



Corso di laurea in ECONOMIA AZIENDALE

STATISTICA I

Inferenza statistica

9° settimana

A.A. 2008/2009

Carla Rampichini

Inferenza statistica

- Campioni casuali
- Distribuzioni campionarie
- Stima puntuale
- Intervalli di confidenza
- Test delle ipotesi statistiche

Principi di inferenza statistica

■ Statistica Descrittiva

- Raccogliere, presentare, e descrivere i dati

■ Statistica Inferenziale

- Estrarre conclusioni e/o prendere decisioni riguardanti una popolazione sulla base dei dati campionari

Principi di inferenza statistica *(continuazione)*

- Rispetto al calcolo delle probabilità l'ottica è capovolta: i **risultati** della prova sono **noti**, ma non si sa da quale **popolazione** provengono!
- Scopo dell'inferenza statistica è utilizzare i risultati dell'esperimento campionario per giungere alla conoscenza della popolazione che ha generato quei risultati

dai dati osservati per un campione si giunge ad affermazioni che riguardano la popolazione (induzione probabilistica)

Popolazioni e Campioni

- Una **popolazione** è l'insieme di tutte le unità o individui oggetto di studio

□ **Esempi:** Tutti i potenziali votanti nelle prossime elezioni
Tutti i pezzi prodotti oggi
Tutti gli scontrini di novembre

- Un **campione** è un sottoinsieme della popolazione

□ **Esempi:** 1000 votanti selezionati a caso per un'intervista
Alcuni pezzi selezionati per un test di distruzione
Scontrini selezionati a caso per una verifica

Perché usare un campione?

- **Richiede meno tempo** di un censimento
- **Meno costoso** da amministrare di un censimento
- È possibile ottenere risultati statistici con **precisione sufficientemente alta** sulla base dei campioni.

Popolazione e parametri

- Una popolazione **finita** è un **insieme di unità** su cui si può osservare un certo carattere. (es: gli investimenti annui di tutte le aziende di un paese; il numero di figli di ogni famiglia italiana)
- I parametri della popolazione sono delle **costanti** che descrivono aspetti caratteristici della distribuzione del carattere nella popolazione stessa.

Esempio:

- media della popolazione
 $\mu = (1/N) \sum x_i$
- Varianza della popolazione
 $\sigma^2 = (1/N) \sum (x_i - \mu)^2$

- Una popolazione **infinita**: è composta da tutte le unità **potenzialmente osservabili** e non necessariamente già esistenti fisicamente.
- Il carattere d'interesse può essere rappresentato da una variabile casuale con una certa distribuzione di probabilità. In questo caso si indicherà con "popolazione X" la **v.c. X**.

Esempio:

- media della popolazione $\mu = E(X)$
- Varianza della popolazione:
 $\sigma^2 = E[(X - E(X))^2]$

Campione casuale semplice



- È lo schema di campionamento più semplice: corrisponde all'estrazione da un'urna (tipo numeri della tombola)
- Le unità vengono scelte **A CASO** dalla lista e ogni unità ha la stessa probabilità di entrare a far parte del campione.
- **A caso però non vuol dire a casaccio**. Il concetto di caso è infatti strettamente connesso a quello di probabilità: il caso è un concetto intuitivo strettamente connesso all'idea di impossibilità di previsione.
- Ci sono vari modi per fare un'estrazione casuale, tutti cercano di mimare l'estrazione da un'urna:
 - Tavola dei numeri casuali
 - Generazione di numeri casuali e estrazione con il calcolatore



Statistica inferenziale

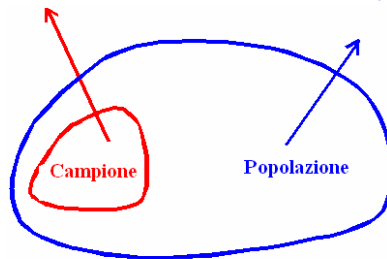
- Facciamo inferenza sulla popolazione esaminando i risultati campionari

Statistiche campionarie → Parametri della popolazione

(note)

Inferenza

(non noti, ma possono essere stimati usando il campione)



Esempio estrazione da v.c. uniforme continua

Statistica inferenziale

continuazione

trarre conclusioni e/o prendere decisioni riguardanti una **popolazione** sulla base dei risultati del **campione**

■ Stima

- Esempio: stimare il peso medio della popolazione usando il peso medio campionario



■ Verifica delle ipotesi

- Esempio: usare le evidenze nel campione per verificare l'affermazione che il peso medio della popolazione è di 120 libbre



Statistica campionaria

- Una statistica campionaria $T_n = T(X_1, \dots, X_n)$ è una funzione a valori reali del campione casuale (X_1, \dots, X_n) che **non** dipende da altre quantità incognite

Esempi di statistiche campionarie:

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

varianza campionaria corretta: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

- La statistica campionaria T_n è una v.c., in quanto funzione delle v.c. (X_1, \dots, X_n) , cui è associata una distribuzione di probabilità detta **distribuzione campionaria**

Distribuzione Campionaria

- La distribuzione della statistica T_n calcolata sullo spazio di tutti i possibili campioni è detta distribuzione campionaria di T_n
- Quindi, la **distribuzione campionaria** è la distribuzione di tutti i possibili valori di T_n ottenuti da campioni della stessa ampiezza estratti dalla popolazione
- Vari metodi per lo studio della distribuzione campionaria:
 - Derivazione della distribuzione esatta a partire da $X \sim F(x; \theta)$
 - Distribuzione approssimata di tipo asintotico, in base ai teoremi limite del calcolo delle probabilità
 - Distribuzione approssimata di tipo numerico, mediante simulazione.

Sviluppo della distribuzione campionaria

- Assumiamo che ci sia una popolazione ...
- dimensione della popolazione $N=4$
- variabile aleatoria, X , è l'età degli individui
- Valori di X :
18, 20, 22, 24 (in anni)



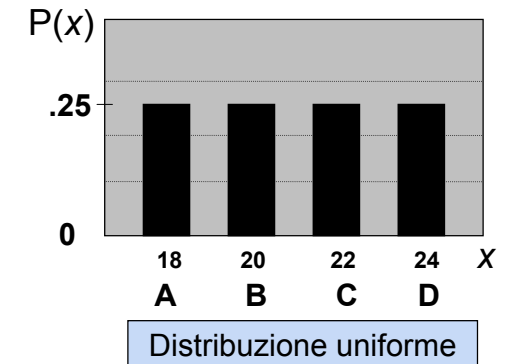
Sviluppo della distribuzione campionaria

(continuazione)

Misure di sintesi della distribuzione della Popolazione:

$$\mu = \frac{\sum X_i}{N} = \frac{18+20+22+24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



Sviluppo della distribuzione campionaria

(continuazione)

Adesso consideriamo tutti i possibili campioni di dimensione $n = 2$

1 ^a Oss	2 ^a Osservazione			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possibili campioni (campionamento con reintroduzione)

16 medie campionarie

1 ^a Oss	2 ^a Osservazione			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

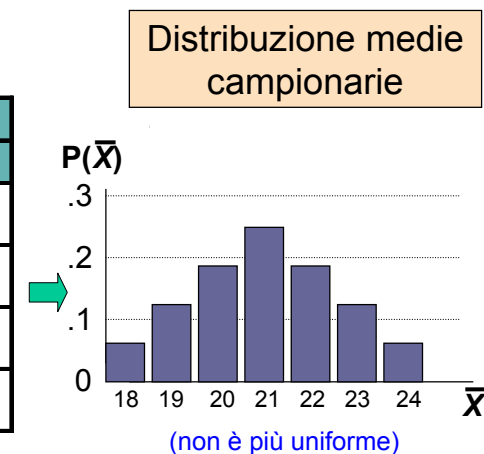
Sviluppo della distribuzione campionaria

(continuazione)

Distribuzione di tutte le medie campionarie

16 medie campionarie

1 ^a Oss	2 ^a Osservazione			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24



Sviluppo della distribuzione campionaria

(continuazione)

Misure di sintesi della distribuzione campionaria

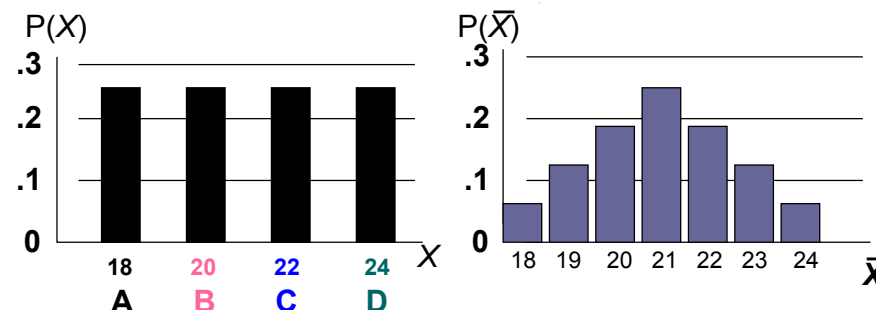
$$E(\bar{X}) = \frac{\sum \bar{X}_i}{N} = \frac{18+19+21+\dots+24}{16} = 21 = \mu$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum (\bar{X}_i - \mu)^2}{N}} = \sqrt{\frac{(18-21)^2 + (19-21)^2 + \dots + (24-21)^2}{16}} = 1.58$$

Confronto tra la popolazione e la distribuzione campionaria della media

Popolazione v.a. X
N = 4
 $\mu = 21$ $\sigma = 2.236$

media campionaria
n = 2
 $\mu_{\bar{X}} = 21$ $\sigma_{\bar{X}} = 1.58$



Errore standard della media campionaria

- Diversi campioni della stessa dimensione estratti dalla stessa popolazione produrranno medie campionarie diverse
- Una misura della variabilità nel valore della media da campione a campione è dato dall'errore standard della media

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Notare che l'errore standard della media diminuisce aumentando la dimensione del campione

Correzione per popolazioni finite

- $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$ o $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$
- Applicare la **correzione per popolazioni finite** se:
 - un elemento della popolazione non può essere incluso nel campione più di una volta (campionamento senza reintroduzione)
 - e
 - il campione è grande rispetto alla popolazione (n è superiore al 5% di N)

Esempio: distribuzione ESATTA della media campionaria ...

- Consideriamo una popolazione finita composta dalle seguenti 5 unità: **8,4,2,11,6**, $\mu=6.2$
- Generiamo tutti i possibili N... $(N-k+1)=5*4=20$ campioni senza ripetizione di dimensione $n=2$, cioè l'**universo dei campioni**, e per ogni campione calcoliamo la **media campionaria**

X_1	8	8	8	8	4	4	4	4	2	2	2	2	11	11	11	11	6	6	6	
X_2	4	2	11	6	8	2	11	6	8	4	11	6	8	4	2	6	8	4	2	11
\bar{X}	6	5	9,5	7	6	3	7,5	5	5	3	6,5	4	9,5	7,5	6,5	8,5	7	5	4	8,5

- Poiché ogni campione ha probabilità $1/20=0.05$ di essere estratto, la distribuzione della media campionaria è

\bar{X}	3	4	5	6	6,5	7	7,5	8,5	9,5
$P(\bar{X})$	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1

→ $E(\bar{X})=6.2$

Correzione per popolazioni finite

- Nell'esempio precedente, $\sigma^2=9.76$
- Quindi applicando la formula per popolazioni infinite si ottiene

$$\text{Var}(\bar{X}) = 9.76/2 = 4.88$$

- Mentre considerando la distribuzione esatta si ha $V(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = 3.66$

Che coincide con la formula 'corretta'

$$\frac{\sigma^2}{n} \frac{N-n}{N-1} = 4.88 \frac{5-2}{5-1} = 3.66$$

Distribuzione della media campionaria per popolazione normale

- Se la popolazione è **normale** con media μ e scarto quadratico medio σ , allora anche la distribuzione campionaria di \bar{X} è **normale** con

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Z per la distribuzione della media campionaria

- v.a. Z standardizzata per \bar{X}

$$Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

- dove:
- \bar{X} = media campionaria
 - μ = media della popolazione
 - σ = scarto quadratico medio della popolazione
 - n = dimensione del campione

Standardizzazione della media campionaria per popolazioni finite

- Se la dimensione del campione n non è abbastanza piccola rispetto alla dimensione della popolazione N , allora usa

$$Z = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

Media campionaria

- data la v.c. $X \sim F(x; \theta)$, si estrae da X un campione casuale (X_1, \dots, X_n)
- La statistica 'media campionaria' è definita da:

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Proprietà della media campionaria

$$E(\bar{X}_n) = \mu \Rightarrow \text{centrata sul parametro}$$

$$V(\bar{X}_n) = \frac{1}{n} \sigma^2 \Rightarrow \lim_{n \rightarrow \infty} V(\bar{X}_n) = 0$$

$$X \sim N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

per la proprietà
riproduttiva
della v.c. Normale



Proprietà riproduttiva della v.c. Normale

- La **combinazione lineare di v.c. normali indipendenti** è ancora una **v.c. normale** con valore atteso pari alla combinazione lineare dei valori attesi e varianza pari alla combinazione lineare delle varianze con i coefficienti al quadrato

$$\left. \begin{array}{l} X_i \sim N(\mu_i, \sigma_i^2) \quad i=1, \dots, n, \text{ indep.} \\ Y = \sum_{i=1}^n a_i X_i \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} Y \sim N(\mu, \sigma^2) \\ \mu = \sum_{i=1}^n a_i \mu_i, \quad \sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2 \end{array} \right.$$

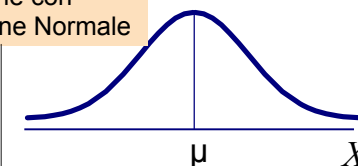
Proprietà della distribuzione campionaria

-

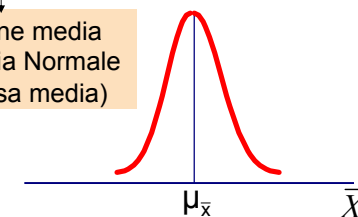
$$\mu_{\bar{X}} = \mu$$

\bar{X} è non distorto
unbiased

Popolazione con
distribuzione Normale



Distribuzione media
campionaria Normale
(ha la stessa media)

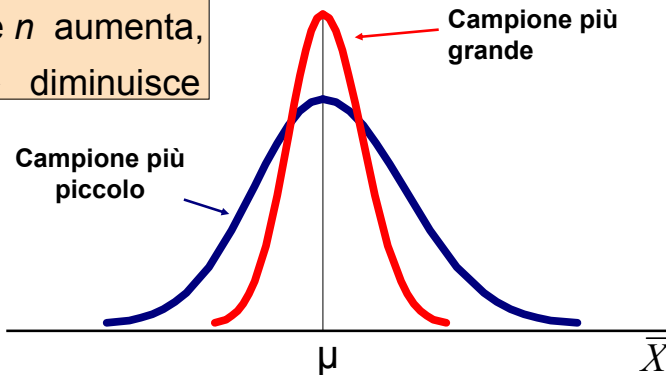


Proprietà della distribuzione campionaria

(continuazione)

- Per campionamento con reintroduzione:

Se n aumenta,
 $\sigma_{\bar{x}}$ diminuisce



Se la popolazione **non** è Normale

- Possiamo applicare il Teorema del limite centrale:

- Anche se la popolazione **non** è normale,
- ...la media campionaria della popolazione sarà **approssimativamente normale** purché l'ampiezza del campione sia abbastanza grande.

Proprietà della distribuzione campionaria:

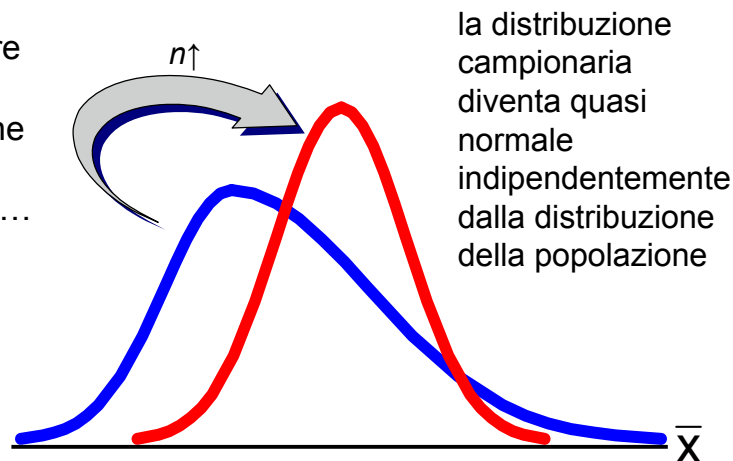
$$\mu_{\bar{x}} = \mu$$

e

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Teorema del Limite Centrale

Al crescere della dimensione del campione...



Se la popolazione **non** è Normale

(continuazione)

Proprietà distribuzione campionaria:

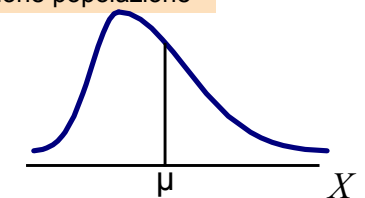
Tendenza Centrale

$$\mu_{\bar{x}} = \mu$$

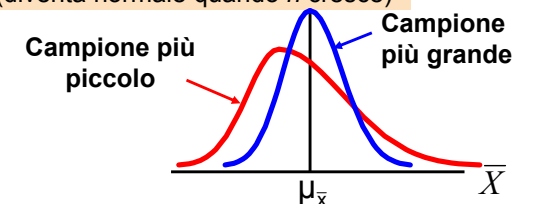
Dispersione

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Distribuzione popolazione



Distribuzione campionaria (diventa normale quando n cresce)



Quanto deve essere grande il campione?

- Per la maggior parte delle distribuzioni, $n > 25$ produce una distribuzione della media campionaria approssimativamente normale
- Per popolazioni con distribuzione normale, la distribuzione della media campionaria è sempre una distribuzione normale

Esempio

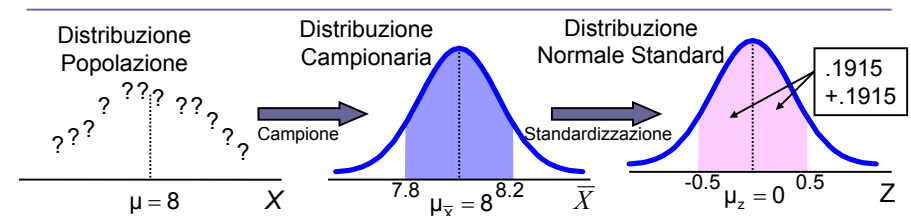
- Supponiamo che una popolazione abbia media $\mu = 8$ e scarto quadratico medio $\sigma = 3$. Assumiamo che sia stato selezionato un campione casuale di dimensione $n = 36$.
- Qual è la probabilità che la **media campionaria** sia compresa fra 7.8 e 8.2?

Esempio: soluzione

- Anche se la popolazione non ha distribuzione normale, poichè $n > 25$ si può fare riferimento al teorema del limite centrale ...
- ... quindi la distribuzione campionaria di \bar{X} è approssimativamente normale
- ... con media $\mu_{\bar{X}} = 8$
- e scarto quadratico medio $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

Esempio soluzione (continuazione)

$$\begin{aligned} P(7.8 < \mu_{\bar{X}} < 8.2) &= P\left(\frac{7.8-8}{\frac{3}{\sqrt{36}}} < \frac{\mu_{\bar{X}} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{8.2-8}{\frac{3}{\sqrt{36}}}\right) \\ &= P(-0.5 < Z < 0.5) = 0.3830 \end{aligned}$$



Media campionaria e legge dei grandi numeri

- Abbiamo visto che la *media campionaria varia da campione a campione*, perchè allora questa è una stima ragionevole della media μ (*incognita*) della popolazione?
- La risposta risiede nel fatto che è possibile dimostrare che *se si aumenta progressivamente l'ampiezza del campione*, la media campionaria *sicuramente* si avvicina sempre di più al valore del parametro μ .
- Questo fatto importante è chiamato *legge dei grandi numeri* ed è un risultato valido per qualsiasi tipo di popolazione e non soltanto per alcune classi particolari come quella delle distribuzioni Normali.

Distribuzione asintotica della media campionaria

- Per il Teorema del limite centrale, la distribuzione della media campionaria tende alla distribuzione Normale per $n \rightarrow \infty$

Teorema del limite centrale (Lindeberg e Lévy)

- Se X_n è una successione di v.c. indipendenti e identicamente distribuite con valore medio μ e varianza $0 < \sigma^2 < +\infty$, allora la v.c. somma standardizzata Z_n tende ad avere una distribuzione Normale standardizzata

$$Z_n = \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} Z \sim N(0,1)$$

Distribuzione della media campionaria

- Se $X \sim \text{Bernoulli}(\pi)$ allora $\bar{X}_n \sim \frac{1}{n} \text{Bin}(n, \pi)$
- Se $X \sim N(\mu, \sigma^2)$ allora $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- Qualunque sia la popolazione, per il **teorema del limite centrale** si ha

$$\lim_{n \rightarrow \infty} P\left(\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \leq z\right) = P(Z \leq z), Z \sim N(0,1)$$

Intervalli di accettazione

- Obiettivo: determinare un intervallo entro il quale verosimilmente cadono i valori della media campionaria, per una data media e varianza della popolazione
 - Dal teorema del limite centrale, sappiamo che la distribuzione di \bar{X} è approssimativamente normale se n è abbastanza grande, con media μ e scarto quadratico medio $\sigma_{\bar{X}}$
 - Sia $z_{\alpha/2}$ il valore di Z che lascia nella coda destra della distribuzione normale standard l'area $\alpha/2$ (cioè l'intervallo da $-z_{\alpha/2}$ a $z_{\alpha/2}$ racchiude una probabilità $1 - \alpha$)
 - Allora $\mu \pm z_{\alpha/2} \sigma_{\bar{X}}$

è l'intervallo che include \bar{X} con probabilità $1 - \alpha$

Utilizzo intervalli di accettazione

- Nel controllo di qualità dei processi produttivi l'intervallo $\mu \pm z_{\alpha/2} \sigma_{\bar{x}}$ viene monitorato a intervalli regolari
- di solito si considera $\alpha < 0.01$
- Esempio: controllo produzione scatole di biscotti peso deve essere 250 gr.
 - Ogni 30' si estrae un campione di $n=5$ scatole
 - Si calcola il peso medio delle scatole estratte
 - Si rappresenta l'andamento del peso medio su una CARTA di CONTROLLO
 - L'intervallo $250 \pm 3\sigma_{\bar{x}}$ comprende il 99% circa delle medie campionarie sotto l'ipotesi di normalità
 - Se la media cade al di fuori \rightarrow problemi nella produzione!

Proporzione della popolazione, π

π = la proporzione della popolazione che possiede le caratteristiche oggetto di studio

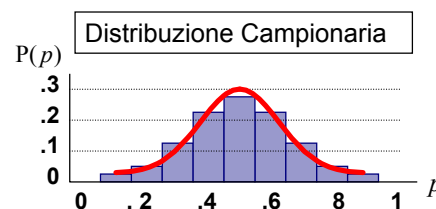
- **Proporzione campionaria p** fornisce una stima di π .

$$p = \frac{X}{n} = \frac{\text{numero di unità nel campione aventi le caratteristiche oggetto di studio}}{\text{dimensione del campione}}$$

- $0 \leq p \leq 1$
- X ha una distribuzione binomiale, ma può essere approssimata da una distribuzione normale quando $np(1-p) > 9$

Distribuzione campionaria di p

- **Approssimazione Normale:**



Proprietà:

$$E(p) = \pi$$

e

$$\sigma_p^2 = \text{Var}\left(\frac{X}{n}\right) = \frac{\pi(1-\pi)}{n}$$

(dove π = proporzione della popolazione)

V. C standardizzata Z per proporzioni

Passiamo da p a Z usando la formula:

$$Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Esempio

- Se la vera proporzione di votanti che sostiene la Proposta A è $\pi = 0.4$, qual è la probabilità che un campione di dimensione 200 produca una proporzione campionaria compresa tra 0.40 e 0.45?

- cioè: **se $\pi = .4$ e $n = 200$, quanto è $P(.40 \leq p \leq .45)$?**

Esempio

(continuazione)

- **se $\pi = 0.4$ e $n = 200$, quanto vale $P(.40 \leq p \leq .45)$?**

Trova σ_p : $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = .03464$

In termini della distribuzione normale standard:

$$P(0.40 \leq p \leq 0.45) = P\left(\frac{.40 - .40}{.03464} \leq Z \leq \frac{.45 - .40}{.03464}\right) = P(0 \leq Z \leq 1.44)$$

Esempio

(continuazione)

- **se $\pi = .4$ e $n = 200$, quanto è $P(.40 \leq p \leq .45)$?**

Dalla tavola della normale standard: $P(0 \leq Z \leq 1.44) = .4251$

