\left(n\cdot\sigma^2 + n\cdot(n-1)\cdot\sigma^*\right) = \frac{\sigma^2}{n} + \frac{(n\text{-}1)\cdot\sigma^*}{n}$$

dove $\cdot\sigma^* = Cov\left(X_i, X_j\right)$ per ogni *i,j*. Se si assume $n=N$, si ha:

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n} + \frac{(n\text{-}1)\cdot\sigma^*}{n} = 0$$

da cui $\tilde{\sigma} = -\frac{\sigma^2}{N-1}$ che sostituito nella precedente espressione da

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n} - \frac{(n\text{-}1)\cdot\sigma^2}{N-1} = \frac{\sigma^2}{n}\cdot\frac{N\text{-}n}{N\text{-}1}$$

dove $\frac{N-n}{N-1}$ viene usualmente detto **fattore di correzione** e fornisce, come si avrà modo di chiarire successivamente, una misura della maggiore efficienza del campionamento esaustivo rispetto al campionamento con ripetizione.

Definendo la **varianza campionaria** (**corretta**) attraverso l'espressione:

$$S^2 = T(X_1, X_2, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

si può verificare, nell'ipotesi di campionamento bernoulliano (campione casuale semplice con ripetizione), che $E(S^2)$ è uguale a σ^2, cioè il valor medio della varianza campionaria è pari alla varianza della popolazione. Mentre la varianza della varianza campionaria corretta S^2 è espressa da:

$$Var\left(S^2\right) = \frac{1}{n}\left(\overline{\mu}_4 - \frac{n-3}{n-1}\sigma^4\right)$$

dove $\overline{\mu}_4$ rappresenta il momento quarto rispetto alla media della popolazione ($\overline{\mu}_4 = E\{(X-\mu)^4\}$) dalla quale viene estratto il campione.

Esempio 2 *(distribuzioni campionarie per campioni estratti da popolazioni discrete)*

Si considerino cinque palline identiche a meno dei contrassegni numerici (1, 3, 5, 7, 9) che su di esse sono riportati. La distribuzione di frequenza relativa alla variabile in questione può essere rappresentata nel modo seguente

Modalità x_i	*Frequenze assolute* n_i	*Frequenze relative* $f_i=n_i/n$ *(probabilità:* p_i*)*
1	*1*	*1/5*
3	*1*	*1/5*
5	*1*	*1/5*
7	*1*	*1/5*
9	*1*	*1/5*

Tab. 1 Popolazione discreta uniforme

Si supponga di aver estratto 100 campioni casuali, di dimensione n = 2, dalla popolazione riportata nella Tab.1 e che l'operazione di campionamento (effettuata reinserendo ogni volta l'unità estratta nella popolazione) abbia dato luogo alle 100 coppie di risultati riportati nella Tab. 2.

Se per ogni coppia di risultati campionari si procede al computo della media

$$\overline{X} = M_1 = T(X_1, X_2) = \frac{X_1 + X_2}{2}$$

dove (X_1, X_2) *rappresenta la coppia degli elementi costituenti il campione, si potrà derivare la distribuzione campionaria sperimentale (relativa ai 100 campioni estratti) della media aritmetica che sono riportati nella Tab. 3 ; dove, evidentemente, la frequenza assoluta* n_i *sta ad indicare il numero dei campioni (su 100 estratti) di due elementi per il quale si è realizzata quella particolare modalità* $\overline{x}_i$ *(media aritmetica dei due elementi campionari).*

N.	*Campione*	*N.*	*Campione*	*N.*	*Campione*	*N.*	*Campione*	*N.*	*Campione*
1	*(3,3)*	*21*	*(5,3)*	*41*	*(3,7)*	*61*	*(5,1)*	*81*	*(1,9)*
2	*(5,3)*	*22*	*(9,3)*	*42*	*(1,7)*	*62*	*(3,5)*	*82*	*(3,7)*
3	*(1,1)*	*23*	*(5,9)*	*43*	*(5,7)*	*63*	*(3,1)*	*83*	*(9,3)*
4	*(7,3)*	*24*	*(7,3)*	*44*	*(7,7)*	*64*	*(7,7)*	*84*	*(9,1)*
5	*(1,5)*	*25*	*(5,5)*	*45*	*(1,9)*	*65*	*(1,1)*	*85*	*(5,9)*
6	*(3,5)*	*26*	*(9,9)*	*46*	*(3,3)*	*66*	*(9,7)*	*86*	*(5,3)*
7	*(5,5)*	*27*	*(9,5)*	*47*	*(3,7)*	*67*	*(1,3)*	*87*	*(1,9)*
8	*(5,7)*	*28*	*(9,7)*	*48*	*(3,1)*	*68*	*(9,5)*	*88*	*(9,5)*
9	*(9,3)*	*29*	*(7,3)*	*49*	*(1,1)*	*69*	*(3,5)*	*89*	*(1,9)*
10	*(3,3)*	*30*	*(3,7)*	*50*	*(1,7)*	*70*	*(9,7)*	*90*	*(5,5)*
11	*(5,7)*	*31*	*(3,1)*	*51*	*(1,5)*	*71*	*(9,7)*	*91*	*(9,3)*
12	*(7,3)*	*32*	*(5,5)*	*52*	*(9,1)*	*72*	*(1,3)*	*92*	*(1,1)*
13	*(3,7)*	*33*	*(9,1)*	*53*	*(7,7)*	*73*	*(1,5)*	*93*	*(3,3)*
14	*(3,3)*	*34*	*(5,9)*	*54*	*(7,3)*	*74*	*(7,1)*	*94*	*(1,3)*
15	*(1,7)*	*35*	*(5,9)*	*55*	*(5,9)*	*75*	*(3,5)*	*95*	*(5,1)*
16	*(5,9)*	*36*	*(9,1)*	*56*	*(3,5)*	*76*	*(5,5)*	*96*	*(1,5)*
17	*(9,1)*	*37*	*(3,1)*	*57*	*(9,7)*	*77*	*(3,5)*	*97*	*(1,5)*
18	*(3,9)*	*38*	*(7,1)*	*58*	*(5,7)*	*78*	*(9,5)*	*98*	*(7,1)*
19	*(7,3)*	*39*	*(7,7)*	*59*	*(5,1)*	*79*	*(7,1)*	*99*	*(7,1)*
20	*(7,5)*	*40*	*(7,9)*	*60*	*(1,3)*	*80*	*(9,5)*	*100*	*(3,5)*

Tab. 2 Prospetto dei risultati relativi a 100 campioni di dimensione 2, estratti casualmente dalla popolazione riportata nella Tab. 1

Media campionaria $M_1 = \bar{x}_i$	*Frequenza assoluta* n_i	*Frequenza relativa* $f_i = n_i/100$
1	*4*	*0,04*
2	*8*	*0,08*
3	*13*	*0,13*
4	*18*	*0,18*
5	*25*	*0,25*
6	*10*	*0,10*
7	*15*	*0,15*
8	*6*	*0,06*
9	*1*	*0,01*

Tab. 3 Distribuzione campionaria sperimentale della media aritmetica relativa ai risultati riportati nella tab. 2

*La **distribuzione campionaria sperimentale** della variabile riportata nella Tab. 3 costituisce una approssimazione della distribuzione campionaria (teorica) di $\overline{X}$. Se si procedesse all'estrazione di una seconda serie di 100 campioni, di dimensione 2, si*

otterrebbe una diversa distribuzione campionaria sperimentale di $\overline{X}$*, tale da costituire anche essa un'approssimazione della distribuzione campionaria teorica di* $\overline{X}$*. Considerando le due serie di esperimenti ad un tempo (cioè 200 campioni di dimensione 2) si dovrebbe ottenere una distribuzione campionaria sperimentale di* $\overline{X}$ *più vicina alla distribuzione teorica di quanto non siano le due distribuzioni considerate separatamente.*

Per determinare la ***distribuzione campionaria teorica*** *della variabile casuale* $\overline{X}$ *si può seguire la via sotto indicata.*

a) - Si considerano tutte le possibili coppie di valori (X_1, X_2) *estraibili (con ripetizione) dalla popolazione riportata nella Tab. 1, che sono*

(1,1)	(3,1)	(5,1)	(7,1)	(9,1)
(1,3)	(3,3)	(5,3)	(7,3)	(9,3)
(1,5)	(3,5)	(5,5)	(7,5)	(9,5)
(1,7)	(3,7)	(5,7)	(7,7)	(9,7)
(1,9)	(3,9)	(5,9)	(7,9)	(9,9)

e su queste coppie di valori vengono calcolate le medie aritmetiche;

b) - Si determina la probabilità relativa a ciascuna coppia (X_1, X_2)*. Essendo il campione estratto con ripetizione da una popolazione uniforme si avrà*

$$P\left[\left(X_1 = x_1\right) \cap \left(X_2 = x_2\right)\right] = P\left(X_1 = x_1\right) \cdot P\left(X_2 = x_2\right) = \frac{1}{25} \quad \text{per i, j = 1, 2, 3, 4, 5,}$$

c) - Si sommano le probabilità relative alle coppie di valori che danno luogo alla stessa media.

Il risultato delle operazioni indicate ai punti a), b), c), possono essere riassunti nella tabella seguente

Modalità $\overline{x}_i$	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Probabilità $f(\overline{x}_i) = p_i$	*0,04*	*0,08*	*0,12*	*0,16*	*0,20*	*0,16*	*0,12*	*0,08*	*0,04*

Tab. 4 - Distribuzione campionaria (teorica) della media aritmetica per campioni di dimensione 2 estratti dalla popolazione uniforme riportata nella Tab. 1

Il confronto tra i dati relativi alla distribuzione campionaria teorica e quelli

relativi alla distribuzione campionaria empirica è riportato nella figura seguente

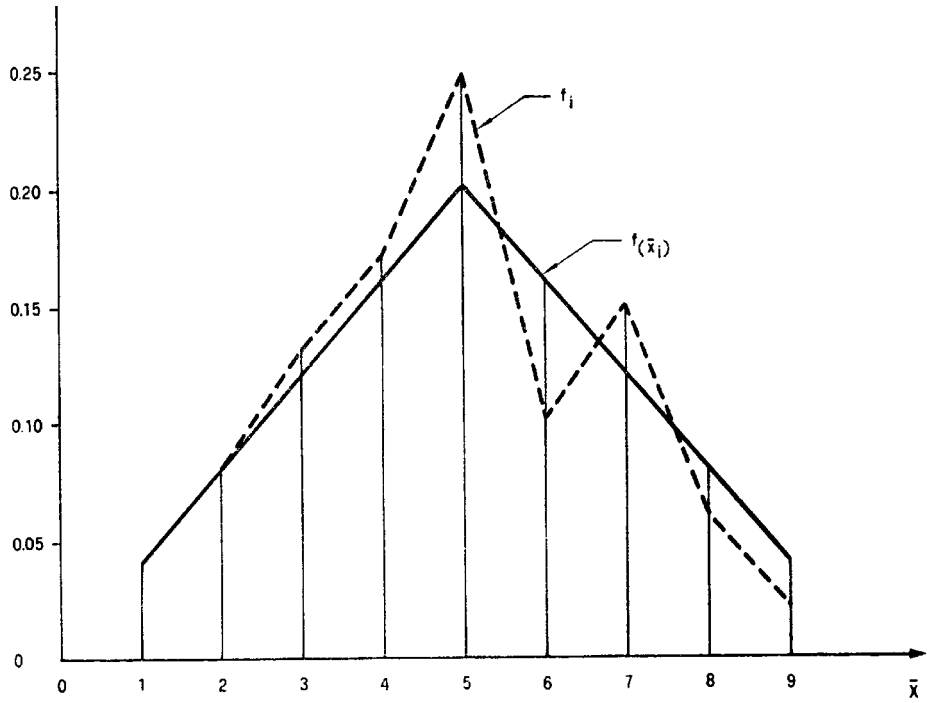

Fig. 2 - Distribuzione campionaria sperimentale (come da Tab. 3) e teorica (come da Tab. 4) per la media calcolata su campioni di dimensione 2 estratti dalla popolazione uniforme X: 1, 3, 5, 7, 9

Utilizzando i dati riportati nella Tab. 4 si derivano le uguaglianze

$$E\left(\overline{X}\right)=\mu = 5$$

$$Var\left(\overline{X}\right)=\sigma^2_{\overline{x}} = \frac{\sigma^2}{2} = 4$$

che verificano empiricamente la proprietà che ha il valor medio (valore atteso) della variabile casuale media campionaria $\overline{X}$ *di essere uguale al valor medio (media aritmetica) della variabile casuale relativa all'intera popolazione e della varianza che risulta essere pari alla varianza della popolazione divisa per la numerosità del campione.*

Esempio 3 *(distribuzioni campionarie per campioni estratti da popolazioni discrete)*

Si considerino 6 palline identiche a meno dei numeri su di esse riportati: {①, ①, ①, ③, ③, ⑨}. La funzione di massa della v.c. X = "risultato dell'estrazione di una pallina" è allora data da

$$f(x) = \begin{cases} 1/2 & x = 1 \\ 1/3 & x = 3 \\ 1/6 & x = 9 \\ 0 & altrimenti \end{cases}$$

Per tale v.c. è facile derivare i principali momenti. Il seguente prospetto riassume il calcolo di $\mu = E(X) = 3$ *e* $\sigma^2 = V(X) = E(X^2) - E(X)^2 = 17 - 3^2 = 8$.

x	$f(x)$	$x\,f(x)$	$x^2\,f(x)$
1	1/2	1/2	1/2
3	1/3	1	3
9	1/6	3/2	27/2
	1	3	17

Tab. 5 – Prospetto di calcolo di E(X) e V(X).

Si considerino ora tutti i possibili campioni $\underline{x} = (x_1, x_2)$ *di dimensione n = 2 che possono essere estratti con reimmissione dalla v.c. in oggetto. La "lista" di questi campioni forma l'**universo dei campioni** che possono essere estratti dalla v.c. X. L'universo dei campioni può a sua volta essere rappresentato dalla v.c. doppia* $\underline{X} = (X_1, X_2)$*, i cui valori e la cui distribuzione sono riportati nella Tab. 6 (la probabilità di ciascuna coppia è semplicemente il prodotto delle probabilità dei singoli, dato che le estrazioni sono indipendenti).*

$\underline{x} = (x_1, x_2)$	(1,1)	(1,3)	(1,9)	(3,1)	(3,3)	(3,9)	(9,1)	(9,3)	(9,9)	tot
$f(\underline{x})$	1/4	1/6	1/12	1/6	1/9	1/18	1/12	1/18	1/36	1

Tab. 6 – Funzione di massa della v.c. doppia $\underline{X} = (X_1, X_2)$.

Qualunque statistica calcolata su $\underline{X} = (X_1, X_2)$ *è una v.c. e ha di conseguenza una sua **distribuzione campionaria**.*

***Media campionaria**:* $\overline{X} = (X_1 + X_2)/2$

La seguente tabella riporta, per ogni campione, la relativa media campionaria con la sua probabilità

$\underline{x} = (x_1, x_2)$	(1,1)	(1,3)	(1,9)	(3,1)	(3,3)	(3,9)	(9,1)	(9,3)	(9,9)	tot
$f(\underline{x})$	1/4	1/6	1/12	1/6	1/9	1/18	1/12	1/18	1/36	1
$\bar{x}$	1	2	5	2	3	6	5	6	9	

Tab. 7 – Prospetto per la costruzione della funzione di massa della media campionaria.

La funzione di massa della media campionaria è riportata nella tabella seguente

$\bar{x}$	1	2	3	5	6	9	tot
$f(\bar{x})$	1/4	1/3	1/9	1/6	1/9	1/36	1

Tab. 8 – Funzione di massa della media campionaria $\overline{X}$ *.*

Si può verificare che $E(\overline{X}) = 3$ *e* $V(\overline{X}) = 4$*.*

Varianza campionaria corretta*:* $S^2 = [(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2]/(2 - 1)$

La tabella che segue riporta, per ogni campione, i valori assumili dalla varianza campionaria corretta con le relative probabilità

$\underline{x} = (x_1, x_2)$	(1,1)	(1,3)	(1,9)	(3,1)	(3,3)	(3,9)	(9,1)	(9,3)	(9,9)	tot
$f(\underline{x})$	1/4	1/6	1/12	1/6	1/9	1/18	1/12	1/18	1/36	1
s^2	0	2	32	2	0	18	32	18	0	

Tab. 9 – Prospetto per la costruzione della funzione di massa della varianza campionaria corretta.

Da tale tabella si ricava facilmente la funzione di massa della varianza campionaria corretta, nella quale si sommano le probabilità relative alle coppie di valori uguali.

s^2	0	2	18	32	tot
$f(s^2)$	7/18	1/3	1/9	1/6	1

Tab. 10 – Funzione di massa della varianza campionaria corretta S^2*.*

Utilizzando un prospetto di calcolo simile a quello utilizzato per calcolare i momenti di f(x), si può verificare che $E(S^2) = 8$ *e* $V(S^2) = 144$*.*

***Minimo campionario**: $x_{(1)} = min\{X_1, X_2\}$*

La seguente tabella riporta, per ogni campione, il relativo minimo campionario con la sua probabilità

$\underline{x} = (x_1, x_2)$	(1,1)	(1,3)	(1,9)	(3,1)	(3,3)	(3,9)	(9,1)	(9,3)	(9,9)	tot
$f(\underline{x})$	1/4	1/6	1/12	1/6	1/9	1/18	1/12	1/18	1/36	1
$x_{(1)}$	1	1	1	1	3	3	1	3	9	

Tab. 11 – Prospetto per la costruzione della funzione di massa del minimo campionario.

Da tale tabella si ricava facilmente la funzione di massa del minimo campionario, nella quale si sommano le probabilità relative alle coppie di valori che danno luogo allo stesso minimo.

$x_{(1)}$	1	3	9	tot
$f(x_{(1)})$	3/4	2/9	1/36	1

Tab. 12 – Funzione di massa del minimo campionario $x_{(1)}$.

Si può verificare che $E(x_{(1)}) = 1,6$ e $V(x_{(1)}) = 2,\overline{2}$.

***Massimo campionario**: $x_{(2)} = max\{X_1, X_2\}$*

La tabella seguente riporta, per ogni campione, il relativo massimo campionario con la sua probabilità

$\underline{x} = (x_1, x_2)$	(1,1)	(1,3)	(1,9)	(3,1)	(3,3)	(3,9)	(9,1)	(9,3)	(9,9)	tot
$f(\underline{x})$	1/4	1/6	1/12	1/6	1/9	1/18	1/12	1/18	1/36	1
$x_{(2)}$	1	3	9	3	3	9	9	9	9	

Tab. 13 – Prospetto per la costruzione della funzione di massa del massimo campionario.

Da tale tabella si ricava facilmente la funzione di massa del massimo campionario, nella quale si sommano le probabilità relative alle coppie di valori che danno luogo allo stesso massimo.

$x_{(2)}$	*1*	*3*	*9*	*tot*
$f(x_{(2)})$	*1/4*	*4/9*	*11/36*	*1*

Tab. 14 – Funzione di massa del massimo campionario $x_{(2)}$.

Si può verificare che $E(x_{(2)}) = 4,\overline{3}$ e $V(x_{(2)}) = 10,\overline{2}$.

3.4 Campionamento da popolazioni normali

Per campioni estratti da popolazioni normali vale il seguente teorema:

Teorema 1 Se $X_1, ..,X_n$ costituiscono un campione casuale di elementi relativi ad una popolazione normale, di media μ e varianza σ^2, allora la variabile casuale campionaria:

i) $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

è distribuita normalmente con media μ e varianza σ^2/n;

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

è distribuita come un χ^2 con $g = n$ gradi di libertà;

iii) $$V = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

è distribuita come un χ^2 con $g = (n - 1)$ gradi di libertà.

Dimostrazione

i) - La funzione generatrice dei momenti della v.c $\overline{X}$ è data da

$$m_{\bar{x}}(t) = E\left(e^{\bar{X}t}\right) = E\left(e^{\frac{1}{n}\sum_{101}^{n} X_i t}\right) =$$

$$\left(per\ l'indipendenza\ delle\ v.c.\ X_i\right)$$

$$= \prod_{i=1}^{n} E\left(e^{\frac{1}{n} X_i t}\right) =$$

$$\left(per\ la\ normalità\ delle\ v.c.\ X_i\right)$$

$$= \prod_{i=1}^{n} e^{\frac{1}{n}\mu t + \frac{t^2}{n^2}\sigma^2} = e^{\mu t + t^2 \frac{\sigma^2}{n}}$$

che è la f.g.m. di una v.c. normale di media μ *e* var*ianza* σ^2 / n.

ii) - La funzione generatrice dei momenti della v.c Y è data da

$$m_y(t) = E\left(e^{Yt}\right) = E\left(e^{t\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}\right) =$$

$$\left(per\ l'indipendenza\ delle\ v.c.\ X_i\right)$$

$$= \prod_{i=1}^{n} E\left(e^{t\left(\frac{X_i-\mu}{\sigma}\right)^2}\right) =$$

$$\left(\begin{array}{l} per\ la\ normalità\ delle\ v.c.\ X_i\ e\ ricordando\ che\ il\ quadrato \\ di\ una\ v.c.\ normale\ s\tan dardizzata\ ha\ distribuzione\ \chi_1^2 \end{array}\right)$$

$$= \prod_{i=1}^{n} (1-2t)^{-\frac{1}{2}} = (1-2t)^{-\frac{n}{2}}$$

che è la f.g.m. di una v.c. chi quadro con n *gradi di libertà* (χ_n^2).

iii) - La funzione generatrice dei momenti della v.c Y è data da

$$m_y(t) = E\left(e^{Yt}\right) = (1-2t)^{-n/2} = E\left(e^{t\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2}\right) =$$

$\left(per\ l'indipendenza\ delle\ v.c.\ X_i\right)$

$$= \prod_{i=1}^{n} E\left(e^{t\left(\frac{X_i-\mu}{\sigma}\right)^2}\right) = \prod_{i=1}^{n} E\left(e^{t\left(\frac{X_i-\bar{X}+\bar{X}-\mu}{\sigma}\right)^2}\right) = \prod_{i=1}^{n} E\left(e^{t\left(\frac{X_i-\bar{X}}{\sigma}\right)^2} \cdot e^{t\left(\frac{\bar{X}-\mu}{\sigma}\right)^2}\right)$$

$\left(se\ si\ ipotizza\ l'indipendenza\ tra\ la\ v.c.\ scarto\ X_i-\bar{X}\ \text{e la v.c.}\ \bar{X}-\mu\ si\ ha\right)$

$$= \prod_{i=1}^{n} E\left(e^{t\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}\right) \cdot E\left(e^{t\left(\frac{\bar{X}-\mu}{\sigma}\right)^2}\right) = E\left(e^{t\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}\right) \cdot E\left(e^{t\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2}\right) =$$

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

da cui

$$E\left(e^{t\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}\right) = E\left(e^{tV}\right) = (1-2t)^{-\frac{n-1}{2}}$$

che è la f.g.m. di una v.c. chi quadro con n-1 *gradi di libertà* (χ^2_{n-1}).

Si dimostra ora l'indipendenza tra il vettore delle v.c. scarto $\left[\left(X_1-\bar{X}\right),\left(X_2-\bar{X}\right),.....,\left(X_n-\bar{X}\right)\right]$ e la v.c. $\left(\bar{X}-\mu\right)$.

Si consideri la f.g.m. del vettore casuale a $n+1$ dimensioni $\left[\left(\bar{X}-\mu\right),\left(X_1-\bar{X}\right),\left(X_2-\bar{X}\right),.....,\left(X_n-\bar{X}\right)\right]$

$$m_{(\bar{x}-\mu),(x_1-\bar{x}),(x_2-\bar{x}),.....,(x_n-\bar{x})}\left(t,t_1,t_2,.....,t_n\right) = E\left(e^{\left(\bar{X}-\mu\right)t+\left(X_1-\bar{X}\right)t_1+\left(X_2-\bar{X}\right)t_2+.....+\left(X_n-\bar{X}\right)t_n}\right) =$$

$$= E\left(e^{\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)t+\sum_{i=1}^{n}X_i t_i-\sum_{i=1}^{n}t_i\frac{1}{n}\sum_{j=1}^{n}X_j}\right) = E\left(e^{\sum_{i=1}^{n}\left(\frac{t}{n}+t_i-\bar{t}\right)X_i}\right)\left(dove\ \bar{t}=\frac{1}{n}\sum_{i=1}^{n}t_i\right)$$

$\left(per\ la\ normalità\ e\ l'indipendenza\ delle\ v.c.\ X_i\right)$

$$= e^{\sum_{i=1}^{n}\left[\left(\frac{t}{n}+t_i-\bar{t}\right)\cdot\mu+\frac{\left(\frac{t}{n}+t_i-\bar{t}\right)^2\sigma^2}{2}\right]} = e^{t\mu*\frac{t^2\sigma^2}{2n}} \cdot e^{\sum_{i=1}^{n}(t_i-\bar{t})^2\sigma^2/2}$$

dove $e^{t\mu+\frac{t^2\sigma^2}{2n}}$ *è la f.g.m. della v.c. distribuita normalmente* $\bar{X}$ *e*

$e^{\sum_{i=1}^{n}(t_i-\bar{t})^2\sigma^2/2}$ *è la f.g.m. del vettore casuale a* n dimensioni $\left(X_1-\bar{X}\right),\left(X_2-\bar{X}\right),.....,\left(X_n-\bar{X}\right)$

Nella Fig. 3 è riportato l'andamento della funzione di densità della variabile casuale χ^2 per diversi valori assunti dal parametro caratteristico (*g = 1, 5, 10, 50* gradi di libertà); si può osservare la tendenza della distribuzione alla normalità al crescere dei gradi di libertà.

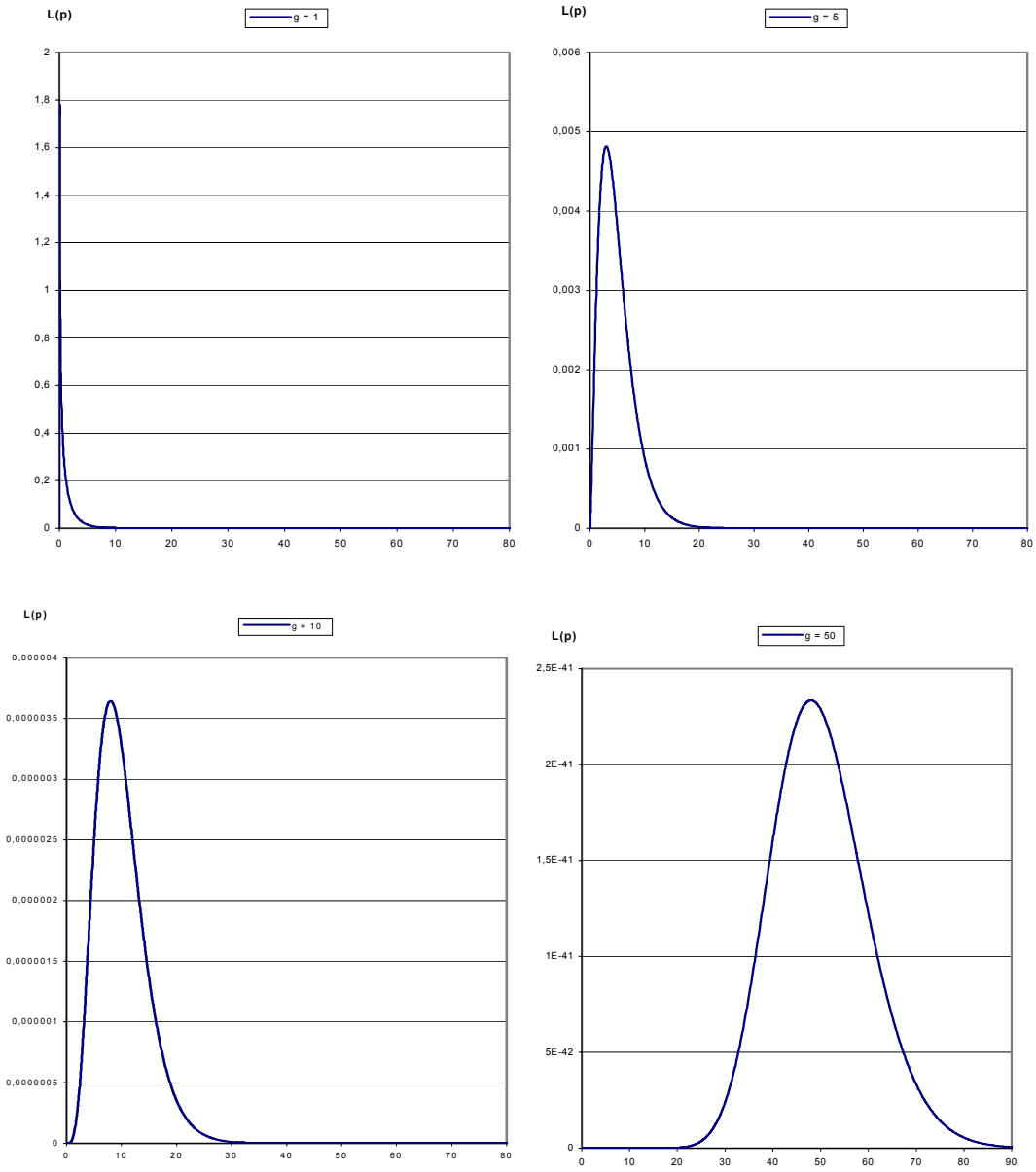

Fig. 3 – Funzione di densità di probabilità della variabile casuale χ^2 per *g = 1, 5, 10, 50*.

Essendo le variabili casuali $\overline{X}$ e V statisticamente indipendenti, ne deriva che la variabile casuale campionaria

$$W = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}}} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

dove

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

ha una distribuzione del tipo *t* di Student con (*n - 1*) gradi di libertà essendo basata sul rapporto tra una variabile casuale normale standardizzata e la radice quadrata di una variabile del tipo χ^2 divisa per i propri gradi di libertà.

Sul concetto di **gradi di libertà** si avrà modo di tornare in seguito, qui basta sottolineare che i gradi di libertà relativi alla variabile casuale campionaria Y sono n perché n sono le variabili casuali indipendenti $(X_1 - \mu), (X_2 - \mu), \ldots (X_n - \mu)$ che entrano nel suo computo. Mentre i gradi di libertà relativi alla variabile casuale campionaria V sono $(n - 1)$ in quanto, pur essendo n gli elementi, le n variabili casuali scarto $\left(X_1 - \overline{X}\right), \left(X_2 - \overline{X}\right), \ldots \left(X_n - \overline{X}\right)$ che entrano nel suo computo , soltanto $(n - 1)$ sono tra loro indipendenti , infatti, le n variabili scarto sono (per costruzione) soggette al vincolo

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$$

3.5 Campionamento da popolazioni non normali

Nei casi in cui l'evidenza empirica o ragioni teoriche escludono la normalità della popolazione cui si riferisce il campione (casuale) di dati a disposizione, e non si hanno altre informazioni sulla popolazione stessa, si può fare ricorso al **teorema del limite centrale** che individua la normale come distribuzione approssimata della variabile

casuale media campionaria. Si riporta di nuovo l'enunciato del teorema nella sua forma più semplice adeguandolo al contesto del campionamento

Teorema 2 (del limite centrale) - Se $X_1, X_2,...,X_n$ costituiscono un campione casuale semplice di n elementi relativi ad una qualunque popolazioni di media μ e varianza (finita) σ^2, allora la variabile casuale media campionaria

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

per n sufficientemente elevato ha una distribuzione approssimativamente normale, con media μ e varianza σ^2/n.

Va sottolineato, inoltre, che la tendenza alla normalità della variabile casuale $\overline{X}$, si realizza anche quando le osservazioni campionarie si riferiscono ad n popolazioni distinte, purchè esse abbiano media e varianza finita. Si avrà pertanto che (ricordando quanto detto a proposito di combinazioni di variabili casuali normali indipendenti) la distribuzione campionaria di una qualsiasi combinazione lineare di medie calcolate su un gruppo di campioni indipendenti tende alla normalità al crescere della numerosità di ciascuno dei campioni considerati.

Benchè il teorema del limite centrale riguardi grandi campioni, nelle applicazioni empiriche più frequenti, l'approssimazione normale risulta soddisfacente anche per campioni di modeste dimensioni. Se le osservazioni campionarie si riferiscono a popolazioni distinte, si avrà una buona approssimazione per i piccoli campioni $(n \leq 30)$ solo quando le distribuzioni di tali popolazioni non si discostano troppo dalla distribuzione normale e le loro varianze non sono molto diverse.

Tornando al problema dell'approssimazione della distribuzione della media campionaria per campioni riferiti ad una stessa popolazione non normale, si deve osservare che la bontà dell'approssimazione dipende, oltre che dalla dimensione campionaria anche dalla natura e dalla forma della distribuzione originaria dalla quale il campione è stato estratto.

Nella Fig. 4 è riportata la distribuzione della media campionaria standardizzata per campioni di diverse dimensioni estratta da popolazioni continue definite dai modelli:

a) *X:* $-\sqrt{3} \le x \le \sqrt{3}$, $f(x) = \frac{\sqrt{3}}{2}$

b) *X:* $x > -1$, $f(x) = e^{-x-1}$

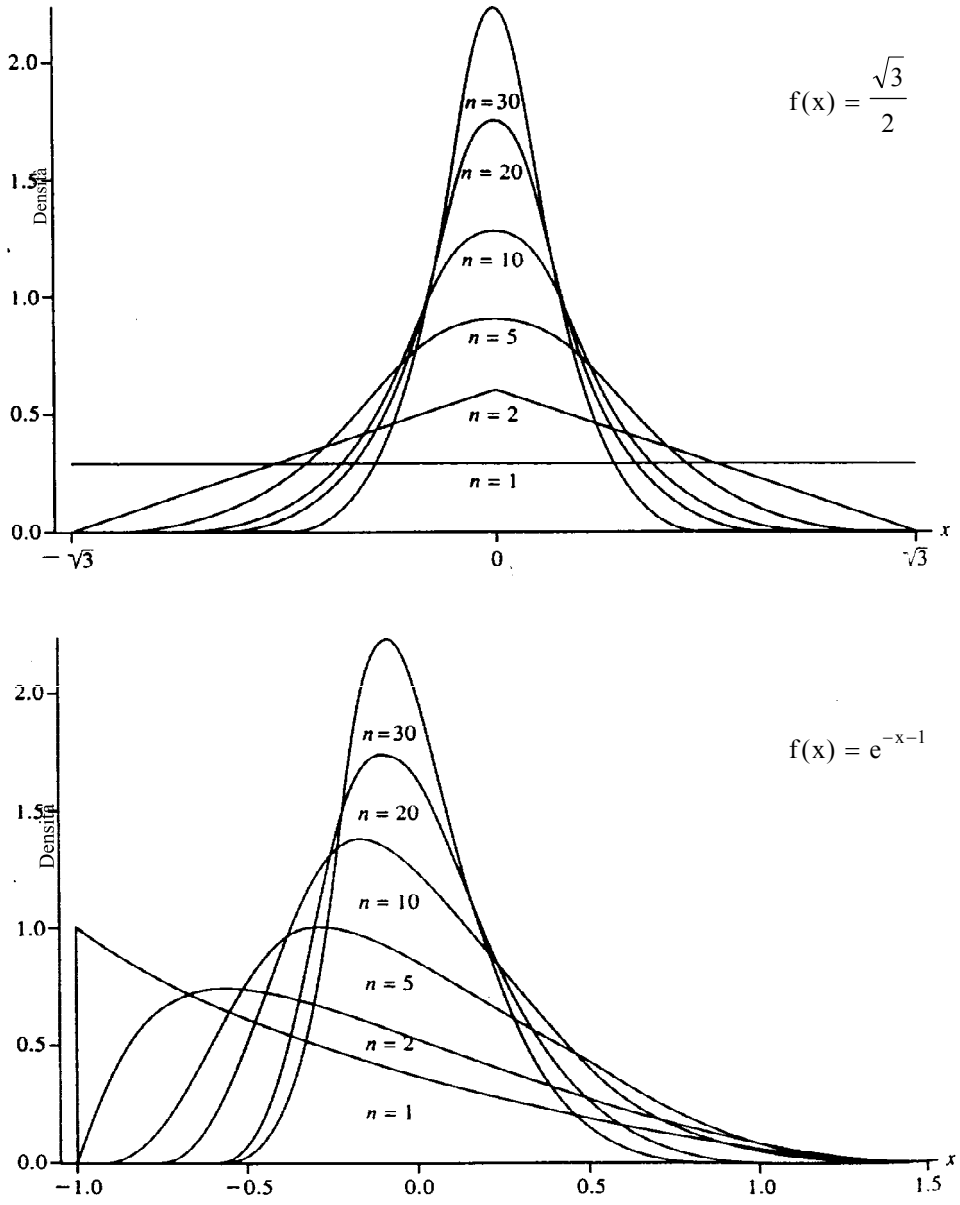

Fig. 4 -Distribuzione della media campionaria per campioni di diverse dimensioni estratti da due diverse popolazioni continue.

Come si può facilmente desumere osservando le figure, a parità di dimensione

campionaria, l'approssimazione migliore è quella relativa alla distribuzione uniforme (distribuzione simmetrica); in generale, si osserva che l'approssimazione della distribuziome normale è tanto più buona quanto più la distribuzione è simmetrica.

Nelle pagine precedenti sono state calcolate media e varianza delle variabili casuali, media campionaria $\overline{X}$ e varianza campionaria (corretta) S^2 , associate a campioni estratti da una qualunque popolazione. Di queste due variabili, di loro trasformazioni e del rapporto tra loro particolari trasformazioni, è stata derivata anche la distribuzione campionaria nel caso di campionamento da popolazioni normali. Si è, inoltre, data indicazione della distribuzione asintotica (cioè della distribuzione cui si perviene facendo tendere ad infinito la dimensione del campione) della media campionaria. Si procederà ora alle stesse elaborazioni in riferimento a proporzioni, a differenze tra medie campionarie e tra proporzioni campionarie e al rapporto tra varianze campionarie con riferimento, in particolare, a campioni estratti da popolazioni normali.

Si supponga di estrarre un campione casuale semplice di dimensione *n* da una popolazione di tipo dicotomico, cioè da una popolazione caratterizzata dalla presenza o meno di un determinato carattere; si supponga inoltre che la proporzione delle unità che possiede il carattere di interesse sia pari a p , mentre $1 - p = q$ è la proporzione delle unità che non possiede il carattere in questione. La popolazione dalla quale viene estratto il campione di dati può essere, in base a quanto detto, rappresentata da una variabile casuale bernoulliana caratterizzata dal parametro $\theta = p$ del tipo

$$X : x_0 = 0 , \; x_1 = 1$$

$$P(X = x_0) = q , \quad P(X = x_1) = p$$

il cui valor medio e varianza sono rispettivamente $\mu = p$ e $\sigma^2 = p\,q$.

Ora, se si considera il punto campionario $(X_1, X_2,...,X_n)$ si vede come, nell'universo dei campioni, ciascuna componente X_i $(i=1, 2, ..., n)$ sia una variabile casuale del tutto simile alla variabile casuale *X* che rappresenta la popolazione.

Si avrà pertanto che la variabile casuale campionaria

$$P = T(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} \frac{X_i}{n}$$

che indica la proporzione delle unità che nel campione presentano quel determinato

carattere, avrà una distribuzione di tipo binomiale **(variabile casuale binomiale relativa**), con valor medio $E(P) = \mu = p$ e varianza $\sigma_p^2 = p\,q/n$. Questa conclusione consente d'interpretare la variabile casuale binomiale relativa, ottenuta attraverso una combinazione lineare di variabili casuali di bernoulli indipendenti, come distribuzione campionaria di proporzioni o percentuali.

Ovviamente, se si definisce come variabile casuale campionaria

$$X_T = \sum_{i=1}^{n} X_i$$

cioè il totale di successi nelle n estrazioni campionarie indipendenti effettuate, tale variabile è esattamente una variabile casuale binomiale con parametri caratteristici n e p , con media $\mu = n\,p$ e varianza $\sigma^2 = n\,p\,q$; il che consente d'interpretare la variabile casuale binomiale come somma di n variabili casuali di bernoulli indipendenti caratterizzate da uno stesso parametro p.

Nelle Figg. 5 e 6 è riportata la distribuzione binomiale (opportunamente standardizzata) per diversi valori di n e di p e la relativa approssimazione con la distribuzione normale. Come si può facilmente desumere osservando le figure, a parità di dimensione campionaria l'approssimazione è tanto più buona quanto più p è prossimo al valore $0,5$ (distribuzione simmetrica); ovviamente l'approssimazione migliora al crescere della dimensione campionaria.

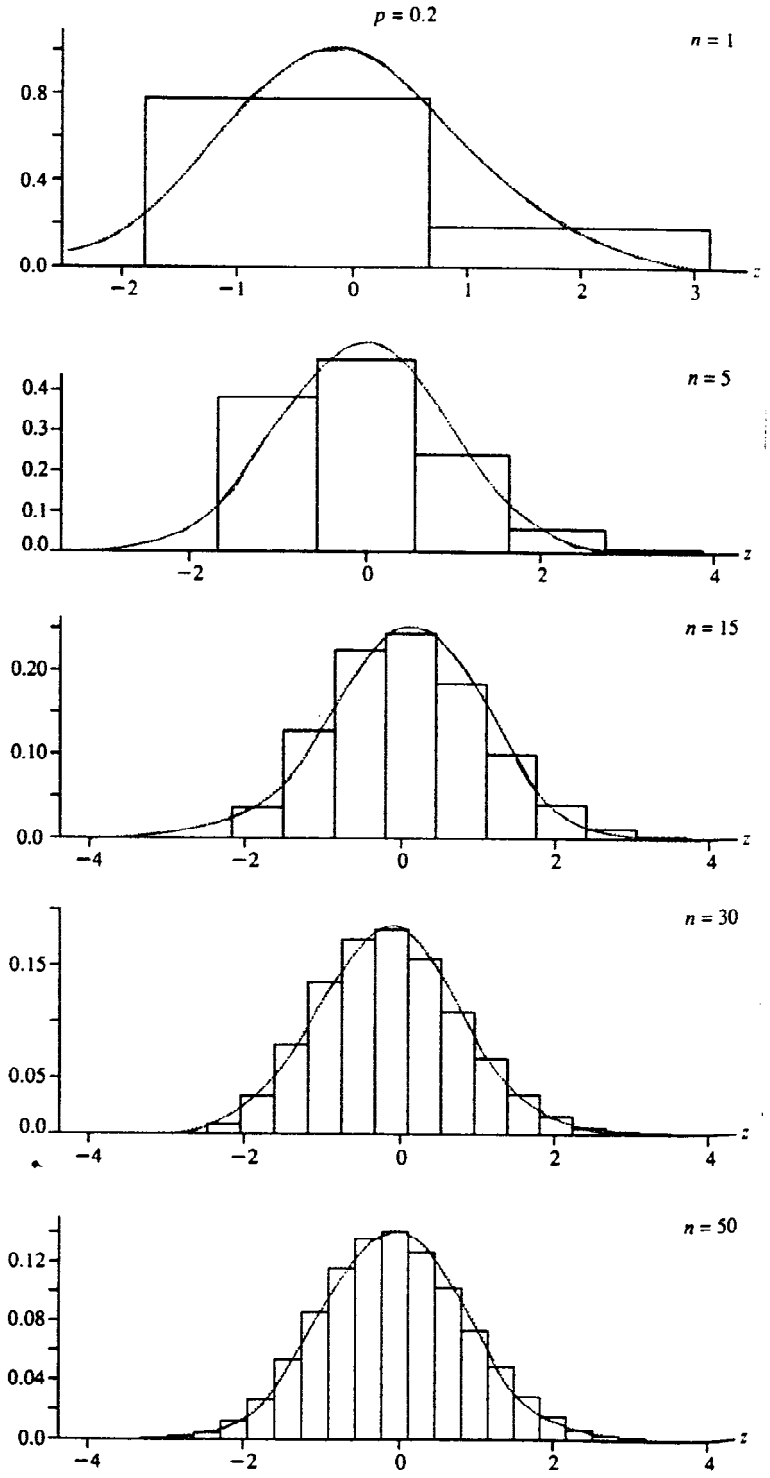

Fig. 5 - Istogrammi della distribuzione binomiale per p = 0,2 e diversi valori di n e relativa approssimazione con la variabile casuale normale standardizzata.

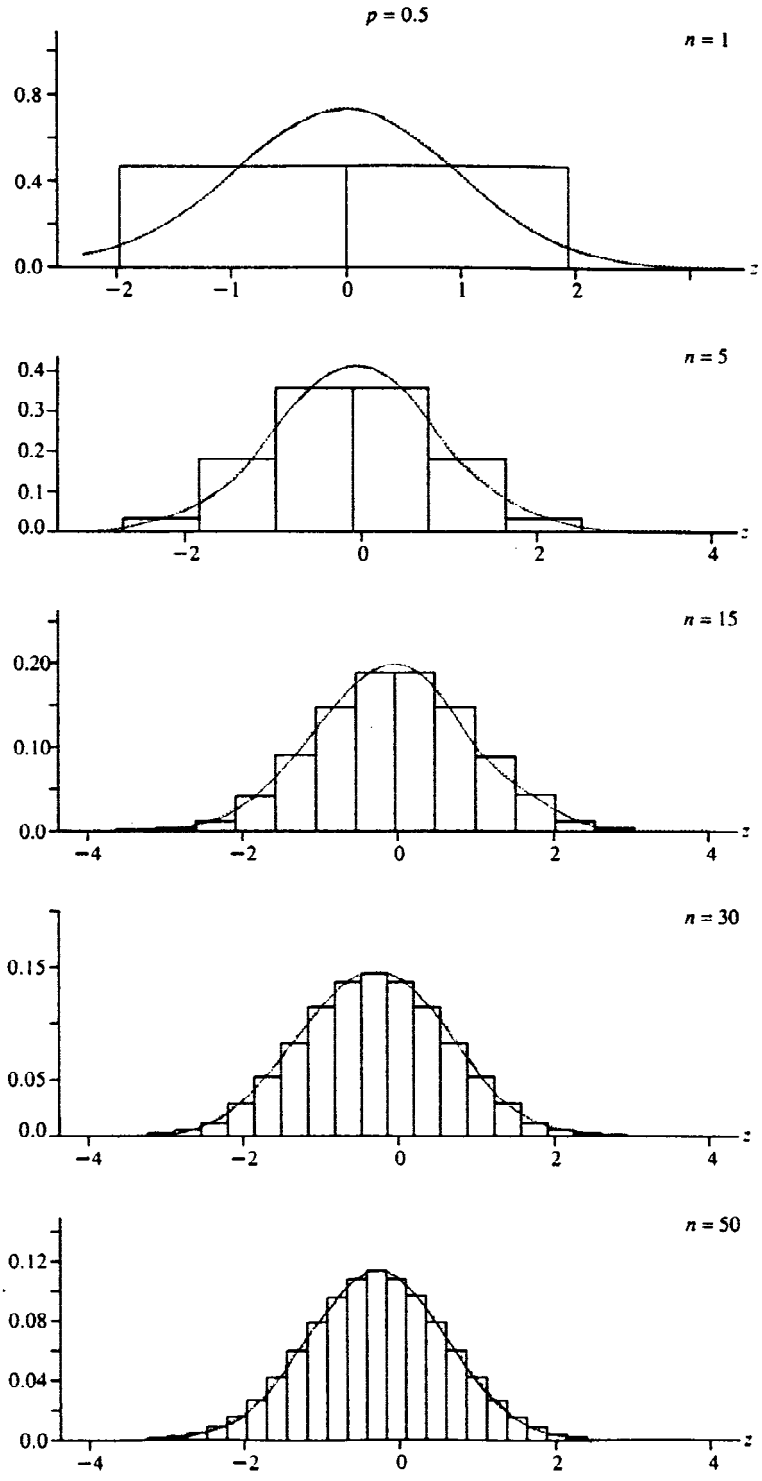

Fig. 6 - Istogrammi della distribuzione binomiale per p = 0,5 e diversi valori di n e relativa approssimazione con la variabile casuale normale standardizzata

3.6 Campionamento da due popolazioni indipendenti

Si supponga ora di estrarre con ripetizione due campioni casuali indipendenti, di dimensione m ed n, da due popolazioni distinte rappresentate dalle variabili casuali X e Y, il cui valore medio e varianza sono rispettivamente μ_x, σ_x^2, μ_y, σ_y^2.

Sugli elementi campionari ($X_1, X_2,...,X_m$) e ($Y_1, Y_2,...,Y_n$) si calcolino le quattro funzioni

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i \quad ; \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$S_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2 \quad ; \quad S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

si calcolino, cioè, le due medie campionarie e le due varianze campionarie corrette, e si definiscono le nuove entità (differenza tra medie campionarie e differenza tra varianze campionarie corrette)

$$V = \overline{X} - \overline{Y}$$
$$S^2 = S_x^2 - S_y^2$$

Le due variabili, nell'universo dei campioni, hanno medie e varianze espresse dalle uguaglianze seguenti

$$E(V) = \mu_x - \mu_y$$
$$Var(V) = \sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2 = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$$
$$E(S^2) = \sigma_x^2 - \sigma_y^2$$
$$Var(S^2) = Var(S_x^2) + Var(S_y^2)$$

Inoltre, se i due campioni sono estratti da popolazioni normali indipendenti vale il seguente teorema

Teorema 3 Se $X_1, X_2, ..., X_m$ costituisce un campione casuale estratto da una popolazione normale di media μ_x e varianza σ_x^2, $Y_1, Y_2, ..., Y_n$ un campione casuale estratto da una popolazione normale di media μ_y e varianza σ_y^2, allora la variabile casuale campionaria:

i) $$U = \overline{X} - \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i$$

è distribuita normalmente con media μ_x - μ_y e varianza $\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}$;

$$V = \sum_{i=1}^{m}\left(\frac{X_i - \mu_x}{\sigma_x}\right)^2 + \sum_{i=1}^{n}\left(\frac{Y_i - \mu_y}{\sigma_y}\right)^2$$

è distribuita come una variabile casuale χ^2 con $m+n$ gradi di libertà;

$$W = \frac{(m-1)S_x^2}{\sigma_x^2} + \frac{(n-1)S_x^2}{\sigma_x^2} = \sum_{i=1}^{m}\left(\frac{X_i - \overline{X}}{\sigma_x}\right)^2 + \sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sigma_y}\right)^2$$

è distribuita come una variabile casuale χ^2 con $m+n-2$ gradi di libertà;

$$F = \frac{\frac{(m-1)S_x^2}{\sigma_x^2}/(m-1)}{\frac{(n-1)S_y^2}{\sigma_y^2}/(n-1)} = \frac{S_x^2}{S_y^2}\cdot\frac{\sigma_y^2}{\sigma_x^2}$$

è distribuita come una variabile casuale F di Fisher-Snedecor con $m-1$ ed $n-1$ gradi di libertà.

Le considerazioni svolte a proposito delle distribuzioni campionarie degli indici sintetici media e varianza, possono essere naturalmente estese ad altri indici caratteristici quali mediana, quartili, scostamento quadratico medio, coefficiente di variazione, ecc.

A proposito della varianza calcolata sulle distribuzioni campionarie di indici sintetici va detto che la sua radice quadrata positiva (scostamento quadratico medio o deviazione standard) viene usualmente denominata **errore standard** o **errore di campionamento**, volendo con ciò sottolineare la sua particolare caratteristica di misura della *bontà* di una stima in termini di variabilità. Su questo punto si avrà comunque modo di soffermarsi a lungo successivamente.