Corso di laurea in

SCIENZE DEL TURISMO AD INDIRIZZO MANAGERIALE

Analisi dei Dati e Revenue Management (Modulo di Statistica)

Docente: Prof.ssa Maria Spano **email:** maria.spano@unina.it

Docente: Prof.ssa Maria Spano (maria.spano@unina.it)

Definizione di campione casuale

In una popolazione infinita, il carattere d'interesse che vogliamo stimare può essere rappresentato da una variabile casuale **X** che potrà essere sintetizzata da due principali parametri:

- **valore atteso** $E(X) = \mu$
- **varianza** $V(X) = \sigma^2$

Data una popolazione infinita **X**, il processo di estrazione di un sottoinsieme di unità statistiche genera il **campione casuale**

una n-pla di v.c. $\mathbf{C=(X_1, X_2, \ldots, X_n)}$

la cui particolare realizzazione è il ***campione osservato***

una n-pla di osservazioni $\mathbf{C_k=(x_1,x_2,\ldots,x_n)}$

Definizione di campione casuale (2)

	1ª oss.	2ª oss.	3ª oss.	4ª oss.	5ª oss.	6ª oss.	7ª oss.	8ª oss.
1° campione	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}
2° campione	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}
3° campione	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
:								
∞° campione	$x_{\infty 1}$	$x_{\infty 2}$	$x_{\infty 3}$	$x_{\infty 4}$	$x_{\infty 5}$	$x_{\infty 6}$	$x_{\infty 7}$	$x_{\infty 8}$

- Se il campione è formato da n elementi, ogni suo elemento può essere considerato come la realizzazione della variabile casuale X_i, indicando con X_i la i-esima estrazione della v.c. X.
- Ciascuna v.c. osservazione campionaria, X_i, avrà la stessa distribuzione e gli stessi parametri della variabile X nella popolazione

Definizione di campione casuale (3)

- Le X_i sono ***indipendenti e identicamente distribuite (i.i.d.)***

- Ciò significa che le X_i saranno tra loro indipendenti
- E che ogni avrà X_i la stessa distribuzione di probabilità, identica a quella della popolazione

In sintesi:

*un **Campione Casuale** è una n-pla di v.c. indipendenti e identicamente distribuite con distribuzione di probabilità pari a $F(X)$*

Statistica campionaria

Poiché ciascuna osservazione campionaria X_i è una variabile casuale, ogni funzione che dipenda dalle osservazioni campionarie sarà essa stessa una variabile casuale

Si definisce **statistica campionaria** *una qualunque funzione a valori reali delle osservazioni campionarie* $(X_1, X_2, \dots, X_n)$

In altre parole una statistica campionaria è rappresentata da una qualunque funzione calcolata sui dati campionari

$$T_n = t(X_1, X_2, \dots, X_n)$$

Il valore della statistica T_n calcolata sul campione osservato $(x_1, x_2, \dots, x_n)$ costituisce la statistica calcolata $t_n = T(x_1, x_2, \dots, x_n)$.

Distribuzione campionaria

Una statistica T_n di variabili casuali (il campione casuale), sarà anch'essa una variabile casuale che assumerà valori diversi al variare delle realizzazioni campionarie.

La probabilità che la statistica T_n assuma un esatto valore t_n è pari. alla probabilità complessiva di tutti i campioni per i quali si ottiene il valore t_n

La distribuzione di probabilità di una statistica campionaria è detta ***distribuzione campionaria***

Università degli Studi di Napoli Federico II
DiSES Dipartimento di Scienze Economiche e Statistiche

Distribuzione campionaria (2)

Dato un *campione casuale*:

$(X_1, X_2, \dots, X_n)$

e definita la statistica:

$T_n = T(X_1, X_2, \dots, X_n)$

ottenuta come elaborazione delle osservazioni campionarie, la distribuzione di probabilità della statistica $T_n(X_1, X_2, \dots, X_n)$ viene definita distribuzione campionaria di T_n.

Ogni statistica è, dunque, una sintesi delle variabili casuali campionarie.

- **media campionaria** ⟶ $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **varianza campionaria** ⟶ $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$
- **varianza campionaria corretta** ⟶ $\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$

Statistiche campionarie e parametri

Le statistiche campionarie non devono essere confuse con i parametri della popolazione

Un ***parametro*** si riferisce all'intera popolazione e ne rappresenta una caratteristica (es. media o varianza della popolazione)

Una ***statistica campionaria*** dipende unicamente dalle osservazioni campionarie e, siccome queste sono variabili casuali, anch'essa sarà una variabile casuale con una propria distribuzione di probabilità

Università degli Studi di Napoli Federico II
DiSES Dipartimento di Scienze Economiche e Statistiche

Ricapitolando……

Prima dell'estrazione…

- Campione casuale $(X_1, X_2, \ldots, X_n)$
- Statistica campionaria $T_n = t(X_1, X_2, \ldots, X_n)$

Dopo l'estrazione…

- Campione estratto $(x_1, x_2, \ldots, x_n)$
- Realizzazione della statistica $t_n = t(x_1, x_2, \ldots, x_n)$

Teorema del Limite Centrale

Siano $X_1, X_2, \ldots, X_n$ una n-pla di variabili casuali indipendenti identicamente distribuite, con media pari a μ e varianza σ^2; posto

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

si ha che la distribuzione della v.c. Z_n

$$Z_n = \frac{(\bar{X}_n - \mu)}{\sigma / \sqrt{n}}$$

Converge, per n->∞, alla v.c. Normale standardizzata

$$\lim_{n \to \infty} Z_n \to Z \sim N(0,1)$$

Teorema del Limite Centrale

Il **teorema del limite centrale** (TLC) afferma quindi che la somma (o la media) di un elevato numero di variabili casuali è approssimativamente normale, indipendentemente dalla distribuzione soggiacente

Il TLC costituisce il fondamento di un gran numero di procedure statistiche.

Difatti, richiamando la definizione di campione casuale (una n-pla di v.c. i.i.d.) allora è intuibile come per alcune statistiche campionarie calcolate su campioni «grandi» <u>diventa approssimativamente nota la distribuzione di probabilità, pur non sapendo nulla a proposito della forma della distribuzione della popolazione</u>

Media campionaria

Tra le statistiche di più frequente utilizzo, la media campionaria ricopre un ruolo di primaria importanza per le desiderabili proprietà campionarie di cui gode.

- Sia X una popolazione infinita con media pari a μ e varianza pa σ^2
- Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale di dimensione

Allora

Si definisce **media campionaria** la seguente statistica:

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

Media campionaria

- Il **Valore atteso della media campionaria** è pari alla media della popolazione

- La **Varianza della media campionaria** è pari alla varianza della popolazione divisa per la dimensione del campione

- L'**Errore standard della media campionaria** (ossia la radice quadrata della varianza di una statistica) è pari alla radice della varianza della media campionaria

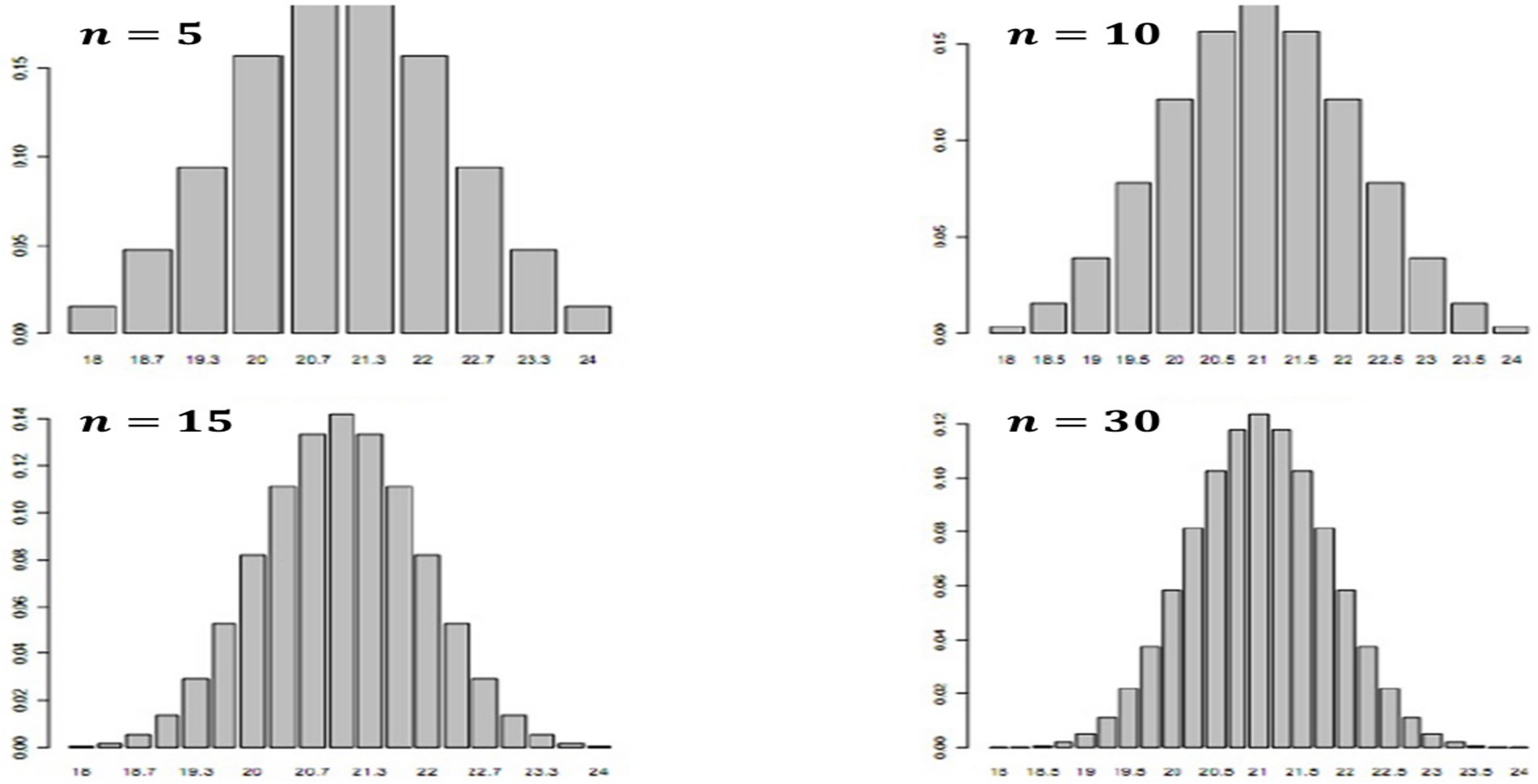
Distribuzione della media campionaria

Se la popolazione ha una distribuzione Normale allora la distribuzione della media campionaria sarà ancora una Normale (a prescindere dalla dimensione campionaria)

$$X \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \sigma^2/n)$$

È interessante osservare come la distribuzione della media campionaria è meno variabile della distribuzione della popolazione, e che inoltre la riduzione della variabilità è tanto più forte quanto maggiore è la dimensione del campione

Relazione tra dimensione del campione e variabilità della media campionaria



Distribuzione della media campionaria

Se ***non è nota la forma della distribuzione della popolazione*** allora si può pervenire ad una buona conoscenza approssimata della distribuzione della media campionaria ***attraverso il Teorema del Limite Centrale***

Quindi ***per n abbastanza grande, la media campionaria si distribuisce approssimativamente come una v.c. Normale***

Solitamente si raggiunge un sufficiente grado di approssimazione con campioni di dimensione n>=30

Media campionaria per popolazioni finite

Sia X una popolazione finita di dimensione N con media pari a μ e varianza pari a σ^2
Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale senza ripetizione di dimensione n

La media campionaria $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ **avrà:**

Valore atteso $E(\bar{X}) = \mu$ **Varianza** $V(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

Quando la dimensione del campione è abbastanza grande e la frazione di campionamento abbastanza piccola , allora per il TLC
la distribuzione della media campionaria può essere approssimata a quella Normale

Campionamento *con* reintroduzione (Popolazioni infinite)

$$E(\bar{X})=\mu \quad ; \quad Var(\bar{X})=\frac{\sigma^2}{n}$$

Campionamento senza reintroduzione (popolazioni finite)

$$E(\bar{X})=\mu \quad ; \quad Var(\bar{X})=\frac{\sigma^2}{n}\cdot\left(\frac{N-n}{N-1}\right)$$

- Quando è $n=1$, i risultati ottenuti con lo schema di campionamento con reintroduzione coincidono con quelli ottenuti nel campionamento senza reintroduzione;
- Quando è $n=N$, la varianza della media campionaria nello schema di campionamento senza reintroduzione è nulla. In questo caso, infatti, il campione coincide con la popolazione e non si ha più alcuna incertezza legata al campionamento;
- Nel caso più comune in cui è $n<N$, il fattore di correzione utilizzato nello schema senza reintroduzione è <1. Questo vuol dire che *la varianza della media campionaria nello schema senza reintroduzione è minore di quella che si ottiene nello schema con reintroduzione*; *(Perché?)*
- Quando è $n<< N$, il fattore di correzione per lo schema di campionamento senza reintroduzione è prossimo a 1. *La differenza tra i due schemi di campionamento può quindi essere considerata trascurabile*. *(Perché?)*

Docente: Prof.ssa Maria Spano (maria.spano@unina.it)

Esempio: media campionaria

Si immagini la variabile «età in anni» per una popolazione

Popolazione X

Distribuzione di X

Unità	Età (X)
A	18
B	19
C	20
D	21

Parametri di X

$$E(X) = \mu = 19.5$$

$$Var(X) = \sigma^2 = 1.25$$

Esempio: media campionaria

- Definiamo lo ***spazio campionario***

L'insieme di tutti i possibili campioni casuali semplici con ripetizione di dimensione

$$C_N^n = N^n \rightarrow C_4^2 = 4^2 = 16$$

Campione		
(A, A)	18	18
(A, B)	18	19
(A, C)	18	20
(A, D)	18	21
(B, A)	19	18
(B, B)	19	19
(B, C)	19	20
(B, D)	19	21
(C, A)	20	18
(C, B)	20	19
(C, C)	20	20
(C, D)	20	21
(D, A)	21	18
(D, B)	21	19
(D, C)	21	20
(D, D)	21	21

Ω

Esempio: media campionaria

Campione			
(A, A)	18	18	18
(A, B)	18	19	18.5
(A, C)	18	20	19
(A, D)	18	21	19.5
(B, A)	19	18	18.5
(B, B)	19	19	19
(B, C)	19	20	19.5
(B, D)	19	21	20
(C, A)	20	18	19
(C, B)	20	19	19.5
(C, C)	20	20	20
(C, D)	20	21	20.5
(D, A)	21	18	19.5
(D, B)	21	19	20
(D, C)	21	20	20.5
(D, D)	21	21	21

Da è possibile derivare la distribuzione della **media campionaria**

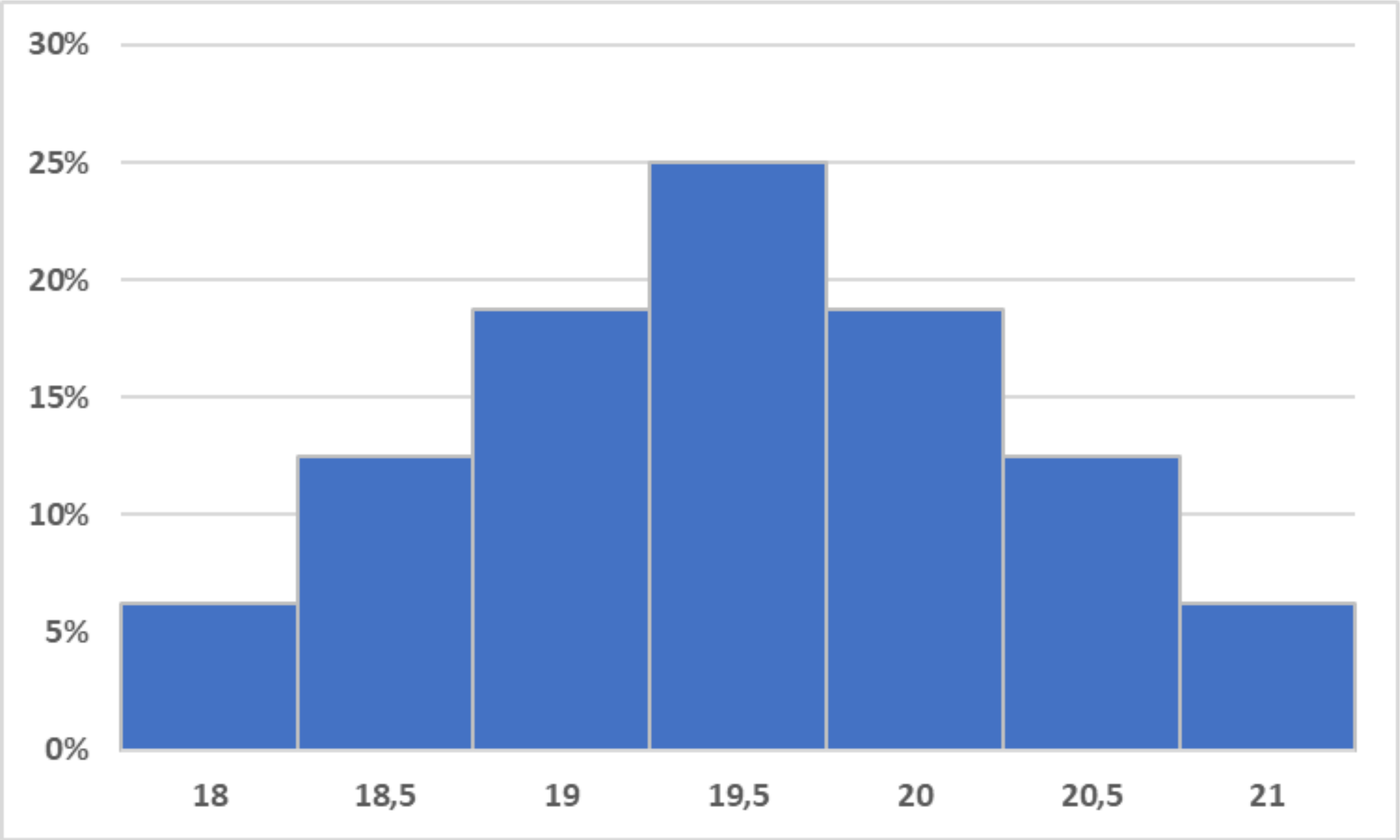
Media Campionaria

$$E(\bar{X}) = \mu = 19.5$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{1.25}{2} = 0.625$$

NB. <u>L'intera colonna dei valori rappresenta la distribuzione della v.c. media campionaria</u>, mentre <u>un singolo valore è la realizzazione della media campionaria in un particolare campione estratto</u>
(es. per il secondo campione, la media è 18.5)

Esempio: distribuzione della media campionaria



Varianza campionaria

Sia una popolazione Normale con media pari con media pari a μ e varianza pari a σ^2

Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale di dimensione *n*

Si definisce **varianza campionaria non corretta** la seguente statistica:

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

mentre

Si definisce **varianza campionaria corretta** la seguente statistica:

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

Varianza campionaria

La statistica S_n^2 si dice non corretta in quanto il suo valore atteso non è uguale alla varianza della popolazione, ma differisce da esso della quantità σ^2/n

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

Al contrario, il valore atteso della statistica $\tilde{S}^2$ è proprio uguale alla varianza della popolazione *X*

$$E(\tilde{S}^2) = \sigma^2$$

La differenza tra le due statistiche diventa trascurabile quando la dmensione campionaria è abbastanza grande

$$n \to \infty \quad \frac{n-1}{n} \to 1 \Rightarrow E(S^2) = E(\tilde{S}^2) = \sigma^2$$

Docente: Prof.ssa Maria Spano (maria.spano@unina.it)

Distribuzione della Varianza campionaria

Attraverso una opportuna trasformazione della varianza campionaria, si ottiene la seguente quantità

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_v \qquad \text{con } v = n-1$$

Che si distribuisce come una chi-quadrato con n-1 gradi di libertà

Campione			
(A, A)	18	18	**0**
(A, B)	18	19	**0.5**
(A, C)	18	20	**2**
(A, D)	18	21	**4.5**
(B, A)	19	18	**0.5**
(B, B)	19	19	**0**
(B, C)	19	20	**0.5**
(B, D)	19	21	**2**
(C, A)	20	18	**2**
(C, B)	20	19	**0.5**
(C, C)	20	20	**0**
(C, D)	20	21	**0.5**
(D, A)	21	18	**4.5**
(D, B)	21	19	**2**
(D, C)	21	20	**0.5**
(D, D)	21	21	**0**

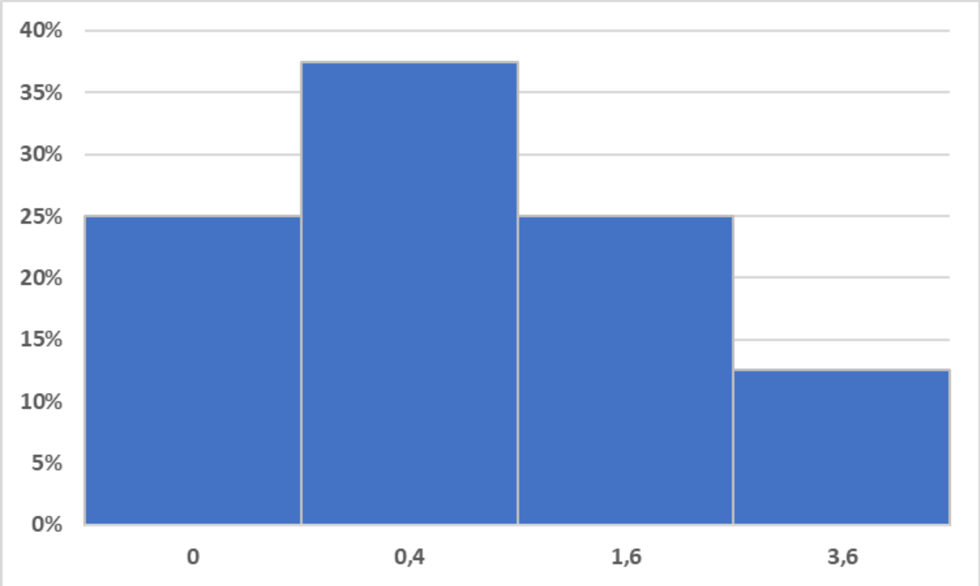
Continuando l'esempio precedente deriviamo la

Varianza Campionaria

Valore atteso

$$E(S^2) = \sigma^2 = 1.25$$

Forma della distribuzione



Proporzione campionaria

- $X \sim B\left(n\pi; \sqrt{n\pi(1-\pi)}\right)$: numero di successi in n prove
- $\frac{X}{n} \sim B\left(\pi; \sqrt{\frac{\pi(1-\pi)}{n}}\right)$: proporzione di successi in n prove

Si definisce **proporzione campionaria** la seguente statistica:

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

Docente: Prof.ssa Maria Spano (maria.spano@unina.it)

Proporzione campionaria

Una popolazione bernoulliana assume solo due valori: 1 (il successo) e 0 (l'insuccesso) e la probabilità del successo ne rappresenta il valor medio.

La statistica P_n ***rappresenta quindi la media campionaria di un campione estratto da una popolazione bernoulliana.***

Essendo tale campione una sequenza di successi e insuccessi, il suo valore medio equivale alla frequenza relativa (o proporzione) dei successi (il numero successi rapportato al numero delle prove)

Proporzione campionaria

Analogamente alla media campionaria

- Il **Valore atteso della proporzione campionaria** è pari alla media della popolazione

- La **Varianza della proporzione campionaria** è pari alla varianza della popolazione divisa per la dimensione del campione

- L'**Errore standard della proporzione campionaria** è pari alla radice della varianza della proporzione campionaria

Distribuzione della Proporzione campionaria

E' possibile derivare la distribuzione approssimata approssimata della distribuzione della proporzione campionaria ***attraverso il Teorema del Limite Centrale***

In questo caso il campione casuale è una successione di v.c. di bernoulli i.i.d, quindi…

per n abbastanza grande, la proporzione campionaria si distribuisce approssimativamente come una v.c. Normale

$$\mathrm{P} \underset{n\to\infty}{\sim} N\left(\pi; \sqrt{\frac{\pi(1-\pi)}{n}}\right) \Rightarrow Z = \frac{\mathrm{P} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0;1)$$